Alma Mater Studiorum - Universitá di Bologna

DOTTORATO DI RICERCA IN

ECONOMIA

Ciclo 30

Settore Concorsuale: 13/A1

Settore Scientifico Disciplinare: SECS-P/01

Emergency Medical Performance: Why and How Should We Care?

Presentata da: Elena Lucchese

Coordinatore Dottorato:

Prof. Marco Casari

Supervisori:

Prof.ssa Margherita Fort

Prof. Matteo Lippi Bruni

Esame finale anno 2018

Emergency Medical Performance: Why and How Should We Care?

PhD candidate: Elena Lucchese*
October 15th, 2018

Abstract

The research in economics is devoting a growing attention to the study of the performance in the Emergency Medical System (EMS) setting. In such context, timely actions are considered of key importance, and higher degree of performance are often directly related to the size of the investment made. However, technological constraints limited the possibility of collecting adequate quality data, and so it was not possible to precisely quantify the relationship between EMS performance and health conditions of patients. Thanks to recent progresses, in some cases is now possible to gain the access to good quality data, and a growing literature is investigating more in deep such relationship. Conditional on the availability of adequate softwares to support data collection, the EMS setting is an interesting context for several reasons. First, the performance level in this context are measured by a set of standard indicators, which improves the external sound of results. Examples of most classical measures are the ambulance response time – i.e. the amount of minutes required to reach the patient once an ambulance is requested – and the severity and mortality of patients observed after different time horizons – usually between one hour and one year. Other common measures are the duration of the waiting time for the patients in the Emergency Department's waiting room,

*Department of Economics, University of Bologna, Piazza Scaravilli 2, 40126 Bologna, Italy. E-mail: elena.lucchese@unibo.it. I wish to thank my supervisors, Margherita Fort and Matteo Lippi Bruni, to which I am indebted. I am thankful to Paolo Roberti, Giacomo Pasini, Joseph Doyle, Marco Bertoni, Jayhanta Bhattacharya, Jerome Adda, Nicola Persico, Decio Coviello, Veronica Grembi, Sara Lazzaroni, Mark Duggan and Vincenzo Atella, and seminar participants at EEA/ESEM, Brucchi Luchino, CRENoS, University of Bologna and Stanford University for helpful comments and suggestions. I acknowledge support from AIES, EIEF and the University of Bologna. A special thank goes to Francesco Bermano, Edoardo Berti Riboli and Domenico Gallo without whom this project would have not be possible and to Andrea Furgani for his constant support.

the duration of the medical visit once a care provider takes care of the patient, the re-admission probability and the mortality rate after several time horizons.

With my work, I obtained the access to two high quality administrative datasets that collect information on all ambulance missions performed in the Italian region Liguria in 2013-2014, accounting for over 1.3 million observations, and on all accesses in the Emergency Departments in Liguria in 2013-2015, accounting for over 1.3 million observations. Liguria is an ideal framework for this analysis for two main reasons. First, in 2012 Liguria was appointed by the European Union as one of the pilot regions in charge of documenting best practices for the EMS, as a part of a project aimed at harmonizing the health system in Europe. Second, in Italy – as well as in most European countries – the health care system is organized and provided at the regional level, and so the regional dimension represents a relevant dimension.

The questions that I investigate with my research, help understanding more about the effect of EMS services on patient conditions. In the first chapter of the thesis, I study the causal relationship between the ambulance response time and patient's severity and mortality rates. I address the problem of endogeneity due to reverse causality by exploiting the hourly amount of rainfall as instrument for the ambulance response time. In the second chapter, I investigate the inefficiency due to the problem of patients' localization. I perform the analysis at the level of each ambulance mission, and I adopt a difference-in-differences identification strategy to clear the effect of interest by the compounding effect of other unobservables that may correlate with the effect of interest. Finally, in the third chapter I study how the productivity of care providers working in the emergency department changes when the end of shift is approaching. This last work contributes to the growing literature that studies the determinants of the performance of intrinsically motivated workers ruled by incomplete contracts and where the activity is scheduled. I make use of a theoretical model with testable implications to support the idea that lower performance near the end of shift is due (at least to) to fatigue.

Chapter 1

The causal effect of ambulance response time on cardiovascular severity and mortality

Elena Lucchese*

Abstract

Ambulance providers devote vast resources to minimizing the time that it takes them to attend the scene of an accident. However, existing empirical evidence reports no meaningful effects of ambulance response time on out-of-hospital patients' health conditions, before receiving in-hospital medical treatments. I revisit this question using a rich administrative dataset from the Emergency Medical System in Liguria. To identify causal effects, I exploit changes in the amount of hourly rainfall in the municipality in which the ambulance performs the mission. Contrary to previous evidence, I find a large and strongly significant effect: a 1 minute increase in response time – 3% at the sample average – leads to a 1.5 percentage points increase in the likelihood of severe health impairment and 0.7 percentage points increase in the likelihood that the patients dies by the time she is transported to the hospital. Given these findings, I estimate the indifference curve of policy maker to invest in policies that improve the performance of emergency care services and show that, although both desirable policies, introducing advance location systems for the caller is substantially more cost-effective than increasing the number of available ambulances.

JEL classification: C26, I12, I18, R41

Keywords: Response time, emergency care, organisational performance, mortality rate

1 Introduction

The study of the determinants of health is a central component of the standard economic model of health (Davis, 1956; Peltzman, 1973; Mokyr, 1993). It is also

*Department of Economics, University of Bologna, Piazza Scaravilli 2, 40126 Bologna, Italy. E-mail: elena.lucchese@unibo.it.

instrumental to support policymakers in the understanding of the intricacies of health care institutions and technologies (Fuchs et al., 1996). The aim of this work is to gain more knowledge about the determinants of health and to improve the understanding of the costs and benefits related to the implementation of reforms in the health care system, such as the Affordable Care Act (ACA) in the United States. The introduction of the ACA, for instance, expanded the access to ambulance transportation to a larger share of the population. This change added strain to the response system and slowed average response time of 19% (Courtemanche et al., 2017). However, it is not clear whether longer responses have a negative impact on patients' outcome. Recent literature on the economics of crime has shown that the response time is a key determinant of police performance (Blanes i Vidal and Kirchmaier, 2017) – a setting that shares the trait of acting during an emergency with the service provided by ambulances. However, although it seems natural to assume that shorter ambulance responses are important for patients outcomes, evidence in support of this belief is lacking.

This paper builds on a recent stream of literature in economics that exploit the ambulance setting to learn more about to effect of hospital care on patient outcome. These research works exploit the observed regularity between ambulance company that attend the scene of an accident and hospital of destination, to study whether patients located close to each other, but transported to different hospitals, achieve better health when transported to higher-spending (Doyle Jr et al., 2015, 2019) and higher-ranking (Hull, 2018) hospitals. Although it is possible that different ambulance companies assign similar response times to patients – even when they transport them to different hospitals – it is worth to directly investigate whether the response time is in itself a crucial determinant of the morbidity and mortality rates of patients before their access to the hospital. This is important as different in-hospital outcomes may be partially driven by the health condition of patients at the time of hospital admission due to response time.

Given the importance attributed to response time, emergency care providers devote vast resources to minimizing it: they track and publicize response time statistics and they include response times as part of the core performance measures by which they are evaluated.[1] The effectiveness of rapid response has, however,

[1]In Europe, where the access to the ambulance service is usually provided free of charge to all citizens, response time targets are generally established by regulation or by national law. Additional information about response targets in Europe can be find in the web page of the European Commission dedicated to public health: ec.europa.eu/health. In the United States, local health care agencies contractually agree response time levels with ambulance providers (Ludwig, 2004) and response time specifications have come to serve as a fundamental measure of the emergency medical system performance (see Swor (1993) and as one of the standards imposed by the Commission on Accreditation of Ambulance Services (CAAS), Glenview, IL: CAAS, 2017).

long been questioned by health scholars. The notion that rapid response has limited beneficial effect on patients health outcomes is one of the most well-established paradigms in the medical literature.[2] Swor and Cone (2002) is more specific, concluding that *regrettably, the scientific basis that documents clinical benefit to this rapid response is virtually absent.* As a result, the argument commonly put forward by specialists is that response time matters only within the first few minutes after the health shock, an unrealistically short interval for even the most efficient ambulance provider (Blackwell and Kaufman, 2002; O'Keeffe et al., 2010).

Existing evidence on the effect of ambulance response time on patient outcome is far from convincing. This is unsurprising, as adequate softwares to support the collection of good-quality data have been developed only recently and are not yet widely adopted. Moreover, public-use mission-level datasets do not provide information on out-of-hospital conditions of patients, and therefore requires the unlikely collaboration of ambulance providers. Additionally, there is the problem of endogeneity in response time. Ambulance missions assigned a higher priority could be characterized with an ex ante higher or lower likelihood of saving the patient that is observed by the ambulance driver and that affects the response time (Wilde, 2013). Furthermore, most critical cases are likely to receive more resources from the health system ex-post. As a result, identifying the causal effect of response time is a challenging exercise.

In this work I estimate the effect of ambulance response time on the health condition of patients by making use of a rich administrative dataset and a research design that exploits discontinuities in response time due to rainfall. The dataset comprises of the 2013-2014 internal records of the Emergency Medical System on each ambulance mission activated to rescue patients affected by cardiovascular problems in the Italian region Liguria, which oversees a population of 1.6 million and a geographic extension of 5410 square kilometers. In Italy, as in most European countries, this geographical dimension is particularly meaningful given that most health care services are organized at the regional level. A particularly interesting feature of the data adopted in this analysis is related to the informative system that was implemented in Liguria in 2012 with the aim of improving the quality of the data collected.[3] After this innovation, more information is collected at the spot, minimizing the likelihood of mistakes and misreporting that might instead occurs in similar collections, when the information is self-reported

[2] See Pons et al. (2005) and the review of the literature by Nichol et al. (1996).

[3] The Emergency Medical Service (EMS) in Italy is one of the most advanced in the world and of inspiration for policymaker of other countries. For instance, Italy was the first, in 1990, to introduce a unique number for emergencies and to centralize the coordination of ambulance activity. In addition, Liguria is one of the regions in Europe that is accounted by the European Union to documents its EMS practices in order to provide insights for other countries and harmonize the health system in Europe.

by ambulance providers at the end of the mission. In addition, the record of the ambulance response time – our main regressor of interest – is automated, with the software that keeps track, among others things, of the exact moment in which the emergency call starts and when care providers reach the scene of the event. Finally, patients affected by cardiovascular problems presents standard symptoms so that the likelihood of misjudgments is little, and the degree of severity is detected by a trained professional usually a paramedic, in three points in time, namely during the emergency call, at the arrival on the scene, and at the hospital admission. After controlling for a rich set of covariates, I make use of the degree of severity observed at the scene to learn more about the direct effect of response time on health. It is particularly interesting to study this relationship, as it gives as useful insights about the speed of health deterioration in the absence of medical treatments. In addition, this result has an external sound validity, as it is not affected by the medical treatment provided to the patient by the ambulance crew and that differs across regions or countries, confounding the effect resulting from a quick response with the effect of the medical treatment provided. Finally, information on the health condition of patient at the time of admission to the hospital – and in particular the mortality rate – give us insights about the persistency over time of the effect of ambulance response on patient's condition.

The analysis considers the ambulance runs to patients affected by cardiovascular problems. This pathology is commonly investigated in the medical literature, being the primary cause of death in Europe and in the United States;[4] it is sensitive to timely treatments;[5] the symptoms are standard and a trained professional can easily identify the nature of the problem even during the call, before visiting the patient, so that the likelihood of misjudgments is small.[6]

The estimated effect of response time on patient's severity is negative, large and strongly significant. I show that one minute increase in the response time (3.6% change at the average) increases the probability of highly severe condition by the time the ambulance crew reaches the patient's side by 1.5 percentage points (over 3% at the average). In addition, one minute increase in the response time

[4]For additional information about the intake of cardiovascular problems in the population and a discussion about the related morbidity and mortality rates, see the report published by the World Health Organization: Mendis et al. (2011). In line with other countries, a large part of all emergency calls in Liguria are due to cardiovascular problems, accounting for one third of all non-deferrable ambulance missions and for 70% of all death before the patient is visited at the hospital.

[5]The introduction of response standards was first discussed during the Conference on Cardiopulmonary Resuscitation and Emergency Cardiovascular Care. The issue of the related *International Guidelines 2000* strongly influenced the design of the EMS services. For additional information, visit: https://doi.org/10.1161/01.CIR.102.suppl_1.I-1.

[6]In Liguria, over 93% of the diagnosis of cardiovascular problems performed by the responder of the emergency call, are confirmed once the patient is actually visited.

increases the probability that the patient dies by the time she is admitted to the hospital by 0.7 percentage points (17.5% at the average). This results show that the effect of ambulance response time on patient conditions is immediately visible and sizable, and that the effect persists over time. (Wilde, 2013) reports a similar effect on 90 days mortality; Avdic (2016) shows that most of deaths in the EMS setting takes place in the short term, by the time of admission to the hospital.

In the last section of the paper, I carry out a cost-benefit analysis of two alternative policies designed to reduce response times. I find that adopting a technology that supports the location of patients would reduce the average response by up to 30% at an almost zero cost. I also compute the maximum cost at which the policy of reducing ambulance response times would be cost-effective.

The rest of the paper is organized as follows. Section 2 describes the framework and the data. In section 3 I describe the empirical methodology. I present the main results in section 4, where I also explore the robustness of the results across model specifications. Finally, on section 6 I discuss the policy implication of the results and I conclude.

2 Data and Framework

2.1 Institutional setting

Emergency medical calls in Liguria are received by centralized contact centers, where nurses specifically trained to manage emergency calls establish the first contact with the caller and collect information about the event in order to assess the urgency of the situation. They assess the presumed pathology and the degree of severity by asking a predetermined set of questions specifically designed with the intent of maximizing the quality of information collected during the emergency call. The procedure adopted is prescribed by the influential text manual *Dispatch*, which is adopted internationally by all developed emergency medical systems. During the call, the nurse obtains information about the symptoms and categorizes the presumed pathology according to a class out of the 17 available, and the degree severity. These information are reported in a form, which is shared with the ambulance crew that performs the mission; after the ambulance dispatch, the crew communicates with the nurse also via radio in the case they need additional information or support to finding the patient. The nurse observes the list of available ambulances and their location, and dispatches the closest one. When the ambulance crew (which always includes a trained professional onboard) reaches the scene and locates the patient communicate this information to the contact center; they assess the pathology of the patient and the degree of severity; they provide first aid; they transport the patient to the closest adequate hospital to provide the

care the patient needs. The choice of the hospital in which to transport the patient is undertaken with the support of the nurse in the contact center, which verifies which hospital is capable to receive the patient in that moment. The service is provided free of charge to all citizens, and the patient is asked to pay a 25 euros ticket in the case in which the event was not an emergency.[7]

2.2 Performance measure and outcomes

The most standard performance measure adopted in the emergency framework is the response time, that is the time required to reach the person in need once it is requested. In 2012 Liguria introduced an innovative informative system to collect higher quality data in the Emergency Medical System (EMS). Thanks to this innovation, the administrative data adopted in this analysis are collected in real time (at the spot) and not self reported at the end of the mission by the ambulance crew as in similar data collection.[8] As such, the quality of data is high and, in particular, information about the timing of the mission is extracted by the automated record of the moment in which started the emergency call, and the moment in which the ambulance crew reaches the scene of the event. The severity of the patient condition is recorded in three point in time: during the emergency call, by the nurse that answered; at the arrival on the scene before attempting medical treatments, by the trained professional that is part of the ambulance team; at the arrival at the hospital, before being admitted to the hospital.

2.3 Cardiovascular problems

Cardiovascular problems are the main pathology class considered by the literature that consider the emergency medical setting (see for example Pons et al. (2005); Wilde (2013); Avdic (2016)).[9]This interest is due to at least three factors. First, a large share of the population is affected by heart problems, which represent the first cause of death in developed countries (for a deeper discussion see for example the report of the World Health Organization, *Global atlas on cardiovascular disease prevention and control* Mendis et al. (2011)). Second, this type of problems are usually sensitive to timely treatment, so that they are of particular interest in the emergency medical framework. Rapid response time by the ambulance is of key importance, and so, given also the spread of this type of problem, this is at the

[7]Some categories, namely pregnant women, kids under the age of 14, disabled people, and those with lower income, are exempted from the payment of the ticket.

[8]As for example in (Wilde, 2013)

[9]See also Nichol et al. (1996) for a review of less recent literature.

origin of the introduction of international response standards for the ambulance.[10] Third, this pathology is characterized by a set of symptoms that are relatively easy to recognize. For this reason, the identification of the pathology and the severity is performed with little margin of error yet during the emergency call. As such, there is a limited likelihood of misjudgment, as opposed to other class of problems that are not as easy to classify, for example the severity of the conditions of people involved in a car accident. In this way, the measures of patient severity are obtained in a setting in which it is, on average, non-misleading identifying the degree of health impairment and the priority assigned to patient to minimize the likelihood of further worsening of patient's condition.

2.4 Descriptive statistics

To perform the analysis in this paper I make use of administrative data that collects data for 30,149 ambulance missions to rescue patients affected by cardiovascular problems in the Italian region Liguria in 2013-2014. The use of these data was previously authorized under a data use agreement with the health regional authority.[11] The data collection in Liguria is supported by a management system, and the information are recorded at the spot, during the mission. Thanks to this unique technology introduced in the region in analysis, data are precise and more accurate than in most similar data collections, where the information is reported by the ambulance crew at the end of the mission, at the arrival at the hospital. Each record includes information about the pathology, age and gender of the patient, date and time when the call started and when the ambulance crew reaches the patient, the type of vehicle that performed the mission (advanced vs basic life support), the severity of the patient at the ambulance arrival on the scene – before providing medical treatments – and at the end of the mission – at the moment of admission to the hospital, and the municipality in which the patient is located. To generate an exogenous variation in the ambulance response time, I adopt the amount of rainfall during the hour and in the municipality in which the emergency mission is performed.

The ambulance response time (RT) is our main regressor of interest and is the performance measure usually adopted by the emergency medical system. RT is the amount of minutes from the moment in which the emergency call started, to

[10] See the *International Guidelines 2000* publicized after the Conference on Cardiopulmonary Resuscitation and Emergency Cardiovascular Care by the American Heart Association and similar recommendations by the European Resuscitation Council (Association et al., 1992; Nolan et al., 2010).

[11] In Italy, as in most European countries, the health care system is organized at the regional level. Along the regional dimension, the characteristics of the service are homogeneous.

the moment in which the ambulance arrives on the scene.[12] The distribution of RT is left skewed, as illustrated in Figure 4. The average RT is 28 minutes and the median is 25.6, as reported in Table 1.

[Tables 12 and 1 about here]

The data on rainfall are provided by the Regional Agency for the Environment of Liguria (ARPAL). The hourly data are collected by 213 land-based weather stations (illustrated by red dots in Figure 3), and each station covers an average area of 20 km^2 (12.4 miles^2). I interpolate the hourly records and I calculate the average rainfall at the municipality level by making use of a Geographic Information System (GIS) software. The municipality is the smallest administrative unit within the region, and the average dimension is about 30 km^2 (18.6 miles^2), accounting for 242 units.[13] The amount of precipitations is measured in term of hourly millimetres (1 mm = 0.04 in), and the average rainfall is about 1.4 mm. About 14% of the missions are performed during a rainfall.

The outcomes of interest are the degree of severity observed in two points in time: at the ambulance arrival on the scene (H1) and when the mission is concluded, at the arrival at the hospital (H2). The degree of severity is ranked accordingly to 4 alternative values, from 1 to 4, greater for higher levels of severity. A degree of 4 indicates a severe health impairment and that the patient is in imminent danger of life. H2 assumes also a value of 5 if the patient dies out-of-hospital. An interesting feature of H1 is that it is recorded before performing the medical treatment: this is particularly useful as it offers the possibility of quantifying the direct effect of RT on observed severity, H1, being so a direct measure of the sensitivity of cardiovascular problems to time. Table 1 reports the empirical distribution of H1 and H2 in the sample. The main outcomes adopted in the analysis, M1 and M2, are obtained by recoding H1 and H2. M1 is coded as a dummy equal 1 when H1 assumes the maximum degree of severity, zero otherwise.

[12] Also Blackwell and Kaufman (2002); Pons and Markovchick (2002); Swor and Cone (2002); Pons et al. (2005); Hollenberg et al. (2009) adopted this measure of performance. Other works use only the driving time of the ambulance (from the ambulance departure to the arrival on the scene), without taking into account the time required to manage the call and activate the ambulance mission. This is usually due to data limitations, as in Wilde (2013). The measure of performance adopted here considers the total time required to reach the patient once it is requested, so from the beginning of the call, as this is the best approximation we have on the amount of time that elapses from the health shock to first-aid, a period of time in which the health condition of the patient deteriorates quickly.

[13] In Liguria there are 233 municipalities with an extension of about 30 km^2. The municipality of Genova, the main city of the region, is instead 240 km^2 large. To make its geographical extension comparable with the other municipalities, I split it in its 9 neighbourhoods, each of which has an extension of about 27 km^2. Finally, as suggested by Agrillo and Bonati (2013), the hourly rainfall is interpolated adopting an inverse distance weighting ratio.

M2 is a dummy that equals 1 when the patient dies out-of-hospital, zero otherwise. The incidence of maximum severity degrees, M1, is 45%, while the incidence of out-of-hospital mortality, M2, is 4%. The large shares of observations characterized by highly severe (M1) or deaths (M2) are in line with the severity and mortality of cardiovascular pathology discussed by the literature. The use of M1 and M2 as outcomes mitigates concerns related to potential misjudgments because there is little ambiguity associated with these conditions – e.g. absence of vital functions. I test the sensitivity of my results to the way the outcome is coded, in Section 4.2.

The estimate results include four sets of controls: mission characteristics, time controls, patient's demographics and location characteristics. Mission characteristics comprise a dummy variable for ambulance high priority dispatches, fixed effect for the contact center that managed the call (there are 5 central operation centers), a dummy when the mission is performed by an Advance Life Support (ALS) vehicle, the distance travelled and its square. Time controls include year and day of the week fixed effects, and a dummy for holidays. Patients demographics include a dummy for male and the age category of patient (i.e., between 50 and 79 years or above 79 years, with the excluded category being the ones younger than 50). Finally, location characteristics include population density at the municipality level and municipality fixed effects to capture for other observable and unobservable characteristics at the municipality level that might have an effect on the outcome.[14] As reported by Table 1, 92% of missions are high priority dispatches and over 18% are performed by ALS vehicles, and the average distance travelled is 20 kilometers. The sample is balanced between males and females, the average age is 70 and about 50% of the individuals in sample are between 50 and 79 years old. Almost half of the missions are performed in highly dense population areas. Population density classifies the municipalities in the sample as highly, medium, or low densely inhabited. This statistic is provided by the Italian National Institute of Statistic, ISTAT, and is merged to the dataset using an unique identifier for municipality as key. Additional detail on the sample construction is discussed and illustrated in Table 12 in the appendix.

3 Empirical methodology

In the standard econometric framework in the literature, the health outcome (M) is modelled as a function of response time (RT). To ease of discussion, as in Wilde (2013), I adopt a linear probability model. I test the sensitivity of the results to functional form choices in Section 4, where I estimate a probit and an ordered

[14] By including municipality fixed effects, the source of variation exploited to obtain the first stage estimates, where rainfall is adopted as instrument for RT, comes by the variations in the rainfall amount at the *hourly* and date level.

probit model.

The simple OLS model is:

$$M_{ipt} = \alpha_0 + \mathbf{X}_{ipt}\alpha_1 + \alpha_2 RT_{ipt} + \epsilon_{ipt}, \tag{1}$$

where M is the health outcome for individual i located in the municipality p and that called an ambulance on the date and hour t. M is a dummy variable that indicates, alternatively, the degree of severity observed in the short term (M1) and the mortality rate (M2) of patients.[15] $\mathbf{X}$ is the vector of mission, time, individual and location controls described in previous section. There are several potential issues to empirically isolate the effect of RT on M by making use of the specification reported by Equation 1. Indeed, as discussed by Wilde (2013), there may be some information unobserved by the researcher but observed by the ambulance driver that affects the behavior of the latter, resulting in shorter responses for most critical cases. The OLS estimator may lead to biased estimates due to reverse causality.[16] To address this problem of non-random selection, I adopt an instrumental variable identification strategy, where I exploit the effect of hourly variation of rainfall amount at the municipality level (Z) on RT. I complement Equation 1 above by estimating the so-called first-stage regression described by Equation 2:

$$RT_{ipt} = \beta_0 + \mathbf{X}_{ipt}\beta_1 + \beta_2 Z_{pt} + v_{ipt}. \tag{2}$$

The standard errors, v, are clustered at the hour, date and municipality level –the level at which the instrument, Z, varies – as suggested in Abadie et al. (2017).[17]

To be a valid instrument, Z must satisfy two conditions: relevance and exclusion restriction. The first one requires the instrument to have a nonzero correlation with the endogenous variable RT. Relevance is tested by the first stage F-statistic which, accordingly to the rule of thumb proposed by Stock and Yogo (2005), must

[15] In Section 4.2 I also show the results for alternative codifications of M1 and M2.

[16] The error term, ϵ, is the sum of an idiosyncratic component – a part of the severity that is not observed by the ambulance driver and neither by the researcher – and by a measure of severity that is communicated to the ambulance driver but not observed by the researcher and that influences the behavior of the driver. As long as this is negatively correlated with the response time – e.g., shorter RT for more life-threatening cases – the OLS estimates are downward biased.

[17] In particular, Abadie et al. (2017) discuss the harm in clustering at a too aggregate level, arguing that this can lead to standard errors that are unnecessarily conservative, even in large samples. They discuss two issues: motivation for the adjustment and the appropriate level of clustering. According to Abadie et al. (2017), I choose the appropriate level of clustering by recognising that, in my setting, the assignment to the treatment (Z) is clustered at the hour, date, and municipality level. In this case clustering is justified and the level is the one of Z: hour, date, and municipality.

exceed 10. Here the F-statistic is 15.9. During a rainfall RT is greater because the ambulance drives safely at a lower speed and because there is an higher probability of traffic congestion that might have a further negative effect on the ambulance speed.[18] Even if the weather could be forecasted, its effects could not be totally mitigated. For this reason, weather shocks are largely adopted in the literature as instrumental variables for endogenous regressors (see Dell et al., 2014 for a review). Rainfall has a large and statistically significant effect on the duration of the first phase of the mission – i.e. RT – and a small and not significant effect on the second phase – i.e., on the way back, when the patient is transported to the hospital. The estimate results are reported on Table 2. Rainfall does not have a relevant effect on the time to go back because during this phase of the mission the ambulance drives slowly in any case, because there is the patient onboard. For this reason, the rainfall has a limited effect on further decreasing the speed of a vehicle that is already driving slowly. The practice of driving slowly on the way back is prescribed by the guidelines that discipline the activity of care providers, in order to limit the adverse consequences that abrupt breaking and accelerations might have on patient's condition.[19] Given the low precision of the estimates of the effect of rain on the way back, however, it still could be that part of the effect on the severity observed at the end of the mission (M2) comes from slower times to the hospital.

The second condition for the instrument validity, the exclusion restriction, requires that Z affects the outcome M only through its effect on RT. In the setting in analysis is likely that this condition is satisfied because of the granularity of the data. As such, is unlikely that variations in the rainfall amount at the hourly and municipality level have a direct effect on the severity of cardiovascular problems. In addition, the analysis includes a rich set of covariates to control for characteristics of the specific ambulance mission, time and place fixed effects, and patient demographics. Finally, influential medical literature that studied the relationship between the severity of cardiovascular problems and weather conditions, mitigates concerns about the existence of such link.[20]

Another important assumption for the internal validity of the my identification strategy is monotonicity, which requires that rainfall has a non-negative effect on

[18] See of the National Academies (2000) and Agarwal et al. (2005) for a detailed discussion about the effect of rain on road conditions and traffic. See **?** for a discussion on the effect of weather conditions specifically on ambulance performance.

[19] See the guidelines that discipline the activity of the EMS in Liguria: della Liguria (2013).

[20] Phillips et al. (2004) show that the apparent effect of weather conditions on the morbidity and mortality rate from cardiovascular problems disappears after controlling for holidays. Indeed, many holidays falls during the winter time, a time of the year in which also adverse weather conditions are more likely. The behavioral change of individuals during holidays, instead, affects the insurgence of cardiovascular problems. In my analysis I control for holidays.

RT. It is easy to believe that this condition is fulfilled, and I also provide evidence in support of this following the approach adopted by Angrist and Imbens (1995) and drawing the cumulative distribution functions of RT with and without rainfalls; these results are futher discussed at the end of Section 4.

4 Results

Tables 3 and 4 report the results of the baseline specification. Column (1) shows the estimates for the simple OLS model, column (2) reports the first stage estimates (FS), column (3) illustrates the intention to treat – the reduced form – estimates (ITT), and finally column (4) reports the second stage least square estimates (IV). All the estimates are obtained by including the full sets of controls presented in Section 2. All the results report clustered standard errors at the hour, date, and municipality level.

[Tables 3 and 4 about here]

Table 3 reports the results for the outcome M1, that refers to the health condition recorded at the ambulance arrival on the scene before performing the medical treatment. Table 4 shows the estimated effects that refers to the outcome M2, the mortality rate observed by the end of the ambulance mission. The first two rows in tables report the estimates of the parameters of interest: response time (RT) and rainfall (Z).

Column (1) of Table 3 reports the correlation between RT and the health condition of the patient, $M1$, and shows that a minute increase in RT increases the probability of $M1$ by 0.9 percentage points – an elasticity at the mean of 0.5. As discussed in Section 3, the OLS estimates may be affected by the unobserved behavior of the ambulance driver which reaches faster the patients in worse conditions, biasing downward the estimates. The FS estimates are reported in column (2). The instrument is relevant (the FS F-statistic is 16.10): each millimeter of rainfall increases RT by 0.34 minutes. Column (3) reports the reduced form estimates on the association between rainfall and the outcome. During rainfalls is more likely that a call results in severe health condition ($M1$) by about 0.5 percentage points: 1% at the sample average. The magnitude of the effect is proportional to the effect of the instrument on RT – column (2). Column (4) reports the results for the instrumented RT, and shows that a minute increase in RT increases $M1$, on average, by 1.5 percentage points. The elasticity of the results at the sample average is 0.9. The effect is about 3% both at the sample average and in terms of standard deviation. The average effect obtained with the 2SLS estimation strategy is about two times larger the effect estimated by OLS.

Column (1) in Table 4 reports the estimates of the correlation between RT and $M2$, suggesting that a minute increase in RT increases the probability of dying by 0.3 percentage points – an elasticity at the mean of 2. The FS estimates are exactly the same of the ones discussed above and reported also in Table 3. Column (3) reports the reduced form estimates on the association between rainfall and the outcome. Is more likely that a call results in death by about 0.3 percentage points per millimeter of rainfall. Column (4) reports the two stage least square estimates (2SLS): one minute increase in RT increases $M2$ by 0.7 percentage points. The elasticity of the results at the sample average is 5. The effect is about 17.5% at the sample average and 3.5% in terms of standard deviation. As for $M1$, also for $M2$ the effect obtained with the 2SLS estimates is over two times larger.

4.1 Sensitivity of results to included covariates and balancing of covariates

I test the sensitivity of the results with respect to the included covariates. The estimate results referred to the outcomes M1 and M2 are reported, respectively, in Tables 5 and 6. The first column in both tables reports the baseline results; columns (2)-(5) report the results obtained by gradually excluding sets of covariates. The results reported in column (2) are obtained by excluding the set of controls for mission characteristics, in column (3) I also exclude time fixed effects, in column (4) I exclude the demographics of the patient, in column (5) I exclude fixed effects about the characteristics of the location in which the ambulance mission in performed. The point estimate of the effect of RT on the outcomes of interest does not change substantially by changing the set of controls included in the regression and the first stage f-statistic is always above the critical threshold of 10. By excluding controls, the effect of RT on $M2$ loses some statistical power – especially to what concerns the unconditional specification in column (5) – but the point estimate is similar across specifications as one would expect when the instrumental variable strategy is internally valid.

[Tables 5 and 6 about here]

Table 7 reports the balancing of the covariates. Column (1) reports means and standard deviations of controls in the absence of rainfall and column (2) reports means and standard deviations of controls in the presence of rainfall. Columns (3) and (4) illustrate, respectively, the difference of the means and the p-value of the difference. As expected, we observe that the difference in the average response time is statistically significant. Other significant differences are related to the characteristics of the location in which the mission of the ambulance is performed: is more likely that a mission takes place in urban and in rural areas when it rains,

and it is less likely that this happens in medium densely inhabited areas. There is also a difference in which contact center manages the emergency call when it rains. These differences are driven by geographical factors, probably because in some areas it is more likely that it rain than in others. As such, I obtain my regression results including also fixed effects for 142 municipalities. In line with this, we do not observe instead any difference with respect to the demographics of patient – and this provides further support in favor of the fulfillment of the exclusion restrictions for the validity of instrument. Also the distance driven by the ambulance and the type of ambulance that performs the mission – being and advance life support type of vehicle or not – do not appear to be affected by rain.

[Table 7 about here]

4.2 Alternative Specifications

The estimates reported in Tables 8 and 9 show how the results change by changing the outcomes specification. As for the main results, the estimates are obtained by including the full set of controls and clustered standard errors at the hour, day, and municipality level. I also show the results by adopting an ordered probit model instead of a linear probability model.

Column (1) of Table 8 reports the baseline results for the effect of RT on M1, where M1 is a dummy equal 1 when the degree of severity observed at the arrival on the scene, H1, is equal 4. Column (2) shows the effect of RT on the probability of H1 to be equal 3 or 4. To what concerns cardiovascular problems, patients are classified as 3 or 4 over 92% of times.[21] Given the little variability in the outcome, it is difficult to identify the marginal effect of RT. Indeed, the magnitude of the estimated effect is small and statistically non distinguishable from zero. In the estimates of column (3), H1 is the outcome and so it assumes alternatively 4 values, from 1 to 4: one minute increase in RT increases H1 by 0.017 points. Finally, in the last two columns I report the estimates obtained by making use of an ordered probit model. The parameters sum to zero, so I report two out of three of them.[22] The marginal effect that results from the ordered probit estimates (column 5) is 1.6 percentage point. The effect is very similar to

[21] Patients affected by cardiovascular problem often present high severity conditions. For this reason, and because of the spread of this type of disease, this class of problems is adopted as target in the configuration of emergency services.

[22] To estimate the ordered probit, I combined together the severities of degree 1 and 2 because of the low numerosity of 1. The included controls are the fixed effects for priority dispatch, type of rescue vehicle, distance driven and the square, patient gender, age category, and fixed effect for day of the week. The standard errors are clustered at the usual level. The same applies also to the estimates reported in Table 9

the baseline results reported in the first column.[23]

[Table 8 about here]

Table 9 reports the results for the effect of RT on the condition of the patient at the end of the ambulance mission (H2). H2 assumes values from 1 to 4 for growing degrees of severity, and is equal 5 when the patient dies out-of-hospital. The baseline results, reported in column (1), are obtained by coding the outcome as a dummy equal 1 when H2 is 5. Column (2) reports the results for the outcome coded as a dummy equal 1 when H2 assumes values 4 or 5. In column (3) the outcome is a dummy equal 1 when H2 is 3, 4, or 5. The magnitude of the effect reported in column (2) is larger, suggesting that, coherently with previous discussion, RT affects negatively not only the likelihood of dying but also of being in extremely severe conditions. Column (4) shows the results when H2 is kept in levels: a minute increase in RT increases H2 by 0.036 points.[24] Finally, columns (5)-(7) report the marginal effects associated to H2 equal 3, 4, and 5 respectively, obtained by estimating the ordered probit model. The parameters sum to zero, so I report three out of the four of them. A minute increase in RT increases the probability of most severe conditions (H2 = 4, column 6) by 1 percentage point and increases the probability of dying (H2 = 5, column 7) by 0.6 percentage points. As before, the ordered probit estimates are very similar to the baseline model.

[Table 9 about here]

4.3 Internal Validity: Monotonicity

The interpretation of the estimates illustrated before as local average treatment effects (LATEs) requires monotonicity, that is stochastic dominance of response time distributions with respect to rain. To perform this test I follow the approach adopted by Angrist and Imbens (1995). The left panel of Figure 5 shows the conditional distribution functions (CDF) of response times with and without rainfall, and the right panel illustrates the difference between the two CDF. As shown in figure, stochastic dominance is met as the two CDF never cross. The difference

[23] The effect reported in column (4), and that refers to severity of degree 3, is -1.5. It almost cancels out the effect reported in column (5) indicating that, coherently with the result in column (2), greater RTs shift patients from degree of severity 3 to 4. The order probit is estimated at the sample average. The average values of relevant regressors – namely, the average response time, patient's age, and distance travelled by the ambulance – are reported in the bottom of the table.

[24] The magnitude of this effect is about double than in H1, suggesting that the negative effect of RT on health might increase over time, given that H2 is recorded later in time with respect to H1.

of the CDF shows the compliers, which are concentrated in the response times between 10 and 45 minutes.

[Figure 5 about here]

5 Cost-Benefit analysis

I now perform a cost-benefit analysis of two hypothetical policies: (1) implementing a system to improve the quality of the directions on patient location and provided to the ambulance driver when dispatched, and (2) buying one additional ambulance. In the discussion that follows, I devote the first subsection to quantify the benefit from reducing the response time. I value the benefit in terms of lives saved and I concentrate on the population in analysis – namely, patients affected by cardiovascular problems – ignoring other benefits such as the effect that a quicker response might have also on the other patients served by the ambulance. I discuss alternative measures of value of life proposed by the literature, and I indicate the most adequate in this setting. Then, I show the *indifference curve* of policy maker, i.e. the amount of resources that the policy maker is willing to invest to obtain a given reduction on the average response time of ambulance. The policy maker is happy to implement the policies that lie above the indifference curve (i.e., cost-effective policies); is indifferent whether to implement or not the policies on the curve; does not implement the policies with a cost-benefit ratio that locates under the indifference curve. In the second subsection I discuss the nature of the problem of patient location faced by the ambulance driver and I propose and estimate of its magnitude. The third subsection illustrates the cost of alternative policies: the ones aimed at improving patient location and of buying one additional ambulance. Here I also discuss the expected return – in terms of time saved – from these measures. Finally, I show how these alternative measures locates with respect to the indifference curve of the policy maker discussed above.

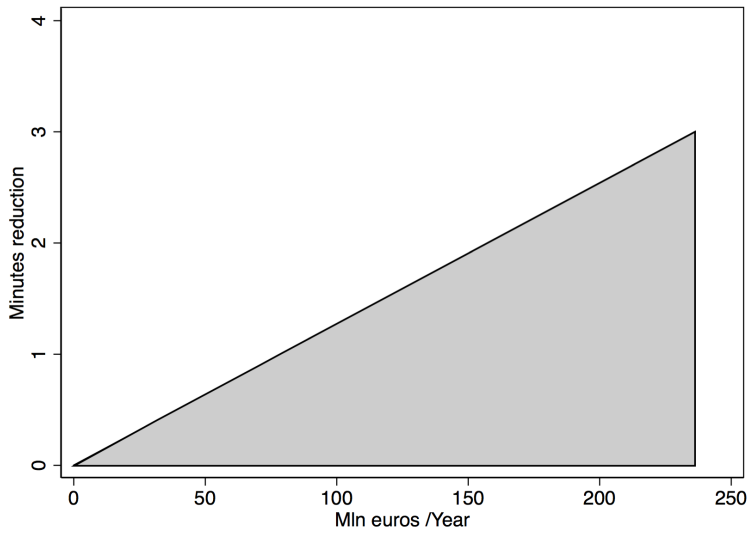
5.1 The value of life and the indifference curve of policy maker

Estimates of the value of a statistical life (VSL) developed by economists provides governments with a reference point for assessing the benefit of risk reduction efforts. Viscusi and Aldy (2003) provide a comprehensive review of the literature to reconcile the broad range of VSL estimates and discuss the application of VSL to public policy decisions. Viscusi (1994a) illustrates how to generate an estimate for the expenditure-induced fatality rate based on the value of a statistical life and the marginal propensity to spend on health (MgH):

$$\text{marginal expenditure per statistical life lost} = \frac{VSL}{MgH}.$$

This approach requires an estimate of VLS and an estimate of MgH. In this analysis I adopt the VLS proposed by Murphy and Topel (2006), as it allows me to estimate the value of one year of life – not the value of the entire life of an individual. The value of a life year is over 4 times the annual income; in my sample, over 75% of the individuals observed are older than 60, so I consider the median annual retirement income in Liguria in 2015 (i.e. 18,660 euros) as a measure of their annual income. Then, Viscusi (1994a; 1994b) proposes a measure of the marginal propensity to spend on health based on an analysis of 24 OECD countries: the resulting value ranges between 0.08 and 0.12. This implies that for every dollar increase in national income, an additional 8 to 12 cents are spent on health care. Assuming a marginal propensity to spend on health of 0.1 and a VSL of 74,640 euros (=18,660*4), the yearly marginal expenditure per statistical life lost is 746,400 euros. By multiplying the statistical value of one year of life by the effect that one minute increase of the ambulance response time has on the probability of dying (0.7 percentage points increase) and by the number of individuals affected by cardiovascular problems each year (30,149/2=15,074), we obtain the value of one year of life on the population of interest, i.e. 7.88 million euros. Then, after dividing this value by the marginal propensity to spend on health, we obtain the yearly willingness to pay to reduce the average response time when rescuing patients affected by cardiovascular problems in Liguria, i.e. 78.8 million euros per minute. In this analysis I do not consider other benefits, such as the effect that a more performing ambulance service would have also on the other patients rescue by ambulances. To ease of analysis, I consider the willingness to pay of policymaker for each minute saved linear. This relationship is illustrated in the following figure:

Figure 1: Indifference curve of policymaker: willingness to pay by minutes saved.



NOTES: the vertical axis reports the number of minutes saved, the horizontal axis shows the amount of investment per minute saved. The solid line illustrates the willingness to pay of policymaker per amount of time saved.

The solid line illustrates the minimum return in terms of EMS performance that the policymaker expects as a result of a given investment. Policies that lie above the solid line are preferred to the ones that lie on the solid line, while investments with a return that lies in the grey are would not be implemented.

5.2 The location of patient

During the emergency call important information is transmitted, including the directions to reach the site of the event. The panic or fear that the caller might be experiencing in that moment can substantially affect the quality of communication, resulting in unclear or inaccurate directions. Good quality directions is a fundamental piece of information for the ambulance driver in order to be able to locate quickly the patient. To quantify the magnitude of the location problem, I consider each ambulance mission includes two driving periods: the way to go – from the ambulance dispatch to the arrival on the scene – and the way to go back – from the departure from the scene to the arrival at the hospital. Each way driven by the ambulance is characterized by a different different of certainty about the location of the final destination: on the way to go, the ambulance driver must be able to locate the patient with the directions communicated during the emer-

gency call; on the way back, instead, the driver knows the location of the hospital. I identify the location problem by calculating the difference between the driving time to go and the driving time to go back at the single event level. Descriptive statistics about driving times and regression results are reported, respectively, in Tables 10 and 11. On average, the ambulance takes 15 minutes to go and 11.5 minutes to drive back. The estimates reported in Table 11 show that driving to the scene of an event takes 3.7 minutes more on the way to go and this measure is taken as an approximation of the location problem.

[Tables 10 and 11 about here]

I adopt the driving time to go back as a benchmark for the driving time to go in the absence of the location problem. However, this might lead to an underestimate of the real location problem for two reasons: (i) the maximum speed of the ambulance is lower on the way back because there is the patient onboard; (ii) if anything, the distance driven by the ambulance on the way back is longer because the patient is reached by the closest ambulance, but is not always transported to the closest hospital, as the choice of the hospital depends on the availability of space – this information is communicated to the driver before attempting transportation – and on the type of treatments that the patient needs.[25] Our approximation of the location problem is good because: (i) data on driving times are quite precise because are collected by a software that automatically records the moment in which the ambulance depart and arrive on the scene; (ii) the difference between driving times at the mission level clears from compounding factors that are fixed at the mission level, such as the characteristics of the driver, of the emergency vehicle, and of the patient.

5.3 The cost-effectiveness of improving location or buying one additional ambulance

The first policy I discuss is the one that concerns the acquisition of a new ambulance. Following Wilde (2009), Pons and Markovchick (2002), and Pons et al. (2005), I evaluate the cost of one additional staffed ambulance at 450,000 euros per year. To calculate the effect that one additional ambulance would have on the average response time of the Emergency medical System in Liguria, I consider that:

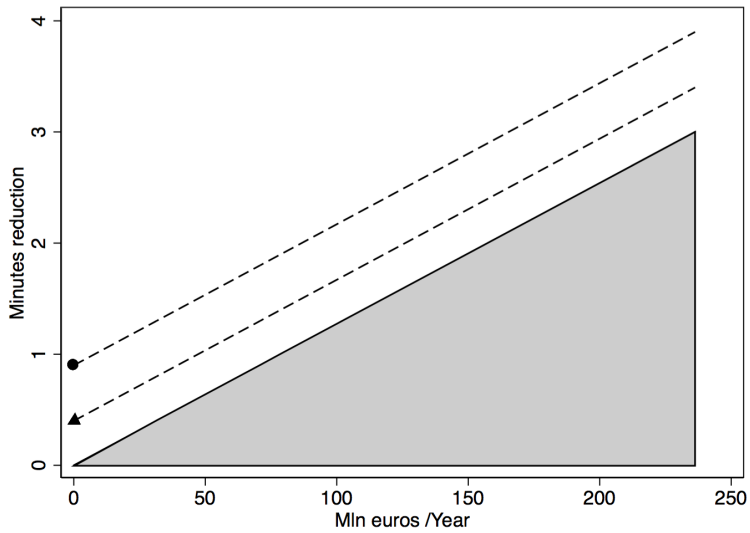
[25] The type of Emergency Medical System (EMS) adopted in Liguria, called the Franco-German model, prescribes the ambulance to reach the patient as fast as possible on order to provide first aid at the scene, before attempt transportation to the hospital. In addition, as discussed previously, the time required to reach the patient is the most important performance measure in the emergency medical system. For additional information, see the Liguria EMS guidelines: della Liguria (2013).

(i) on average, an ambulance in Liguria drives 10.4 kilometers to reach a patient; (ii) the average driving time from the ambulance dispatch to the arrival on the scene is 15.86 minutes; (iii) on average, each day there are 41 ambulance missions for cardiovascular problems in Liguria. One additional ambulance would reduce the average distance covered from 10.4 to 10.15 kilometers. An ambulance covers one kilometer in 1.5 minutes, so reducing the average distance by 0.25 kilometers would reduce the average response time by 0.4 minutes.

The second policy proposed is the introduction of a tool to improve the location of patients. As discussed above, the ambulance driver devotes about 3.7 minutes to locate the patient. The problem of location is not a new for service providers, and nowadays several instruments to support location are being tested around the world. Probably the most interesting for our setting is the availability of a software that have been recently developed and that can be installed in the call center that answers the emergency calls. This softwares automatically collects from the smartphone information about location (longitude, latitude and altitude) by making use of all the location technologies available in the phone (gps, wi-fi, and cells) and the information is share in automatic and it does not require an active cooperation from the caller. The adoption of this tool might take place at a virtually zero cost, and the only cost can be seen in training the emergency call responders to make use of such information. Given that this is an easy, user friendly tool, it is reasonable to estimate that the call responders can be trained with a total of 50 hours of training, and that each hour costs 50 euros, so that the cost of introducing this tool would be about 2,500 euros. If we assume that half of the calls will be provided with this additional information (the technology works only for the calls performed with a smartphone), and that this information will halved the *searching time* for the ambulance driver, the average response time would be reduced by 0.9 minutes.

The following figure compares the cost-effectiveness of the above mentioned policies. We can see that both policies seems to be good investments, as the expected return is higher than the critical threshold represented by the solid line. The triangle in figure illustrates the cost-benefit ratio from marginally increasing the number of available ambulances; the circle shows the cost-benefit ratio from improving patient location. The dashed lines illustrates the indifference curve associated to each policy: improving patient location would have a remarkable higher benefit.

Figure 2: Indifference curve of policymaker and two alternative policies.



NOTES: the vertical axes reports the number of minutes saved, the horizontal axes shows the amount of investment per minute saved. The solid line illustrates the willingness to pay curve of policymaker to obtain a given benefit. Circle: cost-benefit ratio from introducing an innovative location system; triangle: the cost-benefit ratio from buying one additional ambulance. Dashed lines: indifference curves of policymaker related to each policy.

6 Conclusions

This paper contributes to the related literature in several ways. The instrument I use, rainfall, has never been exploited – to the best of my knowledge – as shock for ambulance response. I identify the causal effect of RT on the severity and mortality of the patients affected by cardiovascular problems that called an ambulance. This class of diseases has strongly influenced the way in which the EMS services are designed because of their spread and sensitivity to timely care. I show that one minute increase in RT (3.6% change at the average) increases the probability that a patient affected by a cardiovascular problem is in highly severe condition by 1.5 percentage points (over 3% at the average). So the effect of ambulance delays on patient's conditions is immediately detectable and sizable. The effect persists over time, as there is also a 0.7 percentage points (17.5% at the average) increase on the probability of dying before reaching the hospital. The magnitude of the effect is similar to Avdic (2016), which studied the effect of changes in hospital proximity on the probability of dying after cardiovascular problems. By observing

the mortality rates also several months after the initial ambulance call, Avdic (2016) shows that most of mortality takes place out-of-hospital, coherently with the importance of timely treatments associated to this type of problems. Improving the access to EMS services would have sizable returns on health. Finally, I calculate the indifference curve of policymaker to study her willingness to pay for each given change in terms of EMS performance. I show that increasing the number of available ambulances and introducing an innovative system to improve the location of patient are both desirable policies, and the latter is the one that ensures the higher benefit.

References

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. (2017). When should you adjust standard errors for clustering? Technical report, National Bureau of Economic Research.

Agarwal, M., Maze, T. H., and Souleyrette, R. (2005). Impacts of weather on urban freeway traffic flow characteristics and facility capacity. *Proceedings of the 2005 mid-continent transportation research symposium*, pages 18–19.

Agrillo, G. and Bonati, V. (2013). Atlante climatico della liguria. *Agenzia Regionale per la Protezione dell'Ambiente Ligure, ARPAL*.

Angrist, J. D. and Imbens, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American statistical Association*, 90(430):431–442.

Association, A. H. et al. (1992). Guidelines for cardiopulmonary resuscitation emergency cardiac care. *Jama*, 268:2212–2302.

Avdic, D. (2016). Improving efficiency or impairing access? health care consolidation and quality of care: Evidence from emergency hospital closures in sweden. *Journal of health economics*, 48:44–60.

Blackwell, T. H. and Kaufman, J. S. (2002). Response time effectiveness: comparison of response time and survival in an urban emergency medical services system. *Academic Emergency Medicine*, 9(4):288–295.

Blanes i Vidal, J. and Kirchmaier, T. (2017). The effect of police response time on crime clearance rates. *The Review of Economic Studies*, 85(2):855–891.

Courtemanche, C., Friedson, A., Koller, A., and Rees, D. I. (2017). The affordable care act and ambulance response times.

Davis, K. (1956). The amazing decline of mortality in underdeveloped areas. *The American Economic Review*, 46(2):305–318.

Dell, M., Jones, B. F., and Olken, B. A. (2014). What do we learn from the weather? the new climate–economy literature. *Journal of Economic Literature*, 52(3):740–798.

della Liguria, D. I. S. ., editor (2013). *Guida pratica per i soccorritori delle Associazioni convenzionate per il soccorso con il Servizio 118 ligure*.

Doyle Jr, J. J., Graves, A, J., and Gruber, J. (2019). Evaluating measures of hospital quality. *Review of Economics and Statistics*.

Doyle Jr, J. J., Graves, J. A., Gruber, J., and Kleiner, S. A. (2015). Measuring returns to hospital care: Evidence from ambulance referral patterns. *Journal of Political Economy*, 123(1):170–214.

Fuchs, V. R. et al. (1996). Economics, values, and health care reform. *American Economic Review*, 86(1):1–24.

Hollenberg, J., Riva, G., Bohm, K., Nordberg, P., Larsen, R., Herlitz, J., Pettersson, H., Rosenqvist, M., and Svensson, L. (2009). Dual dispatch early defibrillation in out-of-hospital cardiac arrest: the salsa-pilot. *European heart journal*, 30(14):1781–1789.

Hull, P. (2018). Estimating hospital quality with quasi-experimental data.

Kloner, R. A. (2004). The "merry christmas coronary" and "happy new year heart attack" phenomenon.

Ludwig, G. G. (2004). Ems response time standards. *Emergency medical services*, 33(4):44.

Mendis, S., Puska, P., Norrving, B., et al. (2011). *Global atlas on cardiovascular disease prevention and control*. World Health Organization.

Mokyr, J. (1993). Technological progress and the decline of european mortality. *The American Economic Review*, 83(2):324–330.

Muller, J. E., Tofler, G., and Stone, P. (1989). Circadian variation and triggers of onset of acute cardiovascular disease. *Circulation*, 79(4):733–743.

Murphy, K. M. and Topel, R. H. (2006). The value of health and longevity. *Journal of political Economy*, 114(5):871–904.

Nichol, G., Detsky, A. S., Stiell, I. G., O'Rourke, K., Wells, G., and Laupacis, A. (1996). Effectiveness of emergency medical services for victims of out-of-hospital cardiac arrest: a metaanalysis. *Annals of emergency medicine*, 27(6):700–710.

Nolan, J. P., Soar, J., Zideman, D. A., Biarent, D., Bossaert, L. L., Deakin, C., Koster, R. W., Wyllie, J., and Böttiger, B. (2010). European resuscitation council guidelines for resuscitation 2010 section 1. executive summary. *Resuscitation*, 81(10):1219–1276.

of the National Academies, T. R. B., editor (2000). *Highway capacity manual*. Washington, DC.

O'Keeffe, C., Nicholl, J., Turner, J., and Goodacre, S. (2010). Role of ambulance response times in the survival of patients with out-of-hospital cardiac arrest. *Emergency medicine journal*, pages emj–2009.

Peltzman, S. (1973). An evaluation of consumer protection legislation: the 1962 drug amendments. *Journal of political economy*, 81(5):1049–1091.

Phillips, D. P., Jarvinen, J. R., Abramson, I. S., and Phillips, R. R. (2004). Cardiac mortality is higher around christmas and new year's than at any other time. *Circulation*, 110(25):3781–3788.

Pons, P. T., Haukoos, J. S., Bludworth, W., Cribley, T., Pons, K. A., and Markovchick, V. J. (2005). Paramedic response time: does it affect patient survival? *Academic Emergency Medicine*, 12(7):594–600.

Pons, P. T. and Markovchick, V. J. (2002). Eight minutes or less: does the ambulance response time guideline impact trauma patient outcome? *The Journal of emergency medicine*, 23(1):43–48.

Stock, J. H. and Yogo, M. (2005). Testing for weak instruments in linear iv regression. *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, page 80.

Swor, R. A. (1993). *Quality management in prehospital care*. Mosby.

Swor, R. A. and Cone, D. C. (2002). Emergency medical services advanced life support response times: lots of heat, little light. *Academic Emergency Medicine*, 9(4):320–321.

Viscusi, W. K. (1994a). Mortality effects of regulatory costs and policy evaluation criteria. *The RAND journal of Economics*, pages 94–109.

Viscusi, W. K. (1994b). Risk-risk analysis. *Journal of Risk and Uncertainty*, 8(1):5–17.

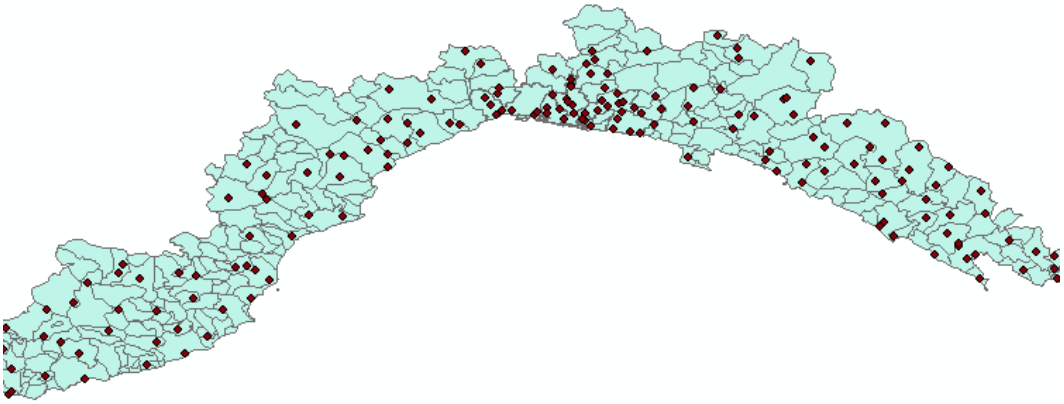
Viscusi, W. K. and Aldy, J. E. (2003). The value of a statistical life: a critical review of market estimates throughout the world. *Journal of risk and uncertainty*, 27(1):5–76.

Wilde, E. T. (2009). Do emergency medical system response times matter for health outcomes?

Wilde, E. T. (2013). Do emergency medical system response times matter for health outcomes? *Health economics*, 22(7):790–806.

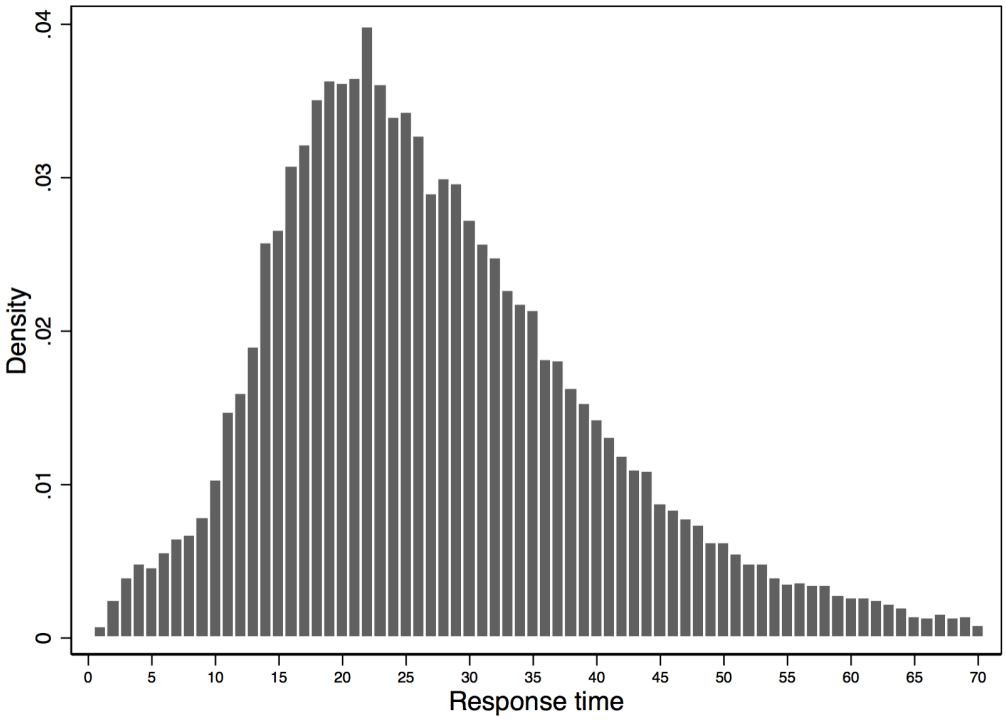
Figures and Tables

Figure 3: Municipality borders in Liguria and location of the weather stations.



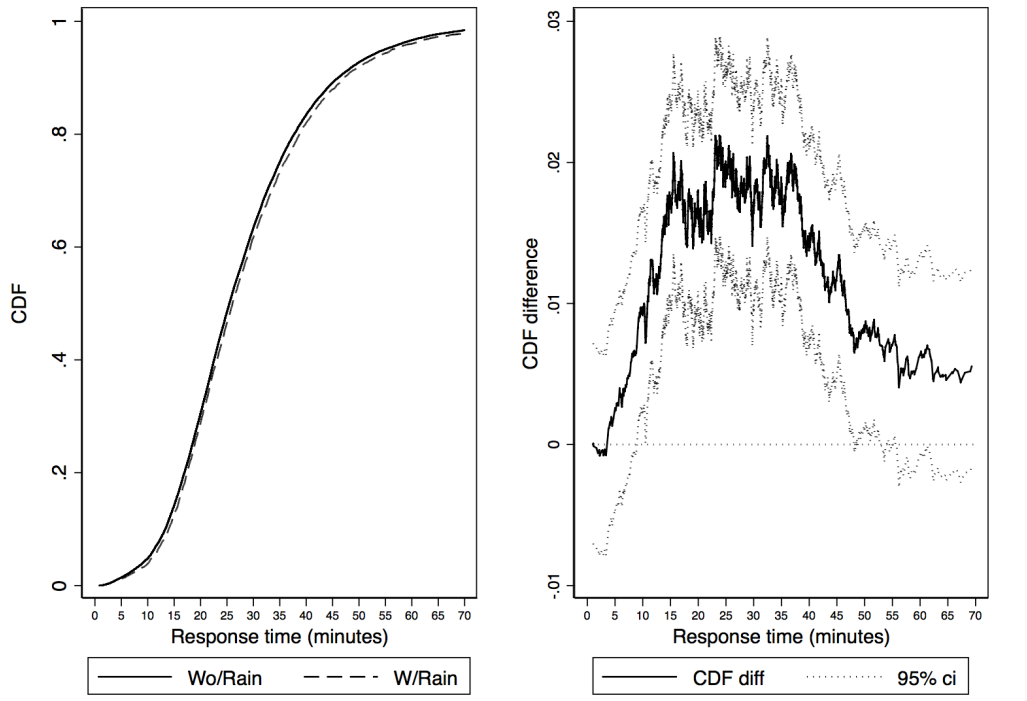
NOTES: the black lines are the borders of 233 + 9 municipalities in Liguria. The additional 9 municipalities are obtained by splitting the municipality of Genova in its 9 districts, in order to obtain a municipality extension similar across the region. The red dots indicate the location of 213 weather stations.

Figure 4: Response times distribution



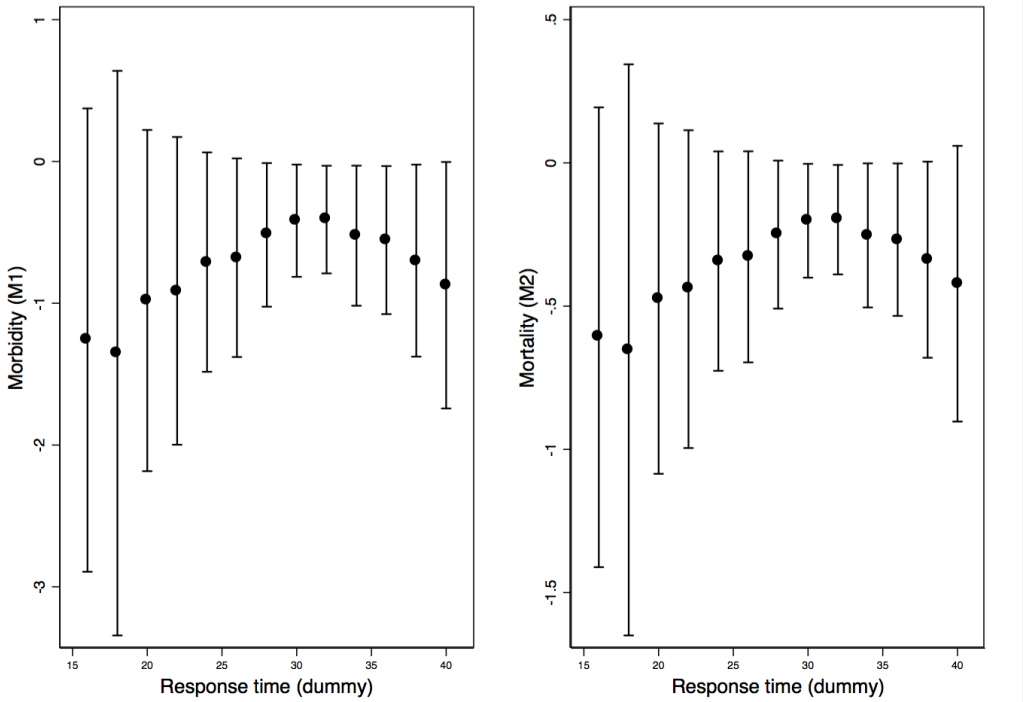
NOTES: empirical probability distribution function of response time for the sample in analysis. The response time is expressed in minutes, and is obtained as the difference between the time in which the ambulance reaches the patient and the time of the emergency call that initiated the mission.

Figure 5: Cumulative distribution function (CDF) and CDF differences of response times during rainfall or not.



NOTES: CDF of response times by weather conditions (in the presence of rainfall or not. Wo/ = without; W/ = with) for the sample in analysis. On the right panel, the difference between the two CDFs (the dotted lines are the 95% confidence intervals).

Figure 6: Plot of the estimates of the effect of response time dummies on Morbidity (M1) and Mortality (M2) rates. The response dummy is instrumented by hourly rainfall at the municipality level.



NOTES: the corresponding estimate results are reported in Table **??**. Response time is recoded as a dummy equal 1 for responses below a certain threshold, zero otherwise, and each dot is the regression results for each threshold in the interval 16-40 minutes. The panel on the left reports the point estimates, the panel on the right includes also the 95% confidence intervals. The figures on top refers to the effect of the response time being within a given threshold on morbidity rate (M1); the bottom figures refers to mortality rate (M2).

Table 1: Summary Statistics

Variables	Mean	Std. Dev.
H1: Level 1 (%)	0.03	1.62
H1: Level 2 (%)	7.48	26.30
H1: Level 3 (%)	47.38	49.93
H1: Level 4 (%) = Morbidity (M1)	45.11	49.76
H2: Level 1 (%)	0.34	5.84
H2: Level 2 (%)	14.38	35.09
H2: Level 3 (%)	65.09	47.67
H2: Level 4 (%)	16.14	36.79
H2: Level 5 (%) = Mortality (M2)	4.04	19.69
Response time	28.22	14.24
Rainfall (%)	14.10	34.80
Rainfall (mm)	1.43	2.50
High Priority Dispatch (%)	91.98	27.16
Patient age: 50-79 years (%)	49.74	50.00
Patient age: 80+ (%)	37.14	48.32
Patient Gender: Male (%)	49.82	50.00
Distance (km)	20.80	22.09
Type of Ambulance: Advance Life Support	18.68	38.97
Population density: High (%)	48.36	49.97
Population density: Medium (%)	43.99	49.64
Population density: Low (%)	7.64	26.57
Day of the week: Monday (%)	14.85	35.56
Day of the week: Tue ay (%)	14.70	35.41
Day of the week: Wedne ay (%)	14.12	34.82
Day of the week: Thur ay (%)	14.11	34.82
Day of the week: Friday (%)	14.58	35.29
Day of the week: Saturday (%)	13.67	34.36
Day of the week: Sunday (%)	13.96	34.66
Number of Observations	30,149	

NOTES: the summary statistics refer to all ambulance missions for cardiovascular problems performed in Liguria in 2 years. H1 is the health condition of patient observed at the arrival on scene, before the medical treatment. H2 is the health condition by the end of the mission, when the ambulance reaches back the hospital. The health condition is classified accordingly to 4 degrees of severity, where 4 is the worse and indicates patients in imminent risk of dying. 5 indicates that the patient is deceased.

Table 2: Effect of rainfall on out-of-hospital time: from the call to the ambulance arrival on the scene (i.e, the response time) and from the arrival on the scene to the hospital (i.e. the way back)

	Go	Back
	(1)	(2)
Rainfall (mm)	0.34***	0.12
	(0.09)	(0.12)
Mission characteristics	✓	✓
Time controls	✓	✓
Patient characteristics	✓	✓
Location characteristics	✓	✓
Average amount of rainfall: 1.44 mm		
Average out-of-hospital time	28	23
Observations	30149	30149

NOTES: the results are obtained by including the full set of covariates, namely mission characteristics (contact center that managed the call, ambulance priority dispatch, dummy for Advance Life Support (ALS) ambulances, amount of kilometers driven by the ambulance and its square), time fixed effects (weekday, holiday, year), individual demographics (gender and age category) and location fixed effects (municipality and population density). Column (1) illustrates the effect of rainfall on the response time (i.e., the time required to reach the patient). Column (2) shows the effect of rainfall on the time required to go back (i.e., to get the patient to the hospital). Clustered standard error at the hour, day and municipality level in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$

Table 3: Effect of ambulance response time on patient's morbidity rate (M1)

Dependent variable:	OLS (M1)	FS (RT)	ITT (M1)	IV (M1)
Response Time	0.009***			0.015**
	(0.0002)			(0.0073)
Rainfall (mm)		0.341***	0.005*	
		(0.0856)	(0.0027)	
High Priority Dispatch	0.383***	2.421***	0.405***	0.368***
	(0.0051)	(0.2616)	(0.0051)	(0.0186)
Type of Ambulance: ALS	0.238***	7.789***	0.309***	0.190***
	(0.0076)	(0.2465)	(0.0076)	(0.0576)
Distance (km)	-0.001**	0.170***	0.001***	-0.002
	(0.0002)	(0.0093)	(0.0002)	(0.0013)
Distance $(km)^2$	0.000***	-0.000***	0.000	0.000**
	(0.0000)	(0.0001)	(0.0000)	(0.0000)
Patient Gender: Male	0.078***	0.576***	0.084***	0.075***
	(0.0051)	(0.1539)	(0.0053)	(0.0067)
Patient Age: 50-79	0.056***	4.140***	0.094***	0.031
	(0.0077)	(0.2338)	(0.0080)	(0.0313)
Patient Age: 80+	0.015*	5.656***	0.066***	-0.020
	(0.0081)	(0.2390)	(0.0083)	(0.0421)
Day of the week: Tuesday	0.004	-0.239	0.001	0.005
	(0.0092)	(0.2835)	(0.0097)	(0.0094)
Day of the week: Wednesday	0.016*	-0.046	0.015	0.016*
	(0.0094)	(0.2895)	(0.0099)	(0.0095)
Day of the week: Thursday	0.019**	-0.084	0.018*	0.020**
	(0.0094)	(0.2827)	(0.0098)	(0.0095)
Day of the week: Friday	-0.007	-0.138	-0.008	-0.006
	(0.0093)	(0.2893)	(0.0097)	(0.0094)
Day of the week: Saturday	0.005	0.360	0.008	0.003
	(0.0094)	(0.2965)	(0.0099)	(0.0100)
Day of the week: Sunday	-0.007	-0.563	-0.011	-0.003
	(0.0180)	(0.5810)	(0.0186)	(0.0190)
Public Holiday	0.018	1.047**	0.027	0.011
	(0.0159)	(0.5193)	(0.0165)	(0.0185)
Year	-0.004	-0.685***	-0.010*	0.000
	(0.0051)	(0.1559)	(0.0053)	(0.0071)
Population density: Medium	-0.912***	-10.289***	-0.560***	-0.402***
	(0.2024)	(1.1275)	(0.0384)	(0.0848)
Population density: Low	-0.744***	-1.409*	-0.310***	-0.289***
	(0.2007)	(0.8365)	(0.0252)	(0.0277)
Contact Center 2 (%)	-0.028	12.497	0.086	-0.105
	(0.1706)	(17.4182)	(0.3224)	(0.1240)
Contact Center 3 (%)	0.165	-6.695	0.105	0.207
	(0.1593)	(5.6771)	(0.1932)	(0.1512)
Contact Center 4 (%)	-0.677***	7.713	-0.605***	-0.723***
	(0.1990)	(6.6581)	(0.2191)	(0.2023)
Contact Center 5 (%)	0.533***	5.611***	0.584***	0.498***
	(0.0197)	(0.6572)	(0.0203)	(0.0458)
Municipality FE	✓	✓	✓	✓
Observations	30149	30149	30149	30149
F statistic				15.92

NOTES: all results are obtained by including the full set of covariates, namely mission characteristics (contact center that managed the call, ambulance priority dispatch, dummy for Advance Life Support (ALS) ambulances, amount of kilometers driven by the ambulance and its square), time fixed effects (weekday, holiday, year), individual demographics (gender and age category) and location fixed effects (municipality and population density). The columns show, respectively, the simple ordinary least square estimates (OLS), the first stage (FS), the intention to treat (ITT) and the instrumented (IV) results. First stage F-statistic on the bottom of the table. Clustered standard errors at the hour, day and municipality level in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 4: Effect of ambulance response time on patient's mortality rate (M2)

Dependent variable:	OLS (M2)	FS (RT)	ITT (M2)	IV (M2)
Response Time	0.003***			0.007**
	(0.0001)			(0.0037)
Rainfall (mm)		0.341***	0.002**	
		(0.0856)	(0.0012)	
High Priority Dispatch	0.026***	2.421***	0.034***	0.016*
	(0.0019)	(0.2616)	(0.0019)	(0.0093)
Type of Ambulance: ALS	0.043***	7.789***	0.067***	0.010
	(0.0040)	(0.2465)	(0.0041)	(0.0290)
Distance (km)	-0.004***	0.170***	-0.003***	-0.004***
	(0.0002)	(0.0093)	(0.0002)	(0.0007)
Distance $(km)^2$	0.000***	-0.000***	0.000***	0.000***
	(0.0000)	(0.0001)	(0.0000)	(0.0000)
Patient Gender: Male	0.014***	0.576***	0.016***	0.012***
	(0.0022)	(0.1539)	(0.0022)	(0.0032)
Patient Age: 50-79	0.001	4.140***	0.014***	-0.016
	(0.0023)	(0.2338)	(0.0024)	(0.0155)
Patient Age: 80+	0.027***	5.656***	0.044***	0.003
	(0.0029)	(0.2390)	(0.0029)	(0.0210)
Day of the week: Tuesday	-0.004	-0.239	-0.005	-0.003
	(0.0039)	(0.2835)	(0.0040)	(0.0041)
Day of the week: Wednesday	-0.002	-0.046	-0.003	-0.002
	(0.0040)	(0.2895)	(0.0041)	(0.0042)
Day of the week: Thursday	-0.001	-0.084	-0.002	-0.001
	(0.0040)	(0.2827)	(0.0041)	(0.0042)
Day of the week: Friday	-0.001	-0.138	-0.001	-0.000
	(0.0039)	(0.2893)	(0.0040)	(0.0041)
Day of the week: Saturday	-0.003	0.360	-0.002	-0.004
	(0.0041)	(0.2965)	(0.0042)	(0.0045)
Day of the week: Sunday	-0.004	-0.563	-0.005	-0.001
	(0.0084)	(0.5810)	(0.0086)	(0.0092)
Public Holiday	0.002	1.047**	0.005	-0.003
	(0.0077)	(0.5193)	(0.0079)	(0.0092)
Year	-0.007***	-0.685***	-0.009***	-0.004
	(0.0022)	(0.1559)	(0.0022)	(0.0033)
Population density: Medium	0.048	-10.289***	0.155***	0.230***
	(0.1126)	(1.1275)	(0.0168)	(0.0421)
Population density: Low	-0.094	-1.409*	0.041***	0.051***
	(0.1121)	(0.8365)	(0.0119)	(0.0135)
Contact Center 2 (%)	-0.046	12.497	-0.008	-0.098
	(0.5595)	(17.4182)	(0.6125)	(0.4885)
Contact Center 3 (%)	-0.152	-6.695	-0.172	-0.123
	(0.1084)	(5.6771)	(0.1153)	(0.1058)
Contact Center 4 (%)	-0.240**	7.713	-0.216*	-0.272**
	(0.1113)	(6.6581)	(0.1165)	(0.1137)
Contact Center 5 (%)	-0.151***	5.611***	-0.134***	-0.175***
	(0.0093)	(0.6572)	(0.0094)	(0.0232)
Municipality FE	✓	✓	✓	✓
Observations	30149	30149	30149	30149
F statistic				15.92

NOTES: all results are obtained by including the full set of covariates, namely mission characteristics (contact center that managed the call, ambulance priority dispatch, dummy for Advance Life Support (ALS) ambulances, amount of kilometers driven by the ambulance and its square), time fixed effects (weekday, holiday, year), individual demographics (gender and age category) and location fixed effects (municipality and population density). M2 is the out-of-hospital patient's mortality rate. The reported results are, respectively, ordinary least square (OLS), first stage (FS), the intention to treat (ITT) and the instrumented (IV) estimations. First stage F-statistic on the bottom of the table. Clustered standard errors at the hour, day and municipality level in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 5: Effect of ambulance Response Time on patient's Morbidity (M1) by included covariates. Response Time is instrumented by rainfall amount at the hourly and municipality level.

Dependent variable: Morbidity at the Arrival on the Scene (M1)

	(1)	(2)	(3)	(4)	(5)
Response Time	0.015**	0.017**	0.017**	0.018**	0.018**
	(0.007)	(0.008)	(0.008)	(0.008)	(0.008)
Population density: Medium	-0.402***	0.016	0.020	0.026	
	(0.085)	(0.057)	(0.056)	(0.057)	
Population density: Low	-0.289***	0.337***	0.330***	0.244**	
	(0.028)	(0.084)	(0.084)	(0.099)	
Patient Gender: Male	0.075***	0.086***	0.085***		
	(0.007)	(0.010)	(0.010)		
Patient Age: 50-79	0.031	0.043	0.041		
	(0.031)	(0.036)	(0.035)		
Patient Age: 80+	-0.020	-0.009	-0.012		
	(0.042)	(0.046)	(0.045)		
Day of the week: Tuesday	0.005	0.004			
	(0.009)	(0.010)			
Day of the week: Wednesday	0.016*	0.019*			
	(0.010)	(0.010)			
Day of the week: Thursday	0.020**	0.023**			
	(0.010)	(0.010)			
Day of the week: Friday	-0.006	-0.002			
	(0.009)	(0.010)			
Day of the week: Saturday	0.003	0.008			
	(0.010)	(0.011)			
Day of the week: Sunday	-0.003	0.005			
	(0.019)	(0.020)			
Public Holiday	0.011	0.008			
	(0.018)	(0.019)			
Year	0.000	-0.003			
	(0.007)	(0.006)			
High Priority Dispatch	0.368***				
	(0.019)				
Type of Ambulance: ALS	0.190***				
	(0.058)				
Distance (km)	-0.002				
	(0.001)				
Distance (km)^2	0.000**				
	(0.000)				
Contact Center 2 (%)	-0.105				
	(0.124)				
Contact Center 3 (%)	0.207				
	(0.151)				
Contact Center 4 (%)	-0.723***				
	(0.202)				
Contact Center 5 (%)	0.498***				
	(0.046)				
Municipality FE	✓	✓	✓	✓	
Observations	30149	30149	30149	30149	30149
F statistic	15.92	14.24	14.91	13.34	13.27

NOTES: column (1) reports the results for the baseline specification, obtained by including controls for mission characteristics (contact center that managed the call, ambulance priority dispatch, dummy for Advance Life Support (ALS) ambulances, amount of kilometers driven by the ambulance and its square), time fixed effects (day of the week, holiday, year), individual demographics (gender and age category), and location fixed effects (municipality, population density). Columns (2)-(5) present the results by excluding, respectively, controls for mission characteristics, time fixed effects, individual demographics and location fixed effects. First stage F-statistic on the bottom of the table. Clustered standard errors at the hour, day and munici-

Table 6: Effect of ambulance Response Time on patient's Mortality (M2) by included covariates. Response Time is instrumented by rainfall amount at the hourly and municipality level.

Dependent variable: Out-of-hospital Mortality (M2)

	(1)	(2)	(3)	(4)	(5)
Response Time	0.007**	0.008**	0.007*	0.007*	0.006
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Population density: Medium	0.230***	0.028	0.024	0.024	
	(0.042)	(0.028)	(0.027)	(0.027)	
Population density: Low	0.051***	-0.150***	-0.153***	-0.144***	
	(0.013)	(0.044)	(0.043)	(0.051)	
Patient Gender: Male	0.012***	0.007	0.008		
	(0.003)	(0.005)	(0.005)		
Patient Age: 50-79	-0.016	-0.021	-0.019		
	(0.015)	(0.018)	(0.018)		
Patient Age: 80+	0.003	0.000	0.003		
	(0.021)	(0.024)	(0.023)		
Day of the week: Tuesday	-0.003	-0.004			
	(0.004)	(0.004)			
Day of the week: Wednesday	-0.002	-0.003			
	(0.004)	(0.004)			
Day of the week: Thursday	-0.001	-0.001			
	(0.004)	(0.004)			
Day of the week: Friday	-0.000	-0.000			
	(0.004)	(0.004)			
Day of the week: Saturday	-0.004	-0.004			
	(0.005)	(0.005)			
Day of the week: Sunday	-0.001	-0.002			
	(0.009)	(0.010)			
Public Holiday	-0.003	-0.003			
	(0.009)	(0.010)			
Year	-0.004	-0.013***			
	(0.003)	(0.003)			
High Priority Dispatch	0.016*				
	(0.009)				
Type of Ambulance: ALS	0.010				
	(0.029)				
Distance (km)	-0.004***				
	(0.001)				
Distance $(km)^2$	0.000***				
	(0.000)				
Contact Center 2 (%)	-0.098				
	(0.488)				
Contact Center 3 (%)	-0.123				
	(0.106)				
Contact Center 4 (%)	-0.272**				
	(0.114)				
Contact Center 5 (%)	-0.175***				
	(0.023)				
Municipality FE	✓	✓	✓	✓	
Observations	30149	30149	30149	30149	30149
F statistic	15.92	14.24	14.91	13.34	13.27

NOTES: column (1) reports the results for the baseline specification, obtained by including controls for mission characteristics (contact center that managed the call, ambulance priority dispatch, dummy for Advance Life Support (ALS) ambulances, amount of kilometers driven by the ambulance and its square), time fixed effects (day of the week, holiday, year), individual demographics (gender and age category), and location fixed effects (municipality, population density). Columns (2)-(5) present the results by excluding, respectively, controls for mission characteristics, time fixed effects, individual demographics and location fixed effects. First stage F-statistic on the bottom of the table. Clustered standard errors at the hour, day and municipality level in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 7: Balancing of covariates in the presence and in the absence of rainfall.

Variables	Wo/Rainfall (1)		W/Rainfall (2)		Difference (3)	Pvalue (4)	
	Mean	SD	Mean	SD			
Response Time	28.08	14.12	28.92	14.58	-0.83	0.000	***
High priority dispatch (%)	91.84	27.38	92.66	26.08	-0.82	0.067	*
Patient Gender: Male (%)	49.75	50.00	49.99	50.01	-0.24	0.769	
Patient Age: < 50 (%)	13.16	33.81	12.89	33.51	0.28	0.622	
Patient Age: 50-79 (%)	49.76	50.00	49.64	50.00	0.13	0.880	
Patient Age: 80+ (%)	37.08	48.30	37.48	48.41	-0.40	0.617	
Distance (km)	15.80	19.75	15.99	20.57	-0.19	0.571	
Type of ambulance: ALS (%)	18.74	39.02	18.58	38.90	0.16	0.803	
Population density: High (%)	48.26	49.97	50.65	50.00	-2.38	0.004	***
Population density: Medium (%)	44.43	49.69	40.94	49.18	3.49	0.000	***
Population density: Low (%)	7.31	26.03	8.42	27.77	-1.11	0.011	**
Day of the week: Monday (%)	14.48	35.20	16.81	37.40	-2.33	0.000	***
Day of the week: Tuesday (%)	14.76	35.47	14.51	35.22	0.26	0.663	
Day of the week: Wednesday (%)	14.12	34.82	14.13	34.84	-0.01	0.982	
Day of the week: Thursday (%)	14.20	34.90	13.83	34.52	0.37	0.517	
Day of the week: Friday (%)	15.08	35.79	11.47	31.87	3.61	0.000	***
Day of the week: Saturday (%)	13.39	34.05	15.35	36.05	-1.97	0.001	***
Day of the week: Sunday (%)	13.97	34.67	13.90	34.59	0.07	0.901	
Contact Center 1 (%)	39.19	48.82	42.56	49.45	-3.37	0.000	***
Contact Center 2 (%)	14.93	35.63	12.32	32.87	2.60	0.000	***
Contact Center 3 (%)	11.99	32.48	13.31	33.97	-1.32	0.015	**
Contact Center 4 (%)	10.09	30.12	9.85	29.80	0.23	0.637	
Contact Center 5 (%)	23.81	42.59	21.96	41.40	1.85	0.008	***
Observations	25,896		4,253				

NOTES: column (1) reports mean and standard deviation of covariates in the absence of rainfall; column (2) reports mean and standard deviation of covariates in the presence of rainfall; column (3) and (4) report, respectively, the difference of the means and the p-value of the difference. *** p<0.01, ** p<0.05, * p<0.1

Table 8: Alternative outcome specifications and ordered probit estimation results for the effect of response time on patient's conditions at the ambulance arrival on the scene.

Dependent variable: Morbidity at the Arrival on the Scene (M1)					
	(1)	(2)	(3)	(4)	(5)
Response Time (RT)	0.015**	0.002	0.017**	-0.015	0.016
	(0.007)	(0.002)	(0.008)	(.)	(.)
Observations	30149	30149	30149	30149	30149
F statistic	15.92	15.92	15.92		
Average RT				28	28
Average age				70	70
Average distance				16	16

NOTES: the estimate results reported in columns (1)-(3) are obtained by including the full set of controls, namely mission characteristics (fixed effect for the contact center that managed the call, ambulance priority dispatch, dummy for Advance Life Support (ALS) ambulances, distance driven by the ambulance and its square), time fixed effects (day of the week, holiday, year), individual demographics (gender and age category), and location fixed effects (municipality and population density). Column (1) reports the baseline results, where the outcome is a dummy equal 1 for maximum degree of severity (level 4), zero for levels 1, 2 or 3. Column (2) report the results where the outcome is coded as a dummy equal 1 for severity degrees 3 or 4, zero otherwise. Column (3) reports the results for the linear model, where the outcome ranges between 1 and 4 (greatest for higher degrees of severity). The results reported in columns (4) and (5) are the marginal effects obtained by estimating the ordered probit model; the set of covariates does not include fixed effects at the municipality level to allow convergence; lower severity degrees (1 and 2) are grouped together because of the numerosity of degree 1. The three parameters estimated by the ordered probit sum up to 1, so I report the results of two of them: the effect of RT on the severity of degree 3 – in column (4) – and on severity of degree 4 – in column (5). The bottom of the table reports the first stage F-statistics and the average values of RT, age and distance driven by the ambulance at which the maximum likelihood estimates are calculated. *** $p<0.01$, ** $p<0.05$, * $p<0.1$

Table 9: Alternative outcome specifications and ordered probit estimation results for the effect of response time on patient's conditions at the end of the ambulance mission, upon arrival at the hospital.

Dependent variable: Out-of-hospital Mortality (M2)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
RT	0.007**	0.018**	0.011**	0.052***	-0.002	0.010	0.006
	(0.004)	(0.007)	(0.005)	(0.011)	(.)	(.)	(.)
Obs	30149	30149	30149	30149	30149	30149	30149
F stat	15.92	15.92	15.92				
Average RT					28.2	28.2	28.2
Average age					70.4	70.4	70.4
Average distance					15.8	15.8	15.8

NOTES: the estimate results reported in columns (1)-(3) are obtained by including the full set of controls, namely mission characteristics (fixed effect for the contact center that managed the call, ambulance priority dispatch, dummy for Advance Life Support (ALS) ambulances, distance driven by the ambulance and its square), time fixed effects (day of the week, holiday, year), individual demographics (gender and age category), and location fixed effects (municipality and population density). Column (1) reports the baseline results, where the outcome is a dummy equal 1 for death (level 5); zero for degrees of severity 1, 2, 3 or 4. Column (2) report the results where the outcome is coded as a dummy equal 1 for severity degrees 4 or 5; zero otherwise. Column (3) report the results where the outcome is coded as a dummy equal 1 for severity degrees 3, 4, or 5; zero otherwise. Column (4) reports the results for the linear model, where the outcome ranges between 1 and 5 (greater for higher degrees of severity). The results in columns (5)-(7) report the marginal effect obtained by estimating the ordered probit model; the set of covariates does not include fixed effects at the municipality level to allow convergence; lower severity degrees (1 and 2) are grouped together because of the numerosity of degree 1. The four parameters estimated by the ordered probit sum up to 1, so I report the results of three of them: the ones associated to: degree 3 of severity, column (5); degree 4 of severity, column (6); degree 5 of severity, column (7). The bottom of the table reports the first stage F-statistics and the average values of RT, age and distance driven by the ambulance at which the maximum likelihood estimates are calculated. *** $p<0.01$, ** $p<0.05$, * $p<0.1$

Table 10: Descriptive Statistics: driving time of the ambulance on the way to go (from ambulance dispatch to arrival on the scene) and on the way to go back (from departure from the scene to arrival to the hospital).

Variables	N. Obs	Mean	Median	Std. Dev.
Ambulance driving time on the way to go	30,394	15.86	13.22	12.20
Ambulance driving time on the way to go	27,729	15.22	12.80	11.28
Ambulance driving time on the way to go back	27,729	11.51	8.52	15.61

NOTES: the first row of the statistics shown in table reports the descriptives for the full sample. The second and the third row report the descriptives for the observations with complete records on both the ambulance driving time to go and to go back. The columns reports, respectively, name of the variable, number of observations, mean, median, and standard deviation in sample.

Table 11: Average difference between the driving time to go and driving time to go back at the mission level

Go	3.70***
	(0.12)
Constant	11.51***
	(0.09)
Observations	55458

NOTES: Driving time is the time required to the ambulance to reach the destination. The number of observations is double with respect to the descriptives reported in table 10: to each event, corresponds two driving times, one on the way to go and the other on the way back. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

7 Appendix

The raw data collection includes 43,713 observations. I eliminate 4,171 observations, for which priority dispatch, response time (RT), health outcomes (H1 or H2), gender or age of the patient are missing. Then, I drop the observations that include outlier values for age (over 100 year), 125 observations, and for RT, the 99th percentile of the *RT* distribution (90 minutes or greater), 394 observations. Finally, when multiple EMS vehicles reaches the scene, I consider the response time of the vehicle that reached first the scene, leaving out of the analysis 8,874 records. However, I keep track of the number and characteristics of the other vehicles that reached the scene. The final sample accounts for 30,149 observations.

Table 12: Sample construction

Sample description or step	Observations
Raw mission data	43713
Drop missing values[1]	39542
Drop patients older than 99 years	39417
Keep only the first vehicle that reaches the scene	35581
Drop 99th percentile of response time distribution	30149

NOTES: This table describes the steps that lead to the working sample. [1] Missing values on the variables: ambulance priority dispatch, response time (RT), health outcomes (H1 or H2), gender or age of the patient.

7.1 Alternative instrument: night work shift

As an alternative instrument for RT, I look at the hour of the day in which the ambulance mission is performed. The dummy N equals one when the mission takes place during the night shift, between 8pm and 7:30am. About 32% of missions take place during this time, with longer average RT because of the fewer personnel on duty. All personnel, in rotation, works during night shifts, with a maximum of 4 nights per month per person. The first stage estimates reported in the bottom panel, column (2), of Table 13, shows that during night shift RT is 2.4 minutes

longer, and the F-statistic is 211.8.[26] Column (1) of Table 13 reports reports the estimates for the effect of RT on M1, where RT is instrumented by rainfall. Column (2) shows the results for the night shift instrument, N. As a robustness check, in column (3) I report the estimates where RT is instrumented with N and the sample is restricted to the missions that take place around the shift cutoff, between 6 and 9 am. Column (4) reports the over-identified estimates, where both instruments are included. The Sargan test and its p-value reported in the bottom of table indicate that the effect of RT on M1 obtained by using each instrument is statistically equivalent.[27] Column (5) shows the results obtained by using rainfall and night-shift to instrument RT and its quadratic term, in order to study the potential non-linear effect of RT on M1. The results are comparable across specification and indicates that in the support of my data RT have a linear effect on M1.

[26]Figure **??** reports the probability of performing an ambulance mission by hour of day. The probability is higher during the morning and decreases during the rest of the day. The identifying assumption for the instrument validity is that the hour of the day is not related with the severity of the cardiovascular problem – not the probability that the event takes place. The hour of the day does not correlates with the severity of the cardiovascular problem (Muller et al., 1989).

[27]Sargan test assumes homoskedastic errors. Given that the standard errors do not change, in order to perform this test I did not cluster the standard errors in the estimates reported in column (4).

Table 13: Effect of ambulance Response Time (RT) on Morbidity rate (M1): alternative instruments and specifications

Dependent variable: Morbidity at the Arrival on the Scene (M1)

	Rainfall (1)	Night Shift (2)	NS 6-9am (3)	Overid. (4)	Quadratic (5)
RT	0.015**	0.023***	0.019***	0.022***	0.049*
	(0.007)	(0.002)	(0.005)	(0.002)	(0.028)
Response Time^2					- 0.000
					(0.000)
First Stage (Dependent Variable: Response Time):					
Rainfall (mm)	0.343***			0.343***	25.467***
	(0.085)			(0.086)	(7.140)
Night Shift		2.399***	3.620***	2.400***	138.304***
		(0.165)	(0.469)	(0.165)	(13.167)
F statistic	16.10	211.79	59.67	112.68	1.96
Sargan Test				0.90	
Pvalue ST				0.34	
Observations	30149	30149	5259	30149	30149

NOTES: all results are obtained by including the full set of covariates, namely mission characteristics (contact center that managed the call, ambulance priority dispatch, dummy for Advance Life Support (ALS) ambulances, amount of kilometers driven by the ambulance and its square), time fixed effects (weekday, holiday, year), individual demographics (gender and age category) and location fixed effects (municipality and population density). Rainfall and night shift are, respectively, amount of hourly rainfall at the municipality level and a dummy equal 1 for missions initiated between 8pm and 7:30am. Column (1) reports the baseline results, where *RT* is instrumented by rainfall. In column (2), work shift is adopted as instrument for RT. Column (3) presents the results where night shift is the instrument and the sample is restricted around the shift threshold (6-9 am). Column (4) shows the overidentification results by using both instruments and on the bottom it reports the result of the Sargan Test. Sargan test assumes homoskedastic errors. Given that the standard errors do not change, in order to perform this test I did not cluster the standard errors in the estimates reported in column (4). Column (5) reports the results for the 2 instruments and two endogenous (RT and RT^2) regression. These estimates returns two first stage F-statistics: the first is identical to the one in column (4), the second is reported on the bottom of column (5). The table reports the F statistics for the instrument relevance; clustered standard errors at the hour, day and municipality level in parentheses. *** p<0.01, ** p<0.05, * p<0.1

7.2 Heterogeneity analysis

To study the heterogeneity of the effect across different population subgroups, I split the sample accordingly to different categories. The estimates of the effect of RT on M1 by gender and age subgroups (where the age quartiles identified the patients younger than 62; 62 $\geq$ age $<$ 75; 75 $\geq$ age $<$ 84; and older than 83) are reported in Table 14. Table 15 reports the results by splitting the sample across the specific classification of cardiovascular problem, namely heart failure; hypertension; and other cardiocirculatory problems. All the estimation results includes the usual full set of controls and clustered standard errors. The estimates are obtained by adopting the work shift instrument; this is the only one with enough statistical power in the subgroups to return significant estimates. The results shows that older patient and males are more sensitive to ambulance delays, so that service providers might take into account these characteristics when assigning the priority to patients.

Table 14: Heterogeneity of the effect on morbidity rate (M1) by gender and age quartile. Instrument: Work Shift

Dependent variable: Morbidity at the Arrival on the Scene (M1)

	Female	Male	Age<62	Age 62-74	Age 75-83	Age>83
	(1)	(2)	(3)	(4)	(5)	(6)
RT	0.018***	0.027***	0.009	0.019***	0.033***	0.026***
	(0.003)	(0.004)	(0.006)	(0.004)	(0.006)	(0.004)
Obs	15141	15008	7770	7845	7287	7247
Fstat	123.32	96.27	31.48	57.29	44.36	97.12

NOTES: all results are obtained by including the full set of covariates, namely mission characteristics (contact center that managed the call, ambulance priority dispatch, dummy for Advance Life Support (ALS) ambulances, amount of kilometers driven by the ambulance and its square), time fixed effects (weekday, holiday, year), individual demographics (gender and age category) and location fixed effects (municipality and population density). Column (1) reports the results only for women, column (2) only for men, and columns (3)-(6) by age quartile: younger than 62; 62 $\geq$ age $<$ 75; 75 $\geq$ age $<$ 84; older than 83. First stage F-statistics on the bottom of the table and clustered standard errors at the hour, day and municipality level in parentheses. *** p<0.01, ** p<0.05, * p<0.1

The results reported in Table 15 show that the individuals suffering from hearth failure are the most sensitive to ambulance delays: 3.6 percentage points increase

in the morbidity rate (M1) for every additional minute (RT). The morbidity rate of patients affected by hypertension and other cardiocirculatory problems increases, respectively, by 1.6 and 2 percentage points for each additional minute.

Table 15: Heterogeneity of the effect on morbidity rate (M1) by type of cardiovascular disease. Instrument: Work Shift

Dependent variable: Morbidity at the Arrival on the Scene (M1)			
	(1)	(2)	(3)
Response Time	0.034***	0.017***	0.023***
	(0.006)	(0.006)	(0.003)
Observations	1457	3378	25314
F statistic	40.63	22.09	168.09

NOTES: all results are obtained by including the full set of covariates, namely mission characteristics (contact center that managed the call, ambulance priority dispatch, dummy for Advance Life Support (ALS) ambulances, amount of kilometers driven by the ambulance and its square), time fixed effects (weekday, holiday, year), individual demographics (gender and age category) and location fixed effects (municipality and population density). Column (1) shows the results for the only patients affected by heart failure; column (2) for hypertension; column (3) includes all the other cardiovascular problems. First stage F-statistics on the bottom of the table and clustered standard errors at the hour, day and municipality level in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Chapter 2

Ambulances get lost: the problem of patient localization in emergency care

Elena Lucchese*

Abstract

Rapid responses in emergency increases the likelihood of clearing crimes for police and saving lives for ambulance. Very little is known, however, about the determinants of response time. In this paper I quantify the localization problem by making use of a difference-in-differences identification strategy. I calculate the difference between driving times at the mission level: on the way to go, the directions adopted by the ambulance driver to reach the scene of the event are provided by the caller; on the way back, the location of the hospital is known by the driver. Then, I compare this difference across type of places: is more likely that the driver knows the location of public places as opposed to private dwellings. Using high quality administrative data on ambulance driving times, I document that the localization problem accounts for 5 minutes of delay for urgent (non-deferrable) missions, 30% at the average driving time. The magnitude of the effect is not affected by the distance travelled by the ambulance. The delay is halved for non-urgent missions, when the caller is less stressed and communicates more clearly. Introducing a technology to enhance location would dramatically improve performance at a virtually zero cost.

JEL classification: D29, D90, I12, I18, R41

Keywords: Ambulance, Emergency, Organizational Performance, Response Time.

1 Introduction

The timing of an intervention is key for success, especially in the case of emergency situations. The rapidity of response is informative about performance, and in the United States local health care agencies contractually agree response times with ambulance providers (Ludwig, 2004). In Europe, ambulance response time targets are regulated by local administrations or settled by national laws.[1] Also, several exceptions to traffic laws were introduced to reduce the response time of emergency vehicles, such as the right-of-way at crossroads and traffic lights and the use of lights and sirens. As a results, the return from shorter responses is the focus of much recent work in the economics

*Department of Economics, University of Bologna, Piazza Scaravilli 2, 40126 Bologna, Italy. E-mail: elena.lucchese@unibo.it

[1]For additional information about response time thresholds in Europe: ec.europa.eu/health.

literature, which show that a rapid reaction increases the probability of clearing the crime by police (Blanes i Vidal and Kirchmaier, 2015) or saving lives by ambulance (Wilde, 2013). Typical approaches have also developed models to optimize the location of ambulances (see Brotcorne et al. (2003) for a review); have supported the introduction of computer-aided dispatch services (Athey and Stern, 2002); and have studied the effect of personnel experience (David and Brachet, 2011) and other environmental characteristics such as population density (David and Harrington, 2010), on patient health outcome.

Current literature provides little insights to understand the determinants of response time, and the aim of this work is to fill this gap. This contribution would allow social scientists and policymakers to make sense of the differing efficiency of alternative policies. The policy implication is important given that the heterogeneity of health care expenditure and patients outcome across countries is remarkable, and having in mind the channels that drive the results would help to improve the efficiency.

During the emergency call important information is transmitted, including the directions to reach the site of the event. The panic or fear that the caller may be experiencing in that moment, however, can substantially affect the quality of communication, resulting in unclear or inaccurate directions. Good quality directions is a fundamental piece of information for the responder, that will not be able otherwise to localize the patient. To quantify the magnitude of the localization problem, I consider two factors. First, an emergency mission includes two driving periods: the way to go – from the ambulance dispatch to the arrival on the scene – and the way back – from the departure from the scene to the arrival at the hospital. The location of the destination at the end of each way is, on average, known with different precision: on the way to go, the ambulance driver must be able to localize the patient by making use of the directions obtained during the emergency call; on the way back, instead, the driver knows the location of the hospital. Secondly, is easier to locate certain type of places than others. Public places (e.g., an office or a street) are usually more visible, and are likely to be more well-known to responders than private dwellings, so that it may be more difficult to localize private residences than the public (non-residential) ones (David and Harrington, 2010).

I identify the effect of interest – the localization problem – by making use of a difference-in-differences identification strategy, where I calculate the difference between driving times and across type of place. The first difference clears the results from the compounding effect of unobservable factors that do not change over time – from the way to go to the way back. Examples are the characteristics of the driver (experience) and of the vehicle (tires level of wear). The second difference clears the results from unobservables that change over time but do not depend on the type of place. An example is the memorisation of the route by the driver. I test whether the estimated effect – the longer time employed by the ambulance on the way to go – is a localization problem by performing a set of exercises. First, I recognise that the localization is a *last mile* problem – an issue that is faced by the driver once in the area of the event. As such, the magnitude of the effect should not depend on the distance driven by the ambulance. I show that the estimates do not change across distance and neither across patient's age or pathology and distance. Finally, I show that the localization problem is greater for urgent rather than

non-urgent calls. This circumstance also highlights the seriousness of the problem, given that patients in most severe conditions are the ones that experience the longer delay. Unlike non-urgent cases, it is less likely that the urgent caller has the time to calm down before calling, so that the communication quality is worse in most urgent situations. The identification strategy adopted requires the parallel trend condition – a condition that cannot be tested because it includes counterfactual outcomes. In particular, it requires that the changes between driving times would have been the same across type of place in the absence of the localization problem. In the specific case in analysis, this condition would be violated if there is an unobservable factor that change over time, drives the results, is heterogeneous across type of place, is not affected by the distance travelled by the ambulance (i.e., it does not depend on the length of *exposure* – the driving time) and is not the localization problem. A circumstance that satisfies all these conditions seems unlikely.

To perform the analysis, I exploit a unique dataset on all ambulance runs performed in the Italian region Liguria in 2013 and 2014. There are several advantages associated with the use of these records. The geographic extension covered by the data, the regional level, is at the same level at which the health care service is managed and provided, allowing for a greater precision of the analysis. In addition, all the data are recorded in real time during the emergency mission, and not self reported at the end, as in other similar data collections. The duration of the steps of each mission (i.e., the ambulance departure and the arrival on the scene, the departure from the scene and the arrival at the hospital) are automatically recorded by an informative system. In this way, the precision and the quality of the data is high, and the likelihood of biases or misreporting is minimized. Finally, there is little risk that the results are affected by selection on patients, given that in the Italian system – as well as in most European countries – the service is provided free of charge. Also, the destination hospital is not chosen by the ambulance provider, but by the Emergency Department that answers the call and that identify the closer hospital with space available, and that communicates this information to the driver. The main analysis considers only non-deferrable calls – the most relevant in the emergency context – and the missions where the patient is transported to the same hospital where the ambulance departed from. With this last restriction I test the robustness of my findings to the characteristics of the street driven by the ambulance, given that in this case it will certainly be similar. A consequence of this last restriction is that most of the missions considered are performed in urban areas. The hospitals are usually located in more populated areas, and the ambulances that return to the hospital of origin have rescued a patient that was likely to be in an urban area too. Apart from population density, all the other observable characteristics are extremely balanced, and the analysis includes numerous hospitals so that it mitigates concerns that the effect is driven by peculiar contingencies associated with a given hospital. Also, the magnitude of the effect does not change by including the entire sample. I show that the location problem accounts for 5 minutes delay – one third of the average driving time to go – and is 3 minutes – one fifth at the average – for deferrable calls. The effect is not affected by the distance covered by the ambulance, and by the age or pathology of the patient

and the distance.

The rest of the paper is organised as follow. Section 2 describes the data. In section 3 I discuss descriptive evidence documenting that driving times on the way to go are longer than driving times on the way back, also after imposing several restrictions to the sample. I illustrate the identification strategy in section 4. I present the main results in section 5, where I also explore the heterogeneity of the effect across observable characteristics and perform some robustness checks. Finally, on section 6 I discuss the policy implication of the results and I conclude.

2 Data

My data source is the universe of emergency ambulance missions from Liguria region, in Italy, in 2013 and 2014. The use of these data was previously authorized under a data use agreement with the health regional authority.[2] The record includes data on pathology, priority dispatch, health condition of the patient, and patients characteristics such as age and gender. In addition to these standard variables, the dataset offers some clear cut advantages in terms of data and the way those are collected.

The data collection in Liguria is supported by a management system that is integrated with the EMS and that enable the collection of information on-the-spot. In this way, the data are precise and more accurate that it would be if reported at the end of the mission, as it appends with similar data collections. In particular, each mission consists in several phases, the duration of which is recorded by the management system. Hence, we can separate the time spent at the phone by the ambulance dispatcher and the time required to prepare the ambulance, from the driving time of the ambulance. As such, we can isolate the time that the ambulance spends driving from all the other phases of the mission. In the same way, also all the other variables in the dataset are collected on-the-spot. [3]

The driving time (DT) tells us how many minutes the ambulance spend driving, from departure to arrival, and it is my dependent variable. Each mission consists of two driving times. The driving time to go, from the ambulance dispatch to the arrival at the scene, and the driving time to go back, from the scene to the hospital. Another important information is given by the patient's type of location. This is classified by a dummy that assumes a value of one for residential locations, zero otherwise (school, office, workplace,

[2] In Italy, as in most European countries, the health care system is organised at the regional level. Data that covers the entire geographic area in which EMS is organized are more accurate in capturing the overall effect of interest.

[3] The sequence of actions is the following: during the call, the management system automatically records the date and time at which the call started and ended. During this time span the nurse asks a set of predetermined questions and collects the information needed. An ambulance is then assigned with a priority dispatch code. When the ambulance is dispatched, the management software records the date and time of the ambulance departure and arrival at the scene. The emergency care provider observes the patient's severity and provide medical treatment. Finally, the management system registers the date and time of departure from the scene and arrival at the hospital. At the end of the mission, upon arrival at the hospital, the patient's health condition is recorded one last time.

street). Finally, the dataset provides also the total amount of kilometres (1 mile = 1.6 kilometres) assigned to each mission.[4]

In addition, I investigate the heterogeneity of the effect of interest by type of pathology and age group. The pathology is classified with respect to the three classes: injury and cardiocirculatory, which on average lead to the most severe health condition treated in emergency care, and all the other pathologies grouped together in the third class.[5] There are three patient's age groups: under 35, 35-64, and 65 or older.

I restrict the sample to urgent, non-deferrable, missions because for these the ambulance reaches the patient as fast as possible and it is likely to expect that it get there directly without any detour. I exclude also the patients who died before reaching the hospital, as for those the urgency is not high anymore. I eliminate the observations for which the relevant variables mentioned above are missing. Finally, I consider only the events where the ambulance started the missions from the hospital in which the patient was, at the end, transported to. This is a large drop in the data, as it eliminates 90% of the sample. However, the big advantage that results from considering only these mission is that the driver travels the same distance on the way to go and on the way back. This circumstance eliminates the concern that the results may be due to a difference in the distance travelled. Table 1 illustrates the steps that lead to the working sample.

Table 2 reports the descriptive statistics for the full sample (196,759 observations) and for the working sample, where origin and destination is the same (18,858 observations).[6] The size of the full sample is 10 times greater than the working sample because most of the ambulances are not parked at the hospital, in order to ensure an adequate coverage of the territory. Reducing the working sample in this way is a strong robustness test on the results. Even by reducing the sample size to 1/10, focusing the analysis on the only missions in which the distance driven in the same in both ways, returns a sizable difference between driving times. Given that hospitals are usually located close to urban areas, 81% of the missions in the working sample are performed in highly populated municipalities, against 53% in the full sample. Also the average distance covered is lower in the selected sample, with 9 km average against 17 km in the full sample. Despite these differences, the distribution of the other characteristics is quite similar across the two groups. On average, the driving time to go is about 15 minutes, and the time back takes 12 minutes (11 in the working sample). About 60% of missions head to residential locations and 38% to other type of locations. Injury and cardiocirculatory problems account each for about 20% of non-deferrable missions, and the rest are other pathologies. Most of the patients observed are in the older age group: 60% are 65 year

[4]In particular, I observe both the amount of kilometres estimated by the system and the actual distance driven by the ambulance. In this setting I adopt the number of kilometres estimated, as it is a measure that is not affected by possible extra kilometres travelled by the ambulance that is not able to locate the patient and drives in circles, and provides a clearer measure on how far a patient is.

[5]On one of the specifications I also adopt an extended classification for pathology, which includes: injury, cardiocirculatory, respiratory, neurologic, psychiatric, toxic, gastro, and other.

[6]The number of observations is the number of missions. In the regression results the sample size is double because the dataset is in long format. As such, for every mission there are two observations: the driving time to go and the driving time back.

or above, 30% are 35-64 year, and about 15% are below 35 years. Finally, there is 1/7th probability that an urgent call takes place in a given day of the week.

3 Descriptive Results

Driving time to go takes, on average, longer than the driving time to go back. Figure 1 reports this relation, where the left panel illustrates the average driving times for the full sample (that includes all non-deferrable missions), and the right panel for the working sample (that is all non-deferrable missions where origin and final destination is the same). As illustrated, this difference persists. This result is unexpected, as the guidelines that regulate the activity of the EMS prescribes ambulances to drive faster on the way to go – when the patient is in urgent need of care – and slower on the way back to avoid abrupt braking and accelerations.[7]

I run a preliminary analysis to check if these guidelines are actually implemented. To do so, I regress rain on driving time.[8] Rain is a dummy variable that equals one when in the hour and municipality in which the mission is performed was raining. During rainfall the road condition deteriorates and the vehicles generally drive safe at a lower speed (Highway Capacity Manual, 2000). Given that, if the driver adjusts the speed and drives slower during rainfall compared to the missions in which is not raining, we expect to observe a positive effect of rain on driving time. Given the different speeds prescribed by the guidelines depending on the phase of the mission, we would expect to find a positive effect of rain on the driving time to go, when the speed is supposed to be higher. On the way back, instead, the speed should be low. In this case so the effect should be close to zero. The results are reported in Table 3, and show that, as expected, the rain has a positive and statistically significant effect on the driving time to go only. The driving time back, instead, is unaffected. The effect is similar in both samples. This result provides suggestive evidence that the guidelines are implemented – that is, the ambulance drive faster on the way to go.

The difference between driving times may arise because of the difficulty to exactly locate the patient. During the emergency call, the spokeperson provides important information to the ambulance dispatcher and the directions provided must be as accurate as possible. Generally the ambulance driver relies uniquely on this information to find the patient. However, accurate directions matter only on the way to go. On the way back the location of the hospital is known by the ambulance driver. As such, even if the ambulance travel the same route, there is a big difference between the way to go and the way back in terms of certainty about the destination.

A way to study the magnitude of this (potential) location problem it to regress a dummy that equals one for the way to go on driving times. Even after introducing a rich set of controls, however, this approach would present a number of problems, as it may be

[7] The guidelines for the region Liguria are reported in *"Guida pratica per i soccorritori delle Associazioni convenzionate per il soccorso con il Servizio 118 ligure"*, 2013.

[8] The data on rainfall are provided by the regional agency for environment (ARPAL), which makes use of 213 land-based weather stations to collect hourly information on rainfall.

that the effect is partially driven by the compounding effect of unobserved variables. A way to solve this problem is to compute the driving time difference at the mission level, which would clear the effect of interest from unobserved characteristics constant at the mission level. However, this approach cannot control for factors that change over time at the mission level. For example, learning the road may play a role, as it is travelled again short after, on the way back. To address the problem that arises from the compounding effect of unobservable factors that change over time, I propose the identification strategy discussed in the next section.

4 Identification Strategy

To identify the magnitude of the location problem I propose an exercise that exploits the fact that, as discussed by David and Harrington (2010), residential destinations are more difficult to be found than non-residential. This is because non-residential locations, such as workplaces, offices, schools or streets, are on average more likely to be well-known to responders than private residencies. As a result, by comparing the driving time differences across these two groups, I can clean the estimates by the compounding effect of unobservable factors that vary over time, as long as they vary in the same way for residential and non-residential destinations. To get back to the example mentioned above, it seems reasonable to expect that the effect of learning the road is similar across type of locations (residential and non-residential). In this setting, the missions to residential locations are the treatment group – as those are the most harder to find – and the non-residential destinations are the control group.

The regression model is specified as follows:

$$DT_{mw} = \alpha_0 + \alpha_1 Go_m + \alpha_2 Residential_m + \alpha_3 Go_m \times Residential_m + u_{mw},$$

where DT is the the driving time for mission m on the way w. Go is a dummy that equals one on the way to go, and $Residential$ is a dummy that indicates the missions to residential locations, against other destinations (namely work place, school, office, sport facility, street and other). The error term u captures the residual components that affect the outcome variable and are not captured by the included regressors.

The constant term, α_0, is the average driving time to go back from non-residential destinations. The parameter α_1 is the driving time difference for non-residential missions. Parameter α_2 is the difference between driving time to go back by location – residential compered to non-residential. The parameter of interest is α_3, and it identifies how much the driving time difference is different between residential and non-residential destinations.

One condition for the validity of the identification strategy is the assumption of common trend between treatment and control group. We cannot test this assumption. If the problem is given by the difficulty to locate the patient, however, we may expect that the treatment (the location problem) takes place at the end of the way to go. Said

in other worlds, we would expect that the location problem is not proportional to the length of the mission, and that it takes place only at the very end, once the ambulance is already in the area. We can name this the last-mile problem. To perform this exercise I run the regression in equation (4) on different distance groups. As I show in the next section, I find that the effect of interest is very similar between distance groups and across different distance definitions.

In the next section I also discuss the heterogeneity of the effect and some robustness checks. Even if the working sample includes only urgent calls – that is, all the patients considered are in great difficulty – there may be some pathologies that has a greater detrimental effect on communication quality than others, and so are more likely to result in a location problem. An example are psychiatric problems. On the other side, we expect that patients in better health conditions – that is non-urgent calls – are able to communicate more clearly their location. Even if the ambulance driver does not try to reach non-urgent patients as fast as she does for the urgent ones, we expect that better communication would result, for non-urgent patients, in smaller delays if better communication reduces the location problem.

5 Results

Main Results

The difference-in-differences estimates based on equation (4) are presented in Table 4. The sample is divided in 4 groups of 5 kilometres each. A fifth group includes all the observations where the distance is 20 kilometres or greater. The distance distribution is shown in Figure 2.[9] The plot of the estimates is illustrated by Figure 3, where 4 panels report the point estimates for each parameter by distance group.

The coefficients on *Go* (α_1) express the difference between driving time to go and to go back for non-residential destinations. The coefficient is generally not significantly different from zero, and the magnitude is relatively small. This is coherent with the fact that for this group the location problem is less of an issue, as discussed by David and Harrington (2010). The coefficients on *Residential* (α_2) are negative and tell that driving back from residential locations takes less time than from non-residential locations. The constant term is the average driving time to go back for non-residential calls and it increases monotonically with the distance, as it is reasonable given that it takes longer to drive more kilometres. The same is true for residential missions, as the average time to drive back is given by the sum of the constant term and the coefficient of *Residential*.

The coefficient on the interaction term, $Go \times Residential$ (α_3), is the difference between driving times, between residential and non-residential missions. The sign is positive, which reflects the fact that, on average, driving time to go is greater than driving time to go back and that the difference is greater for residential compared to non-residential locations. The parameter on the interaction term is stable across distance

[9] The average distance travelled by the ambulances is 9 km (the median in 6). The first 4 groups includes 90% of the observations. The rest in the last group.

groups and ranges between 5 and 6 minutes. This result supports the conjecture that the location problem is a last-mile problem, as discussed above. Also driving time difference for residential (given by $\alpha_1+\alpha_3$) and for non-residential destinations (given by α_1) is relatively constant across distance groups.

Finally, I estimate the effect by adopting a different distance definition. In particular, I compute the groups by quintile of the distance distribution, and the kilometres intervals obtained in this way are respectively: 0-2, 3, 4-6, 7-12 and $\geq$13 kilometres. This alternative definition is interesting because the distance distribution is right skewed, and fixed interval kilometres, other than returning very different groups numerosity, may not capture the existence of possible thresholds in the data effects. However, despite the remarkable difference that results from defining the classes in this way instead of the fixed kilometres interval, the parameters on the interaction term are again relatively constant across groups and range between 4.2 and 6.6 minutes. The results are reported in Table 5 and illustrated in figure 4.[10] In what follows I will mainly focus on the 5-kilometres specification as it returns results that are easier to compare in the case future research will perform a similar exercise with different data.

Heterogeneity and Robustness of the Results

Tables 6 and 7 report the results by age group and gender, and by type of pathology respectively. In estimating this heterogeneity effects I adopt the extended definition for pathology, which divides the pathology in 8 groups, as described in previous section. The effect of the interaction term is, for all specifications, statistically significant and positive. As in the main results presented above, also when I divide the population by age or gender subgroups the parameters values lie in the same value range. A remarkable difference instead appears in the results reported in table 7. Missions activated for patients affected by psychiatric and toxic problems (columns 5-6) results in greater differences in driving times for residential versus non-residential calls. The effect is of 11 minutes for the former and almost 8 minutes for the latter, despite the magnitude of the constant term and of the the parameters do not differ much from the ones of other pathologies. It is likely to assume that patients affected by psychiatric and toxic problems and that are located in residential locations may be more difficult to reach than their counterparts in non-residential places, and communication problems may be one of the factors that drives the results.

Table 8 presents the results by adopting the baseline specification – that is the difference of the driving time difference across type of location, by distance – on non-urgent calls. As before, driving time is defined as the number of minutes spent driving in

[10]The estimates of the parameter *Residential* discussed above can raise concerns on identification as they suggests that it takes less to the ambulance to drive back to the hospital from residential locations than from non-residential ones. One may argue that this is partially driving the difference between driving time to go and to go back for residential calls. This effect, however, should be proportional to the distance driven by the ambulance. Conversely, what I show is that the effect of interest is constant across distances – that is, it is a last-mile problem. The fact that we observe constant effects across distances suggests that there is no violation of the parallel trend in this setting and the results can be interpreted as I suggest, namely as the effect of the patient location problem.

two directions. Even if the activation of the ambulance missions to respond to these calls may be deferred, the driving time does not capture such delay because it is computed from the moment in which the ambulance is dispatched. As we expect that situations in which the patient is better off are also less stressful for the caller, this may also results in higher quality communication. This is likely to be the case because high-priority dispatches are often the result of unexpected events that turns out to be severe situation which results in impetuous calls.

The effect for non-urgent calls is still positive and statistically significant, but 40 percent smaller than the effect for urgent calls (on average the effect is about 3 minutes instead of 5). The results are stable across distance groups and the parameter of interest drops for distances greater than 20 kilometres. This mission lengths are far from the centre of the distance distribution, where the average is 7 kilometres and the median is 4.[11] The effect probably disappears because at the increasing of the distance the difference between driving times between groups (residential and non-residential) gets smaller, as the ambulances that travel the greatest distances may get lost in a similar way across type of destinations. As such, the interaction terms, which returns the difference in the driving times between destinations (residential vs non-residential) assumes a value that is close to zero.

5.1 Other Results and mechanism

In Section 5 I discuss the results obtained by splitting the sample by distance driven by the ambulance. The parameter of interest – i.e. the amount of extra time required to reach a patient at home compared with a patient in other locations – is constant across distances. In the case of location problems we would expect such finding, as the problem of patient location is a last mile problem, i.e. a problem faced by the ambulance driver once in the area of the event and that, as such, is not significantly affected by the total distance driven to get there. In this section I present the results referred to other robustness checks and tests to further support the interpretation of the results. The results are illustrated in Table 10. All the results discussed in what follows are referred, as usual, to the only missions in which origin and final destination of the ambulance are the same.

Columns (1) and (2) in Table 10 report the results obtained by splitting the sample by time of the day in which the mission is performed. The estimates reported in column (1) refer to the missions performed during daylight; the estimates in column (2) refers to the missions performed during night time, from the sunset to the sunrise. We expect that, if the ambulance driver faces location problems, is going to be more difficult to locate the patients during the night than during the day because without light is more difficult to adopt reference points to support location. The parameter of interest in the results, indeed, is about 35% grater for the missions performed during the night.

Columns (3) and (4) report the results obtained by splitting the sample between the missions performed during times of intense traffic conditions versus other times. Given

[11] See figure 8 for the distribution of distance.

that the parameter of interest is a difference, we expect a smaller value for the group of observations referred to the missions performed during traffic times. Indeed, during that time, the ambulance is slower also when driving back, so that the difference in the time required to reach the scene as opposed to going back to the hospital is smaller. We expect a smaller effect for the missions performed during traffic times if we also expect that the ambulance driver does its best to drive the route in the shortest time possible, and so the time of the day in which the mission is performed might affect the response because of different traffic conditions. We find an effect in the interaction terms that is 20% smaller when the mission is performed during traffic times.

Column (5) reports the results for a triple difference where the classes are (i) the direction in which the ambulance is driving (go vs back); (ii) the type of location (home vs other); (iii) whether the ambulance mission is urgent or not. This exercise tests my main results in two ways. First, one would expect that the delay associated to urgent missions is lower – actually someone would expect no delay for such missions – because in this case the patient is in life threatening conditions. On the other side, however, in an urgent situation is more likely that the communication during the emergency call is of poorer quality because of the shock and the hit of the moment, resulting in greater delays. If the rapidity with which the ambulance can get to the scene is affected by communication, we should observe a greater delay for more urgent missions – when a fast response is more needed. The parameter associated to the triple interaction term shows that this is the case, and the delay associated to urgent missions of 0.86 minutes greater than the one to non-urgent one: about 17% of the total delay of the ambulance. Secondly, the triple interaction term allows us to compare the driving times referred to patients that are virtually located in the same place – but affected by different degrees of severity. It is interesting to find a positive effect even in this case, as this clears concerns about the possibility that the results are driven by intrinsic characteristics of home that are not related to the location problem. The parameter associated to the triple interaction is obtained by comparing the driving times assigned to patients all located at home, by affected by different degrees of severity. If the location problem is the mediating factor that drives my results, I expect to find greater delays for patients in most critical conditions and this is what emerges from the results.

Finally, column (6) report the results of the placebo test where the sample is restricted to the only patients at home. Then, I split this group in order to create two artificial classes. I show that the effect of interest drops to zero, supporting the idea that is the location problem that affects patients rescued at home that is driving my results.

6 Policy Discussion and Conclusion

Our results show that mitigate the localization problem would improve the EMS performance by up to 5 minutes, one third of the average driving time to go. The magnitude of such potential performance change is extremely sizeable in a setting like this, where even a single minute makes the difference between life and death. In Wilde (2013) one minute reduction in the out-of-hospital time results in 1 percentage points lower 90-day

mortality rate. This corresponds to 1,130 life years, given a sample of 109,000 patients (Wilde, 2013), that is as much as about 55% of non-deferrable calls in Liguria in the 2 years considered in this work. Lucchese (2017) shows that one minute decrease in the time required to reach the patient decreases the out-of-hospital mortality for patients suffering from cardiocirculatory problems by 1 percentage point.

Technological advance, such as computer-aided dispatch services and mobile geographic information system units on ambulances, allowed ambulances to reach patients far more quickly (Wilde, 2013). However, in the situation under analysis, it is correct to consider technological progress also part of the problem. Indeed, before the massive diffusion of mobile phones, the calls was carried out by land-line phones. The location of land-line phones was easy and it was an automatic procedure in the EMS practice. As such, the localization of the patient was not a big problem as it is now, with mobile phones. Nowadays mobile phones can be tracked by police, which can do it because it has access to special technologies. However, the procedure is not as automatic as it is for land-line calls, and it requires an amount of time that may not be compatible with the timeline of EMS operations.

The adoption of smart phone applications that communicate caller location to the responding dispatch center are being tested around the world as a way to improve patient's localization. This type of solution requires the caller to install a given application in the smart phone, and to activate the emergency call through the app. To appreciate the return from this solution, as much calls as possible should be performed by adopting the application supported by the local emergency system. But at the moment this is not what we observe. As the closer example, the application named *Whereareu* was launched in the Italian region Lombardia in mid 2014 with the support of local authorities. However, at the end of 2017, after several years from its introduction, a negligible number of calls is performed by using it.

The introduction of systems to support patient localization would also reduce the cost of organizational forgetting, as discussed by David and Brachet (2011). David and Brachet discuss that most experienced workers in EMS are more proficient in identifying the correct routes. With labor turnover this is a capacity loss for EMS, that would be mitigated if a software would retain part of this learning. These technologies are currently exploited by other service providers where localization of clients is important, such as food deliveries and taxies. However, these technologies may produce positive returns also in other settings such as emergency missions (being medical care but also for the services provided by police and fire departments), where the location of the victim is a key step in the service provision and where the victim may often be in the condition of not being able to correctly transmit this information, being because of the hit and the shock of the moment, or because this is information is not known by the caller.[12]

[12]With this work I am the first to quantify the location problem in EMS. The efficiency gain obtainable by mitigating this problem is 5 minutes, 30 percent increase on average performance level. The benefit from one minute reduction would save 1,130 life-years in Wilde (2013), and would reduce out-of-hospital mortality rate by 1 percentage point in my previous work that exploited the same dataset adopted here, Lucchese (2017). If we value a life-year as the income that the average person earn in one year in Utah – were the analysis of Wilde is performed – it results that it would worth 23,139 $.[13] As such, one minute

References

(2013). Guida pratica per i soccorritori delle associazioni convenzionate per il soccorso con il servizio 118 ligure. Dipartimento Interaziendale Servizio 118 della Liguria.

Athey, S. and Stern, S. (2002). The impact of information technology on emergency health care outcomes. *Rand Journal of Economics*, 33(3):399–399.

Blanes i Vidal, J. and Kirchmaier, T. (2015). The effect of police response time on crime clearance rates. *The Review of Economic Studies*.

Brotcorne, L., Laporte, G., and Semet, F. (2003). Ambulance location and relocation models. *European journal of operational research*, 147(3):451–463.

David, G. and Brachet, T. (2011). On the determinants of organizational forgetting. *American Economic Journal: Microeconomics*, 3(3):100–123.

David, G. and Harrington, S. E. (2010). Population density and racial differences in the performance of emergency medical services. *Journal of health Economics*, 29(4):603–615.

Highway Capacity Manual, M. (2000). Highway capacity manual. *Washington, DC*.

Lucchese, E. (2017). The causal effect of ambulance response time on patient's health outcomes. Technical report, University of Bologna.

Ludwig, G. G. (2004). Ems response time standards. *Emergency medical services*, 33(4):44.

Murphy, K. M. and Topel, R. H. (2006). The value of health and longevity. *Journal of political Economy*, 114(5):871–904.

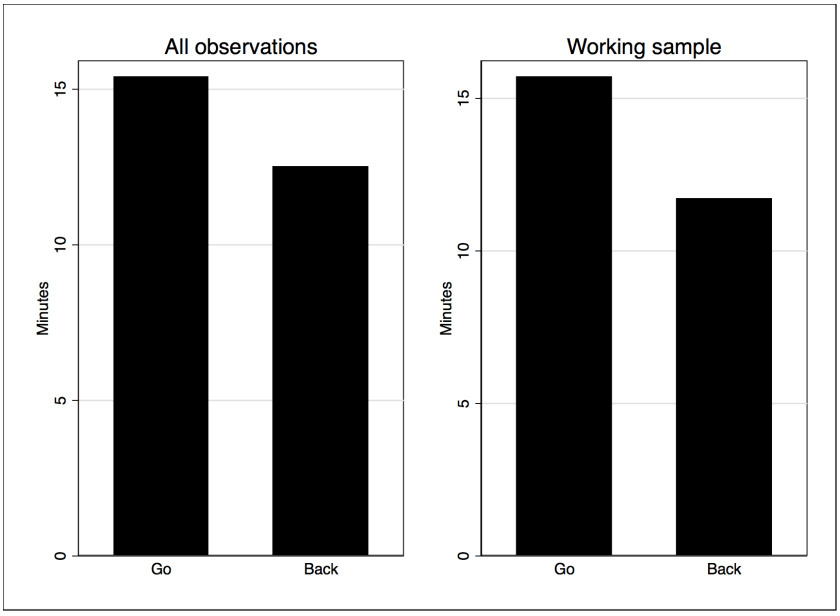
Viscusi, W. K. and Aldy, J. E. (2003). The value of a statistical life: a critical review of market estimates throughout the world. *Journal of risk and uncertainty*, 27(1):5–76.

Wilde, E. T. (2013). Do emergency medical system response times matter for health outcomes? *Health economics*, 22(7):790–806.

reduction in Utah would worth over 26 million dollar (1,130*23,1139). If we consider that estimating the value of a life year in terms of average annual income is strongly underestimated even if compared with the most conservative estimate of the statistical value of life proposed by Viscusi and Aldy or Murphy and Topel.

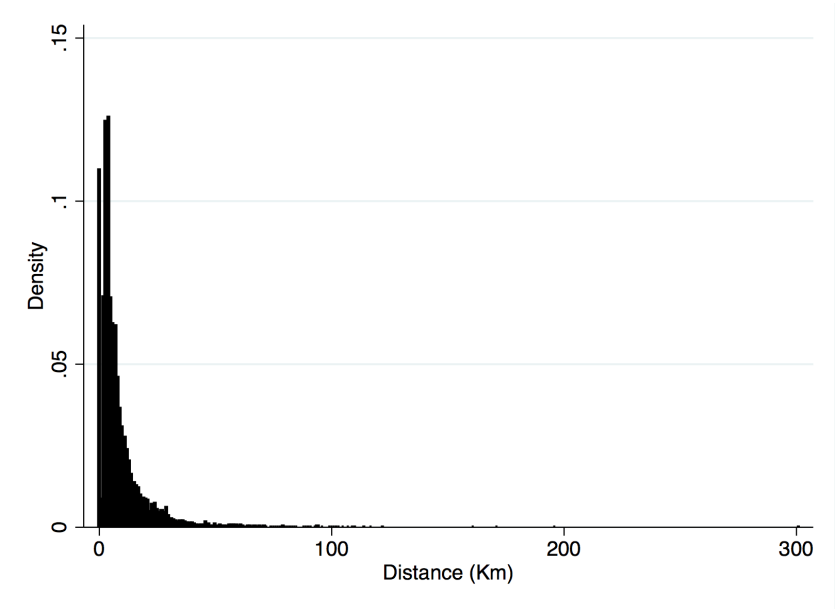
Figures and Tables

Figure 1: Driving time by sample: all observations and only missions with same origin and final destination



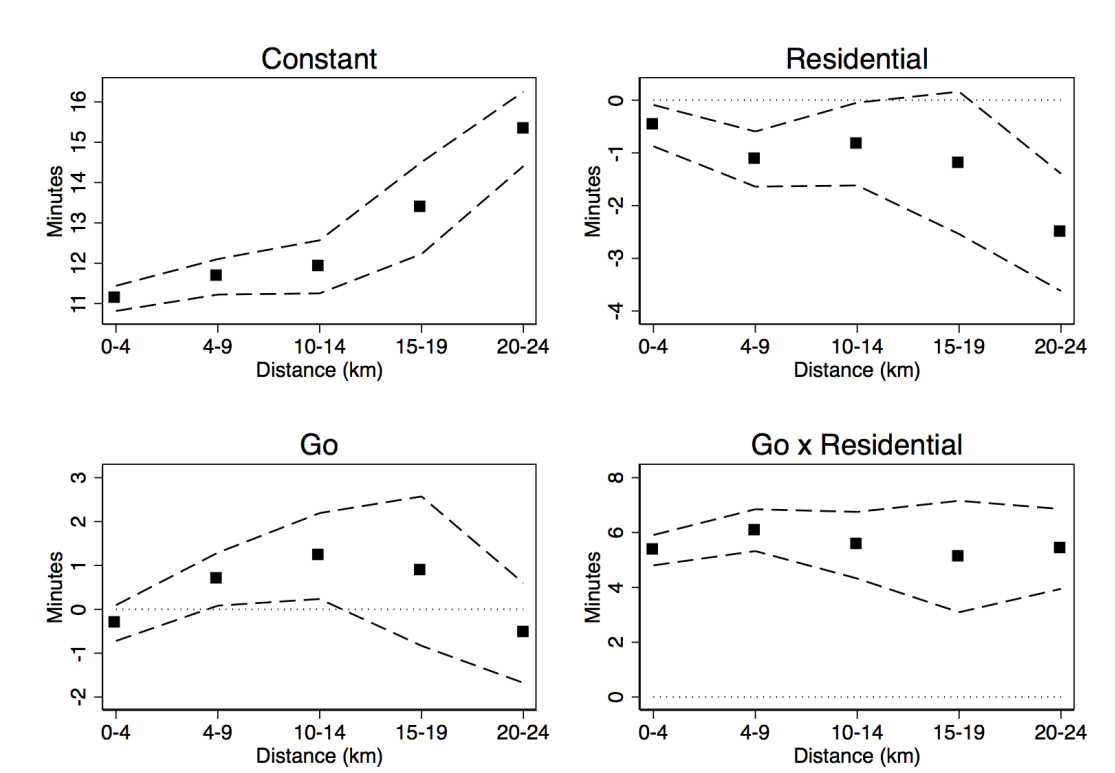
NOTES: the graph on the left hand side reports the average driving times for the entire sample. On the right, the average driving times for the working sample, which includes only the missions which originate and ended in the same location.

Figure 2: Distribution of distance, working sample (non-deferrable calls)



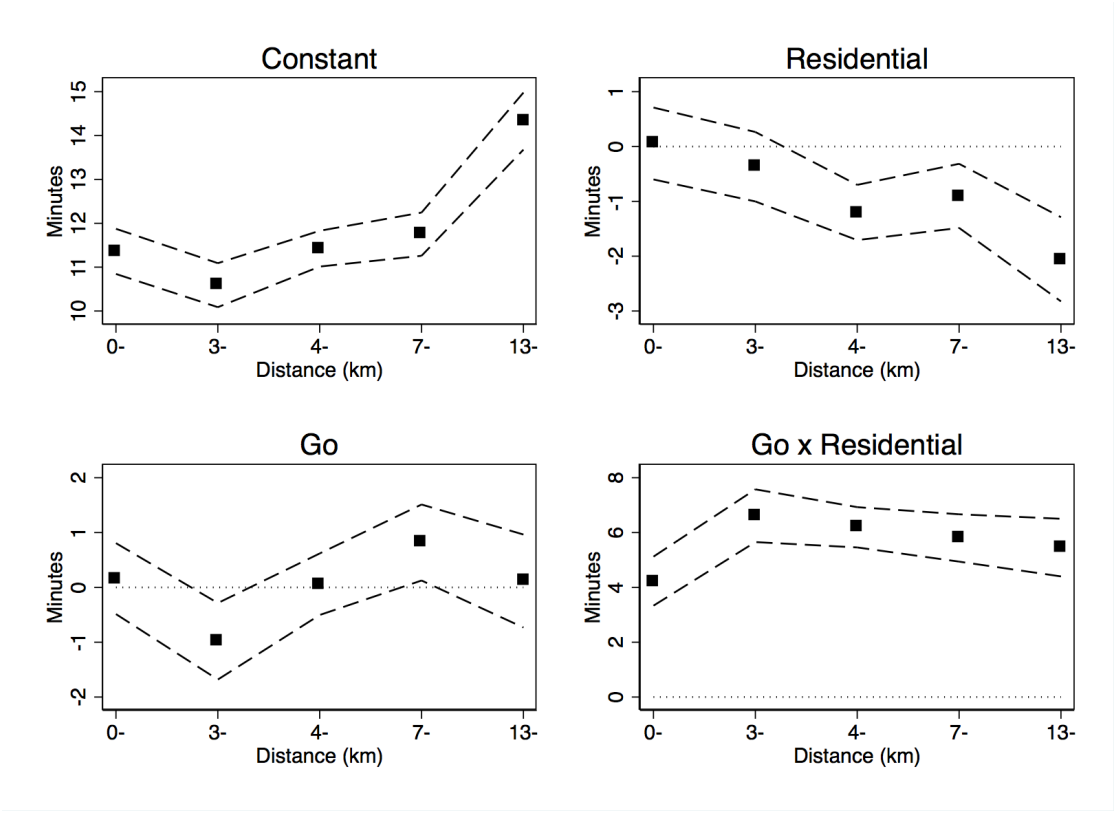
NOTES: distance is the amount of kilometres assigned to each mission.

Figure 3: Plot of the estimated parameters by distance travelled by the ambulance



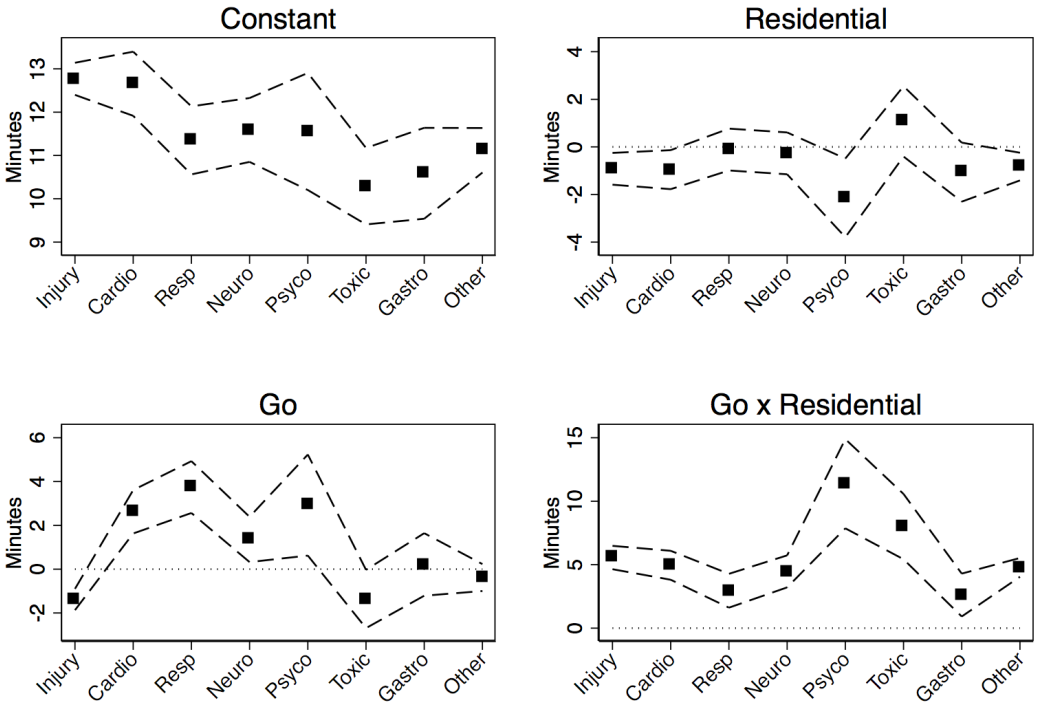
NOTES: the distance intervals are 0-4, 5-9, 10-14, 15-19 and ≥20. The estimate results are reported with the 95% confidence intervals.

Figure 4: Plot of the estimated parameters by distance travelled by the ambulance



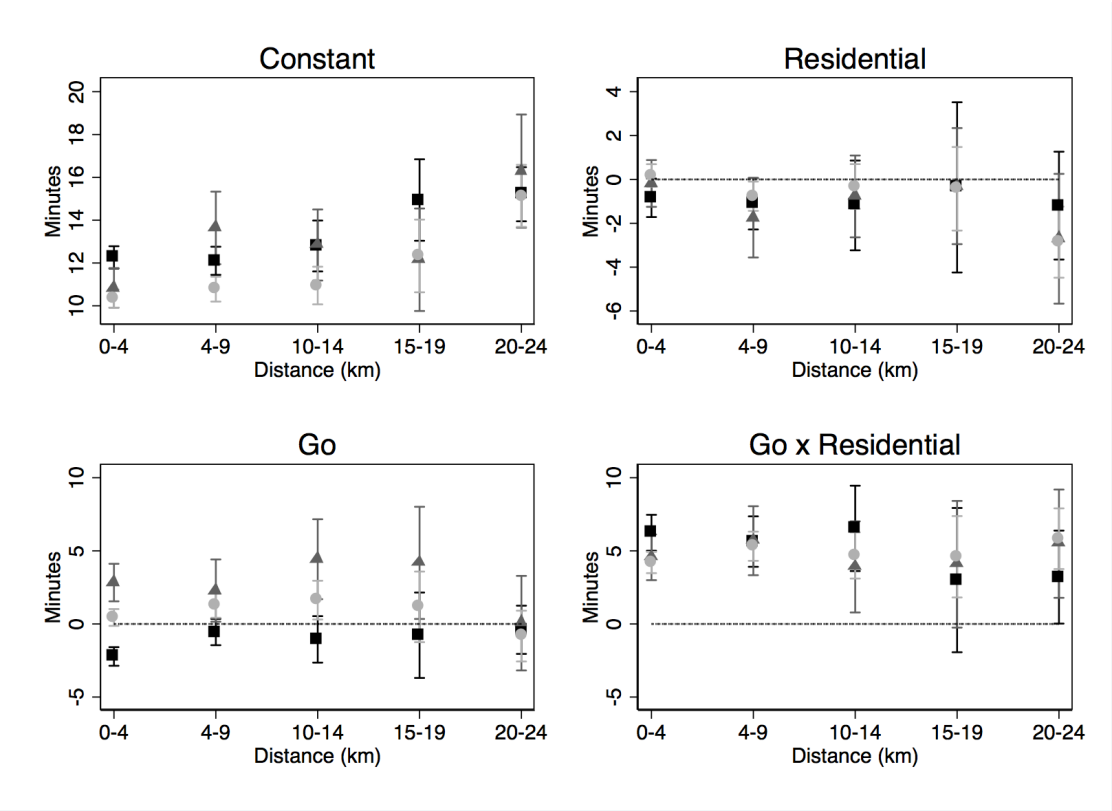
NOTES: the groups are the quintiles of the distance distribution. The estimate results are reported with the 95% confidence intervals.

Figure 5: Plot of the estimated parameters by pathology (extended pathology definition)



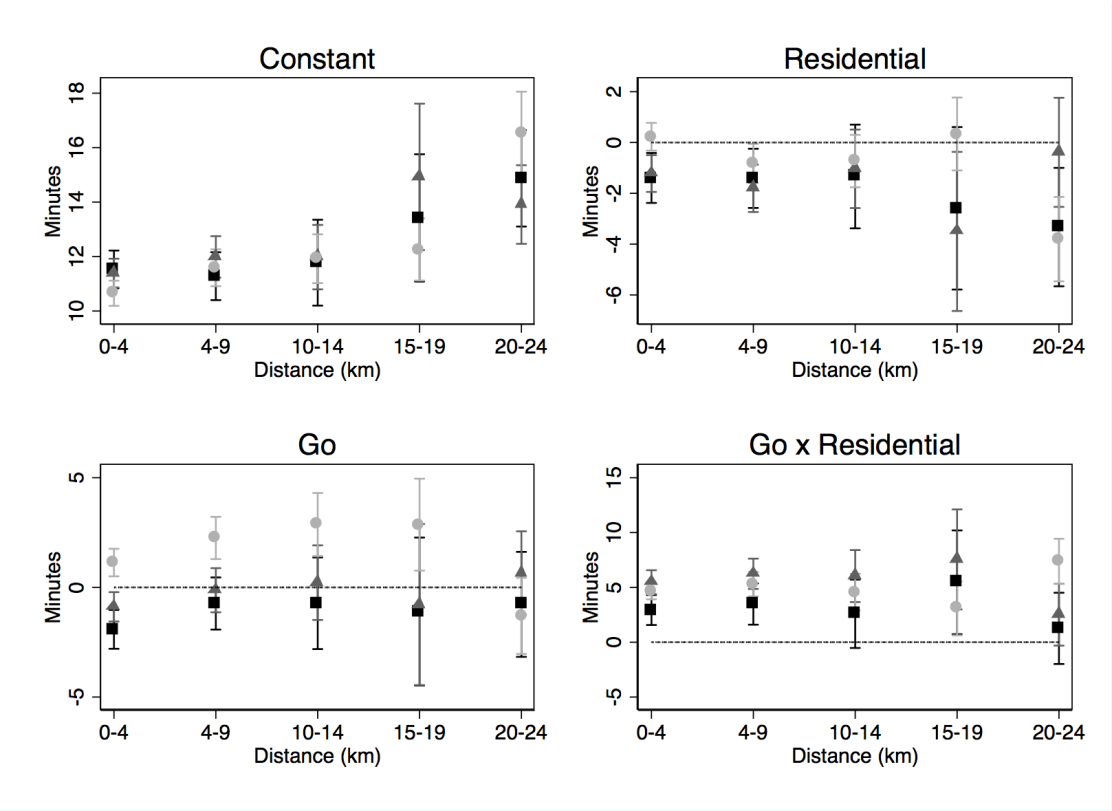
NOTES: The number of observations by type of pathology are reported in the table results, Table 7.

Figure 6: Plot of the estimated parameters by patient's pathology and distance travelled by the ambulance



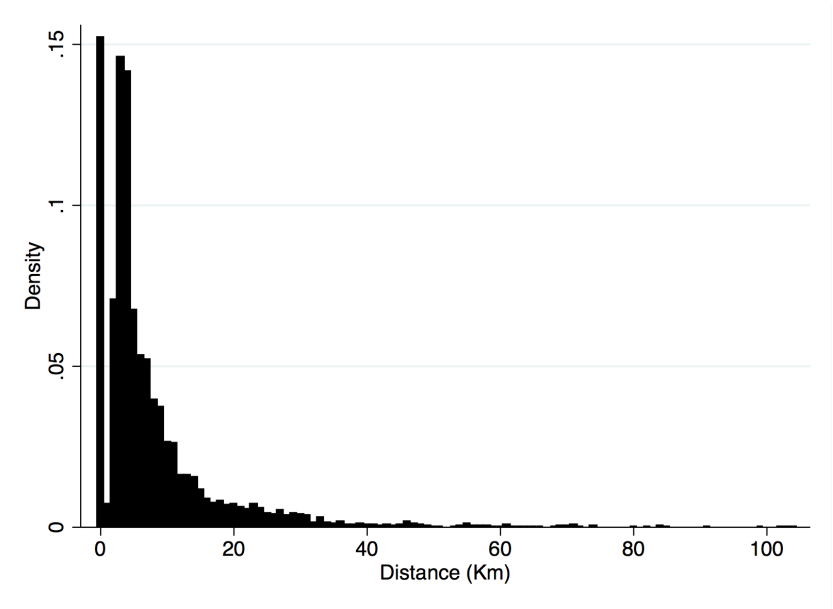
NOTES: the pathology groups are injury (square), cardiocirculatory (triangle), and other (circle). The distance intervals are 0-4, 5-9, 10-14, 15-19 and ≥20. The estimate results are reported with the 95% confidence intervals.

Figure 7: Plot of the estimated parameters by patient's age group and distance travelled by the ambulance



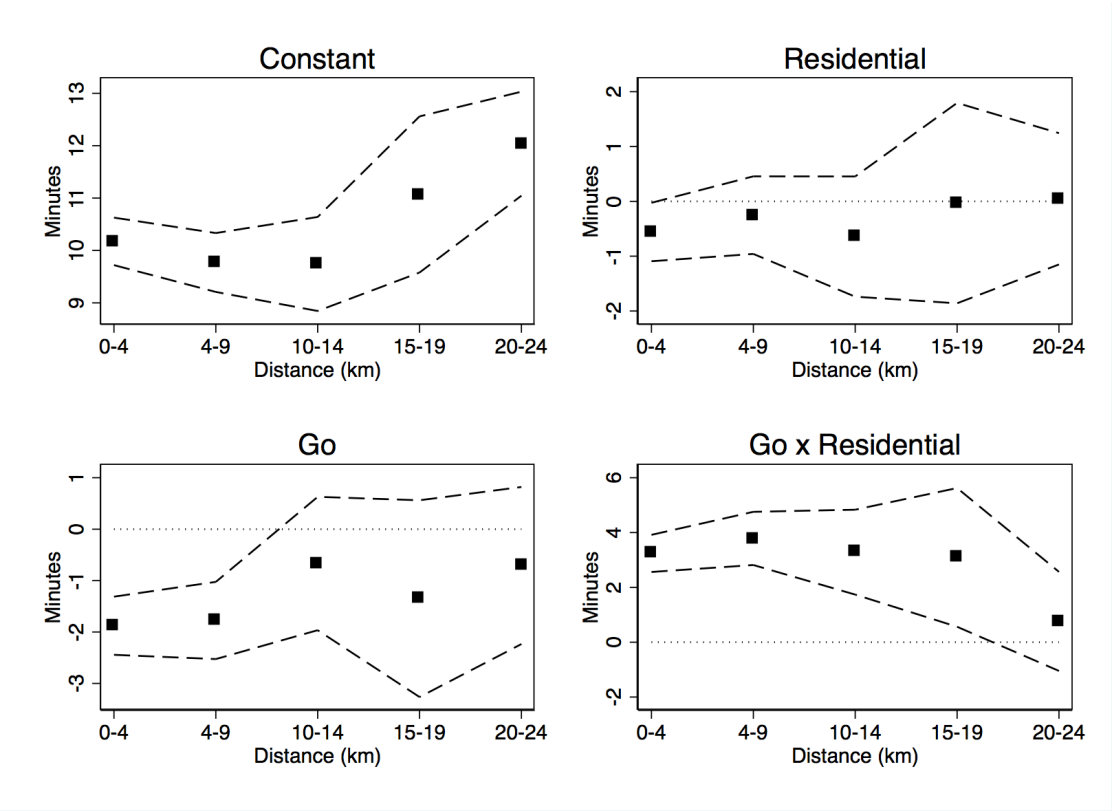
NOTES: the age groups are <35 (square), 35≥age<65 (triangle), and ≥65 (circle) years. The distance intervals are 0-4, 5-9, 10-14, 15-19 and ≥20. The estimate results are reported with the 95% confidence intervals.

Figure 8: Distribution of distance, non-urgent (deferrable) calls



NOTES: distance is the amount of kilometres assigned to each mission.

Figure 9: Plot of the estimated parameters by distance travelled by the ambulance, non-urgent (deferrable) calls



NOTES: the distance intervals are 0-4, 5-9, 10-14, 15-19 and ≥20. The estimate results are reported with the 95% confidence intervals.

Table 1: Sample Definition

Sample description or step	Observations
Raw mission data	646574
Missions where origin and final destination are the same	44727
Drop missing origin and final destination	43460
Drop low priority ambulance missions and out-of-hospital deaths	30943
Drop missing driving times and and extreme values (1- or 120+)[1]	23179
Drop missing distance (expected or imputed)	20352
Drop missing patient's age or pathology	18881

NOTES: This table describes each step in sample construction. [1] I eliminates also driving time shorter than 1 minute and longer that 120 minutes. This accounts for about 7% of the final sample size. I exclude those extreme values because are far from what may be the effect of a relatively standard mission important for the results presented in this work. The exclusion of these observations, however, does not particularly influence the results and only reduces the noise. The estimation results are statistical significant with or without the inclusion of these outliers. Robust standard errors in parenthesis.

Table 2: Descriptive Statistics

	Full sample		Working sample	
Variables	Mean	Std. Dev.	Mean	Std. Dev.
	descr10		descr1	
	mean	sd	mean	sd
Driving time go	15.2	(10.9)	15.0	(10.5)
Driving time back	12.5	(11.9)	11.4	(9.4)
% Residential	61.9	(48.6)	62.2	(48.5)
% Non Residential	38.1	(48.6)	37.8	(48.5)
Distance (Km)	17.5	(22.5)	8.9	(11.3)
% Injury	23.0	(42.1)	20.9	(40.6)
% Cardio	14.5	(35.2)	18.2	(38.6)
% Other	62.5	(48.4)	60.9	(48.8)
% Age (<35)	16.7	(37.3)	14.2	(34.9)
% Age (35-64)	26.2	(44.0)	27.8	(44.8)
% Age (≥ 65)	57.1	(49.5)	58.0	(49.4)
% High	52.8	(49.9)	81.1	(39.1)
% Medium	39.8	(49.0)	18.9	(39.1)
% Low	7.4	(26.1)	0.0	(1.6)
% Monday	14.5	(35.3)	15.5	(36.2)
% Tuesday	14.4	(35.1)	14.6	(35.3)
% Wednesday	14.3	(35.1)	14.2	(34.9)
% Thursday	14.2	(34.9)	14.3	(35.0)
% Friday	14.4	(35.1)	14.2	(34.9)
% Saturday	14.1	(34.8)	13.5	(34.2)
% Sunday	14.0	(34.7)	13.7	(34.4)
Observations	196758		18881	

NOTES: columns (1) and (2) report respectively mean and standard deviation for the full sample, that is all non-deferrable missions. Columns (3) and (4) show mean and standard deviation for the working sample, which considers only the non-deferrable missions in which the origin of the ambulance is also the place in which the patient is transported to at the end of the mission. Robust standard errors in parenthesis.

Table 3: Correlation between rain and driving time by way

Panel A: Full Sample		
Driving time	Go	Back
	(1)	(2)
Rain	0.4906***	0.1306*
	(0.0729)	(0.0782)
Observations	196674	196674
Panel B: Working Sample		
Rain	0.4809**	-0.0609
	(0.2260)	(0.1960)
Observations	18879	18879

NOTES: rain is a dummy equal to one for the missions performed during rainfall. Weather conditions vary at the hour and municipality level. Panel A reports the results for the full sample, panel B for the working sample, that is restricted to the missions where origin and final destination of the ambulance is the same. Column (1) reports the results for the driving time to go, column (2) for the driving time back. Robust standard errors in parenthesis.

Table 4: Results by distance travelled by the ambulance (5 kilometres interval)

	(1)	(2)	(3)	(4)	(5)
Go	-0.31	0.69**	1.22**	0.87	-0.54
	(0.21)	(0.31)	(0.50)	(0.87)	(0.58)
Residential	-0.48**	-1.12***	-0.83**	-1.19*	-2.51***
	(0.20)	(0.27)	(0.40)	(0.69)	(0.57)
Go × Residential	5.36***	6.09***	5.54***	5.13***	5.40***
	(0.28)	(0.39)	(0.62)	(1.04)	(0.74)
Constant	11.13***	11.66***	11.91***	13.37***	15.33***
	(0.16)	(0.22)	(0.34)	(0.58)	(0.47)
Distance (Km)	<5	5-9	10-14	15-19	≥20
Observations	16574	10474	4480	2172	4062

NOTES: the sample is restricted to the missions where origin and final destination of the ambulance coincide. Each column reports the results by distance driven by the ambulance. The kilometres intervals, reported also on the bottom of the table, are respectively: 0-4, 5-9, 10-14, 15-19 and ≥20. Robust standard errors in parenthesis.

Table 5: Results by distance travelled by the ambulance (by distribution quantile)

	(1)	(2)	(3)	(4)	(5)
Go	0.16	-0.98***	0.06	0.82**	0.12
	(0.33)	(0.36)	(0.29)	(0.35)	(0.43)
Residential	0.06	-0.37	-1.20***	-0.90***	-2.06***
	(0.33)	(0.32)	(0.26)	(0.30)	(0.39)
Go × Residential	4.23***	6.61***	6.19***	5.80***	5.45***
	(0.46)	(0.49)	(0.38)	(0.44)	(0.54)
Constant	11.36***	10.59***	11.42***	11.75***	14.33***
	(0.26)	(0.26)	(0.21)	(0.25)	(0.33)
Distance (Km)	0-2	3	4-6	7-12	≥13
Observations	7128	4696	9770	8558	7610

NOTES: the sample is restricted to the missions where origin and final destination of the ambulance coincide. Each column reports the results by distance driven by the ambulance. The kilometres intervals by quantile, reported also on the bottom of the table, are respectively: 0-2, 3, 4-6, 7-12 and ≥13. Robust standard errors in parenthesis.

Table 6: Results by age group and gender

	(1)	(2)	(3)	(4)	(5)
Go	-1.32***	-0.41	1.37***	0.25	0.00
	(0.33)	(0.27)	(0.25)	(0.24)	(0.23)
Residential	-1.68***	-1.34***	-0.73***	-0.88***	-1.04***
	(0.34)	(0.27)	(0.21)	(0.21)	(0.21)
Go × Residential	3.04***	5.65***	5.20***	5.11***	6.38***
	(0.50)	(0.40)	(0.29)	(0.29)	(0.30)
Constant	11.95***	12.04***	11.91***	11.76***	12.14***
	(0.25)	(0.20)	(0.18)	(0.18)	(0.17)
Group	<35	35-64	≥65	Female	Male
Observations	5348	10510	21904	18886	17818

NOTES: the sample is restricted to the missions where origin and final destination of the ambulance coincide. Each column reports the results by age group (columns 1-3) and gender of the patient (columns 4-5). The age groups are <35, 35≥age<65, and ≥65 years. Robust standard errors in parenthesis.

Table 8: Results by distance travelled by the ambulance (5 kilometres interval). Non-urgent (deferrable) calls

	(1)	(2)	(3)	(4)	(5)
Go	-1.8781***	-1.7761***	-0.6679	-1.3514	-0.7068
	(0.2885)	(0.3830)	(0.6618)	(0.9764)	(0.7793)
Residential	-0.5592**	-0.2515	-0.6424	-0.0316	0.0468
	(0.2719)	(0.3612)	(0.5595)	(0.9324)	(0.6111)
Go × Residential	3.2407***	3.7869***	3.2860***	3.0952**	0.7597
	(0.3456)	(0.4950)	(0.7899)	(1.2913)	(0.9217)
Constant	10.1735***	9.7708***	9.7423***	11.0694***	12.0381***
	(0.2311)	(0.2868)	(0.4578)	(0.7602)	(0.5061)
Distance (Km)	<5	5-9	10-14	15-19	≥20
Observations	10084	4870	1954	838	1746

NOTES: the sample is restricted to the missions where origin and final destination of the ambulance coincide. Each column reports the results by distance driven by the ambulance. The kilometres intervals, reported also on the bottom of the table, are respectively: 0-4, 5-9, 10-14, 15-19 and ≥20. Robust standard errors in parenthesis.

Table 7: Results by pathology of the patient (extended)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Go	-1.39***	2.62***	3.75***	1.36**	2.93**	-1.36**	0.22	-0.38
	(0.25)	(0.51)	(0.60)	(0.53)	(1.18)	(0.68)	(0.73)	(0.31)
Residential	-0.92***	-0.96**	-0.11	-0.27	-2.14**	1.07	-1.06*	-0.83***
	(0.34)	(0.42)	(0.45)	(0.45)	(0.84)	(0.75)	(0.63)	(0.30)
Go × Residential	5.56***	4.95***	2.95***	4.47***	11.38***	8.00***	2.60***	4.79***
	(0.47)	(0.58)	(0.68)	(0.64)	(1.79)	(1.31)	(0.86)	(0.38)
Constant	12.77***	12.66***	11.35***	11.59***	11.55***	10.29***	10.59***	11.12***
	(0.19)	(0.38)	(0.40)	(0.38)	(0.69)	(0.45)	(0.54)	(0.26)
Pathology	Injury	Cardio	Respiratory	Neuro	Psyco	Toxic	Gastro	Other
Observations	7880	6880	4526	3960	1142	1024	1502	10848

NOTES: the sample is restricted to the missions where origin and final destination of the ambulance coincide. Each column reports the results by pathology. The pathology is reported on the bottom of the table. Robust standard errors in parenthesis.

Table 9: Results by distance travelled by the ambulance (by distribution quantile). Non-urgent (deferrable) calls

	(1)	(2)	(3)	(4)	(5)
Go	-2.7712***	-1.3140***	-1.8710***	-1.6336***	-0.9269*
	(0.7315)	(0.3807)	(0.4156)	(0.4032)	(0.4835)
Residential	-1.5010**	-0.0652	-0.6889*	-0.3151	-0.4943
	(0.6999)	(0.3540)	(0.3720)	(0.3790)	(0.4131)
Go × Residential	3.1961***	2.8733***	3.8617***	3.9645***	2.2321***
	(0.8134)	(0.4720)	(0.5313)	(0.5156)	(0.5871)
Constant	12.4028***	9.1105***	9.9497***	9.5836***	11.2512***
	(0.6398)	(0.2922)	(0.2891)	(0.3051)	(0.3389)
Observations	3100	4226	4076	4068	4022

NOTES: the sample is restricted to the missions where origin and final destination of the ambulance coincide. Each column reports the results by quantile of distance distribution driven by the ambulance. Robust standard errors in parenthesis.

Table 10: Other results: difference-in-differences by time of the day, traffic conditions, severity of patient conditions, and a placebo.

	(1)	(2)	(3)	(4)	(5)	(6)
Go	-0.01	0.54	-0.22	0.50**	-2.83***	-0.06
	(0.17)	(0.38)	(0.21)	(0.24)	(0.17)	(0.22)
Residential (Res)	-0.62***	-1.40***	-1.17***	-0.62***	-1.60***	
	(0.17)	(0.29)	(0.19)	(0.22)	(0.14)	
Go × Residential	5.00***	6.74***	6.06***	5.02***	4.37***	
	(0.23)	(0.44)	(0.26)	(0.32)	(0.22)	
Go × gourgent					3.43***	
					(0.17)	
Res × Urgent					1.18***	
					(0.13)	
Go × Res × Urgent					0.76***	
					(0.26)	
Dummy						0.02
						(0.24)
Go × Dummy						0.29
						(0.32)
Constant	12.02***	11.52***	12.07***	11.74***	11.41***	11.92***
	(0.13)	(0.26)	(0.16)	(0.18)	(0.10)	(0.16)
	Day	Night	No traffictime	Traffictime	DDD	Placebo
Observations	28404	10440	24080	14764	57476	14710

NOTES: column (1) shows the effect of interest for the only missions perfromed during the daylight; column (2) reports the results where the observations are restricted to the missions performed from sunset to sunrise; columns (3) and (4) illustrate, respectively, the results for missions performed during traffic time and other times; column (5) reports the triple difference with also the severity of the patient; column (6) reports the results of the placebo test. Robust standard errors in parenthesis.

Annexes

Table 11: Results by patient's age group and distance travelled by the ambulance

Panel A: Age < 35					
Distance (Km)	<5	5-9	10-14	15-19	≥20
	(1)	(2)	(3)	(4)	(5)
Go	-2.67***	-1.17*	-2.16*	-1.23	-0.83
	(0.58)	(0.70)	(1.16)	(2.37)	(1.51)
Residential	-1.68***	-0.16	0.33	3.19	1.16
	(0.57)	(0.76)	(1.48)	(3.03)	(1.37)
Go × Residential	2.32***	1.25	3.04	-2.70	-1.45
	(0.74)	(1.05)	(1.88)	(3.53)	(2.01)
Constant	10.37***	9.11***	9.31***	10.06***	10.88***
	(0.44)	(0.52)	(0.86)	(1.96)	(1.03)
Observations	1956	932	328	124	276
Panel B: Age 35 - 64					
Go	-2.29***	-2.23***	-1.00	-5.78***	-2.45**
	(0.41)	(0.63)	(1.27)	(2.05)	(1.24)
Residential	-0.59	0.16	-1.81*	-2.39	-2.44**
	(0.42)	(0.60)	(1.04)	(2.11)	(1.10)
Go × Residential	2.94***	2.69***	2.52	9.20***	1.95
	(0.55)	(0.84)	(1.54)	(2.82)	(1.47)
Constant	9.44***	9.20***	10.16***	12.20***	12.64***
	(0.33)	(0.44)	(0.82)	(1.95)	(0.94)
Observations	3128	1536	608	212	464
Panel C: Age ≥ 65					
Go	-0.64	-1.74***	0.90	0.85	0.64
	(0.54)	(0.65)	(0.92)	(1.19)	(1.26)
Residential	-0.80	-1.29**	-0.29	0.21	0.67
	(0.50)	(0.61)	(0.78)	(0.99)	(0.86)
Go × Residential	2.76***	5.01***	2.64**	1.45	0.11
	(0.59)	(0.78)	(1.05)	(1.53)	(1.42)
Constant	10.99***	11.03***	9.63***	10.83***	12.12***
	(0.46)	(0.54)	(0.69)	(0.76)	(0.72)
Observations	5000	2402	1018	502	1006

NOTES: the sample is restricted to the missions where origin and final destination of the ambulance coincide. Panel A reports the results for patients younger than 35, panel B for patients 35-64 year old, and panel C for 65 year or older. All the results are presented by distance travelled by the ambulance. Robust standard errors in parenthesis.

Table 12: Results by patient's pathology and distance travelled by the ambulance

Panel A: Injury					
Distance (Km)	<5	5-9	10-14	15-19	≥20
	(1)	(2)	(3)	(4)	(5)
Go	-2.48***	-2.05***	-1.40	0.01	-1.13
	(0.43)	(0.59)	(0.91)	(1.14)	(1.07)
Residential	-0.28	-1.24**	-0.02	2.80**	2.24*
	(0.46)	(0.58)	(1.07)	(1.42)	(1.17)
Go × Residential	5.44***	5.74***	4.80***	2.24	-0.05
	(0.58)	(0.80)	(1.32)	(2.06)	(1.55)
Constant	10.41***	10.08***	10.23***	9.91***	11.66***
	(0.36)	(0.46)	(0.74)	(0.66)	(0.67)
Observations	3044	1506	596	280	506
Panel B: Cardiocirculatory					
Go	-0.32	-2.58	4.11***	1.68	-7.86
	(2.07)	(2.27)	(1.25)	(7.70)	(8.45)
Residential	1.86	-2.39	0.58	-3.04	-7.26
	(2.00)	(2.33)	(1.36)	(5.61)	(8.57)
Go × Residential	0.54	8.61***	-2.41	-1.63	11.63
	(2.50)	(2.91)	(2.23)	(9.15)	(9.04)
Constant	9.40***	10.45***	9.72***	14.18***	18.67**
	(1.58)	(2.07)	(0.44)	(3.14)	(8.39)
Observations	184	82	50	24	56
Panel C: Other					
Go	-1.48***	-1.50***	-0.11	-2.60*	0.02
	(0.39)	(0.51)	(0.97)	(1.49)	(1.07)
Residential	-0.63*	0.29	-0.62	-1.49	-0.41
	(0.35)	(0.46)	(0.67)	(1.39)	(0.71)
Go × Residential	2.34***	2.89***	2.50**	4.19**	0.24
	(0.45)	(0.64)	(1.09)	(1.78)	(1.20)
Constant	10.03***	9.47***	9.30***	11.80***	11.98***
	(0.31)	(0.37)	(0.57)	(1.28)	(0.62)
Observations	6856	3282	1308	534	1184

NOTES: the sample is restricted to the missions where origin and final destination of the ambulance coincide. Panel A reports the results for injured patients, panel B for patients affected by cardiocirculatory problems, and panel C for others pathologies. All the results are presented by distance travelled by the ambulance. Robust standard errors in parenthesis.

Chapter 3

Slacking-off or just tired?
Work-shift and workers productivity in the Emergency Department

Elena Lucchese[+], Paolo Roberti[*]
[+] *University of Bologna;* [*] *University of Bergamo*

Abstract

Work schedules is a general form of coordination in organizations. In this study of emergency medical treatments dispensed by care providers working in shifts, schedules induce distortions near the end of shift. Examining how performance changes according to the level of urgency of the patient, we are able to distinguish between leisure (slacking-off) and fatigue motives as potential drivers. Fatigue appears to be an important determinant of performance.

KEYWORDS: shift work, scheduling, slacking-off, fatigue, performance

1 Introduction

The number of workers employed in the provision of personal services have substantially increased over time. The set of information involved to deliver this type of services is, in many instances, continuous and multidimensional, and so it is difficult to draft a contract that precisely defines all the tasks involved. As a result, works such as physician, lawyer, teacher, consultant, or software engineer are often ruled by incomplete contracts: the principal-agent problem first modeled by Hart and Holmstrom (1987). In such setting, intrinsic motivation is particularly relevant, and a now-substantial literature recognizes the key role played by the 'mission' that workers attribute to their job.[1]

Recent literature exploits the emergency medical setting – a context in which contracts are incomplete, workers are intrinsically motivated and the activity is scheduled – to gain deeper insights about the determinants of workers productivity. In addition to that, work-schedule – a general form of coordination in organizations – is often associated with a decrease in worker performance when the end of shift (EOS) is approaching. This might be driven by preferences to enjoy leisure time against working overtime, as proposed by Chan (2018), or by fatigue, in particular at the end of longer shifts, as analyzed by Brachet et al. (2012). However, the effects of either leisure or fatigue on workers performance are in general observationally equivalent. The aim of this work is to build on this literature by proposing a theoretical setup with testable implications and a related empirical analysis to distinguish between leisure and fatigue motives.

Distinguishing between leisure and fatigue is important also in terms of policy, as the implications of the two mechanisms might suggest different sets of actions. For instance, if the decrease in performance near EOS in driven by fatigue, a policy meant to alleviate this decrease in performance would increase the number of shifts in order to lower the number of hours worked per shift. If, instead, this change is due to slacking-off behavior, the total number of shifts should diminish.

In the setup of this paper, health providers care about patients' condition: a common assumption in this framework. In a world in which workers do not value leisure or are not affected by fatigue, their performance near EOS would not change. In the presence of either leisure or fatigue, performance decreases. To distinguish between the two, we look at the degree of motivation and effort required to perform the job proxied by severity of patients (triage). When patients arrive to the Emergency Department they provide basic details about their symptoms and are assigned with a triage code that ranks the severity of their health condition. The care providers do not observe the exact nature of the problem that affects the patients, but know the triage. Low triage code refers to patients that are not in imminent risk of dying, and treating them is in general simpler than higher

[1] Seminal work by Tirole (1986) introduced the role of intrinsic and extrinsic motivation on worker's productivity.

triages. High triage codes are at risk of death and typically require more urgent clinical responses. Medium triages are not in imminent risk of dying, and their diagnoses is usually more complex than low priority patients.

The theory delivers two implications which can help distinguish between leisure and fatigue motives. First, a negative relationship between performance near EOS and lower triage is evidence of leisure motives; a non-monotonic association between the two suggests that providers are tired near EOS. Indeed, whether providers seek leisure or are tired, they always visit high triage patients immediately, as making them wait might have serious adverse consequences on their health. When they seek leisure, they 'slack-off' on lower degrees of urgency, because visiting them increases the probability of working overtime. If they are tired, they still visit low triage patients near EOS as the marginal effort required to take care of them is relatively small. Instead, the decrease in performance of tired physicians should concentrate on medium triage patients because they are not at risk of death and treating them is relatively more demanding. A second implication of the model is that providers driven by increasing fatigue visit fewer patients near EOS when they had faced a large in-flow of medium and high triage patients in the Emergency earlier in the shift. No such relationship should be observed if providers are driven by leisure motives only.

We test the two predictions of the model using an administrative dataset that collects information on all Emergency Department (ED) accesses in the Italian region Liguria over three years. Empirical evidence suggests that fatigue motives is the main driver of lower performance near end-of-shift.

The remainder of the paper is organized as follows. Section 3 sets out the theoretical model. Section 4 describes the data. Section 5 illustrates the first descriptive results.

2 Institutional setting: Emergency care in Liguria

Emergency care in Liguria is publicly funded and the service is provided for free at the point of use for all appropriate accesses, i.e. non-deferrable.[2] Private providers are not involved in emergency care. The majority of care is provided at Emergency Departments (EDs) attached to large, publicly owned hospitals. These major emergency departments are physician-led providers of 24-hour services, based in specifically built facilities to treat emergency patients. Over the period of observation, more than 750,000 patients were responsible for about 1.3 million visits to

[2]Non appropriate accesses are charged with a ticket of 25 euros, but some categories, namely pregnant women, kids under the age of 14, disabled people, and those with lower income, are exempted from the payment of the ticket.

17 emergency departments.[3]

Treatment in the ED follows one of two pathways depending on the method of arrival. Non-ambulance patients register at reception upon arrival, where they provide basic details about their condition. Patient then undergo an initial assessment to establish the urgency of their condition. This triage process is carried out either by a specialist triage nurse or doctor, and includes taking a medical history, and, where appropriate, conducting a basic physical examination of the patient. Patients are then prioritized according to the urgency of their health condition.

Alternatively, patients can arrive at the ED by ambulance following an emergency call out. In 2013/2015, 26% of ED patients arrived by ambulance. For these patients, ambulance staff collect medical details en route, and report these details to hospital staff upon arrival.[4] This information feeds into a separate triage process, where patients will be categorized by urgency.

These triage process sort patients into 'minor', 'medium', and 'major' cases. Minor cases (white codes) require relatively simple treatment, and can mainly be treated in a short time-span. The length of treatment is the difference between the time of discharge and the time of the first visit performed by care providers; the average duration for minor cases is 50 minutes. Medium cases (green and yellow codes, where yellow indicates greater severity) present with more severe symptoms and usually require more intense and accurate treatment and investigations within the ED, and are therefore likely to spend longer time in the ED (on average, 144 minutes). Major cases (red codes) present patients in imminent risk of death (average treatment lasts for 253 minutes).

Following triage, patients are placed into a queue on the basis of their severity and time of arrival. Patients are not aware of their position in the queue. Patients are assigned to individual doctors as they become available. These doctors will carry out a series of further examinations and tests. The nature of these investigations depend on the symptoms presented by the patients, and range from physical examinations to tests such as x-rays or MRI scans. Patients can also receive treatment in the ED, ranging from sutures to resuscitation, before being admitted for further treatment in an inpatient ward, or discharged from the hospital.

Staffing level in the ED varies during day and night time to reflect patient volume. The work shifts do not overlap, and the time of shifts is constant over time in the period observed, being 8am/2pm; 2pm/8pm; 8pm/8am. The medical team works, in rotation, in all shifts.

[3] In addition, 26,000 patients made an additional 35 thousands visits to minor emergency clinics and 'walk in' where simple treatment is provided and that are opened only during day time; as discussed below, we exclude patients from these centers due to the different organizational structure that rules these departments.

[4] Ambulance staff also provide emergency treatment on the scene of the event and in the ambulance to patients when required.

3 Theory

This section develops a theory that helps understanding how different incentives of care providers shape their work decisions in the Emergency Department. In particular the objective of the theory is to provide testable implications to disentangle leisure and fatigue motives. Let us consider a simple model where decision maker (DM_t)[5] has to choose the flow n of patients to visit during her shift t. This decision has to be taken in $I+1$ intervals indexed by $i \in \{0,...,I\}$, where the beginning of shift is denoted by $i=0$. In interval i of shift t the number of visited patients is n_{ti}. All the patients not visited by the DM have to be visited by the DM working in the following shift. The DM starts her shift with a stock s_{t-1I} of patients in the ED, and in every interval faces a stock of patients s_{ti}, where the law of motion of the stock of patients is

$$s_{ti} = (1-p)\left(s_{ti-1} - n_{ti-1}\right) + f_{ti} + \varepsilon_{it}, \qquad s_{t,-1} = s_{t-1,I}, \qquad s_{t,I+1} = s_{t+1,0}.$$

f_i is the expected in-flow of patients in the ED in each interval, p is the share of patients who are not present in the following interval, either because they leave the ED voluntarily, or because they die, ε_{it} are i.i.d. random shocks on the flow of patients with probability density function $g(\cdot)$ with mean 0, and cumulative distribution function $G(\cdot)$ on the support $S \subseteq \mathbb{R}$. The utility gain of the DM from visiting more patients is to positively influence the health of patients. In particular, to the extent that the decision maker is intrinsically motivated, she would like each patient to be visited as soon as possible.[6] Hence she suffers a loss when less patients than the full stock are visited in every period: $-\frac{1}{2(1-k)}\sum_{w\geq t, i\leq I}\left(s_{wi} - n_{wi}\right)^2$, where k parameterizes how much the decision maker cares about the patients' health and is positively linked to the severity of patients.[7] Notice that the decision maker cares about the health of all patients that enter the ED, not only about the patients in her shift.

We consider two possible mechanisms driving the costs of visiting patients, namely *leisure* and *fatigue*. If the decision maker has leisure motives, she considers the leisure time l that she can enjoy instead of working overtime, where the probability of staying overtime q is an increasing function of the number of patients visited by the DM in the last interval of the shift: $\frac{\partial q}{\partial n_{tI}} > 0$. In particular $q(n_{tI}) = \frac{n_{tI}}{\bar{n}}$, where $\bar{n}$ is an exogenous parameter conveniently chosen in order for $n_{tI} \leq \bar{n}$.

[5] The decision maker can be thought as the head of the Emergency Department who internalizes the utilities of the physicians in the Emergency Department. Alternatively one can assume that physicians are sufficiently homogeneous in their utility function with respect to the relevant parameters of this model, so that the decision maker is a representative physician.

[6] Rationales behind this intrinsic motivation are that visiting patients prevents a deterioration of (and can improve) their health, and relatedly, decreases the pain or anxiety of patients.

[7] This issue will be further developed in the next section.

If the decision maker suffers from increasing fatigue, as time i goes by visiting additional patients becomes more costly. Similarly, the stock of visited patients $\sum_{j<i} n_{tj}$ makes new visits more costly : $e(n_{ti}) = (\lambda i)\left(\sum_{j<i} n_{tj}\right) n_{ti}$, $\lambda < 1$. Parameter λ measures how strongly an increase in either time or having visited additional patients affects the cost of visiting patients.

The decision maker working in shift t overall utility is:

$$-\frac{1}{2(1-k)} \sum_{w \geq t, i \leq I} \delta^{(w-t)I+i} (s_{wi} - n_{wi})^2 - \sum_{i=0}^{I} \delta^i (\lambda i) \left(\sum_{j<i} n_{tj} \right) n_{ti} + \delta^I \left(1 - \frac{n_{tI}}{\bar{n}}\right) l,$$

where δ is the discount factor of DM_t. When $l > 0$ and $\lambda = 0$, the only mechanism driving the decision maker choices is the leisure motive. If instead $l = 0$ and $\lambda > 0$ the single driver of agent behavior is fatigue.[8]

Let us consider a benchmark model where neither leisure nor fatigue are present: $l = 0$ and $\lambda = 0$. In each time interval $i \in \{0, ..., I\}$, the first order condition of the decision maker problem is

$$s_{ti} - n_{ti} + \sum_{j>i} [\delta(1-p)]^{j-i} \mathbb{E}_{\varepsilon_{tj}} \left[s_{tj} - n_{tj} \right] + \sum_{w>t, j \leq I} [\delta(1-p)]^{I-i+(w-t-1)I+j} \mathbb{E}_{\varepsilon_{wj}} \left[s_{wj} - n_{wj} \right] = 0. \tag{1}$$

The first order (1) condition shows that without leisure and fatigue motives, the decision maker cares about the difference between stocks and visited patients in the interval in consideration and, with a decreasing magnitude as measured by δ and $1-p$, about these differences in all future intervals. Hence, the optimal number of patients visited in each interval of time by a DM without leisure nor fatigue motives is

$$n^*_{ti} = f_{ti} + \varepsilon_{ti}.$$

We can therefore state the following.

Proposition 1 (Benchmark)**.** *Absent leisure motives and fatigue, the decision maker in each interval visits in expected terms a number of patients equal to the in-flow of patients in the ED:* $\mathbb{E}_{\varepsilon_{ti}} [n^*_{ti}] / f_{ti} = 1$.

Let us now compare the benchmark with the leisure and fatigue motives. When the DM is driven only by leisure motives, the first order condition for the number n^*_{it} of patients visited in

[8] We do not discuss the situation in which both l and λ are strictly positive because the main purpose in this paper is to explore if there is a setting in which the implications of the two different mechanisms are not observationally equivalent. Given this main aim, allowing for both mechanisms to exist simultaneously is not appropriate. However the more general context might be empirically relevant, and we plan to discuss this more general context in future research.

each interval before I is

$$s_{ti} - n_{ti} + \sum_{j>i} [\delta(1-p)]^{j-i} \mathbb{E}_{\varepsilon_{tj}} \left[s_{tj} - n_{tj} \right] + \sum_{w>t, j \leq I} [\delta(1-p)]^{I-i+(w-t-1)I+j} \mathbb{E}_{\varepsilon_{wj}} \left[s_{wj} - n_{wj} \right] = 0. \tag{2}$$

These first order conditions are equal to the ones that solve the benchmark model, as the leisure motive enters only in the last interval of the shift. The first order condition for n^*_{tI} is

$$s_{tI} - n_{tI} + \sum_{w>t, j \leq I} [\delta(1-p)]^{I-i+(w-t-1)I+j} \mathbb{E}_{\varepsilon_{wj}} \left[s_{wj} - n_{wj} \right] - (1-k)\frac{1}{\bar{n}} l = 0. \tag{3}$$

An increase in the number of visited patients in the last interval of the shift decreases the distance between stocks and the number of visited patients which positively affects the utility of the DM, but it also increases the probability of working overtime, which decreases the utility of the DM.

If the DM only suffers from increasing fatigue, the first order condition for the number n^*_{it} of patients visited in interval i is

$$\begin{aligned} s_{ti} - n_{ti} + \sum_{j>i} [\delta(1-p)]^{j-i} \mathbb{E}_{\varepsilon_{tj}} \left[s_{tj} - n_{tj} \right] + \sum_{w>t, j \leq I} [\delta(1-p)]^{I-i+(w-t-1)I+j} \mathbb{E}_{\varepsilon_{wj}} \left[s_{wj} - n_{wj} \right] \\ -(1-k)\lambda i \sum_{j<i} n_{tj} - (1-k)\lambda \sum_{j>i} \delta^{j-i} j \mathbb{E}_{\varepsilon_{tj}} \left(n_{tj} \right) = 0. \end{aligned} \tag{4}$$

As compared to the benchmark, in equation (4) an increase in the number of visited patients in the interval of interest negatively affects the utility of the decision maker because it increases the fatigue of the DM in the interval of interest. Moreover this negative effect increases with i, the number of intervals from the beginning of the shift. Finally n_{ti} also increases the fatigue in all following intervals in the same shift.

Analyzing both mechanisms, as compared to the benchmark, in the last interval of her shift the decision maker has an additional cost from visiting more patients: if driven by leisure motives, she runs the risk of staying overtime and missing on the leisure activity. If the decision maker suffers from increasing fatigue, she is more tired, and every additional patient is more costly. Notice that in both cases, when k approaches 1, the first order conditions become equal to the ones in the benchmark, because the DM cares more about patients' health and less about either leisure or fatigue. Therefore we can state the following proposition.

Proposition 2. *For sufficiently low δ and sufficiently large p, if the decision maker has either leisure motives or suffers from increasing fatigue, and $(1-k) \neq 0, \lambda \neq 0$, in the last interval of a shift the rate of number of visited patients over expected inflow of patients is lower than the rate of the expected number of visited patients over the expected inflow of patients in the first interval of*

the next shift: $\frac{\mathbb{E}_{\varepsilon_{tI}}\left[n^*_{tI}\right]}{f_{tI}} < \frac{\mathbb{E}_{\varepsilon_{t+1,0}}\left[n^*_{t+1,0}\right]}{f_{t+1,0}}$.

This and all following proofs are in the Appendix.[9] Hence, if the decision maker was not driven by leisure nor fatigue motives there would be no difference in the rate of visited over flow of patients between the end of a shift and the beginning of the following one. However, only looking at this difference does not help disentangle whether the DM values leisure time and reduces the number of visited patients at the end of a shift accordingly, or she is tired because she has accumulated fatigue, spending time working and visiting patients.

3.1 Exploiting patients severity to test leisure and fatigue motives

Let us enrich the theory by modeling how the severity of the initial health condition of patients enters the utility function of the decision maker. This will help us disentangle leisure from fatigue motives. In this extension, patients are heterogeneous with respect to the severity (triage) of health when they enter the Emergency Department, $h \in [\underline{h}, 1]$, where $\underline{h} \geq 0$. Let us assume the following.

Assumption 1. *The decision maker cares more about the health of more severe patients:* $\frac{\partial k}{\partial h} > 0$. *In particular* $k = h$.

Assumption 2. *Taking care of a high severity patient is more costly in terms of effort:* $\frac{\partial \lambda}{\partial h} > 0$. *In particular* $\lambda = h$.[10]

Assumption (1) is based on the notion that more severe patients, if not (timely) visited, have worse health outcomes then less severe patients, e.g. if a patient is in a life threatening situation, not visiting her implies her death. Assumption (2) is instead based on the notion that more severe patients request more complex medical procedures, which increase the fatigue of the decision maker. Finally we assume the following.

Assumption 3. *The marginal contribution to the stock of fatigue of visiting patients of severity h is weighted by h:* $\sum_{h,j<i} h n^h_{tj}$.

Assumption (3) simply implies, consistently with Assumption (2), that having visited more severe patients increases more the fatigue of the decision maker, than having visited less severe

[9] This and the following Propositions have been proven for a sufficiently low δ and a sufficiently high p, because in this scenario the decision maker is less affected by dynamic considerations on the evolution of the stock of patients. These are only sufficient conditions, hence under some restrictions on other parameters the Propositions are likely to hold also for $\delta = 1$ or $p = 0$.

[10] The linearity in the relationship between k, λ and h simplifies the analysis and can be generalized to any monotone increasing functions as long as $k \to 1$ when $h \to 1$: when patients have the maximum severity the decision maker cares uniquely about patients health.

patients.[11] For simplicity we assume that there can only three types of severity: low ($h = \underline{h}$), medium ($h = \frac{1}{2}$), high ($h = 1$). Notice that when $\underline{h}$ is strictly larger than 0, taking care of a low severity patient still implies a positive fatigue. If instead $\underline{h} = 0$, the decision maker does not get tired visiting low severity patients. In this extension, the decision maker working in shift t overall utility is

$$-\sum_{h\in\left\{\underline{h},\frac{1}{2},1\right\}}\left\{\frac{1}{2(1-h)}\sum_{w\geq t,i\leq I}\delta^{(w-t)I+i}\left(s^h_{wi}-n^h_{wi}\right)^2+\sum_{i=0}^{I}\delta^i i\left(\sum_{h',j<i}h' n^{h'}_{tj}\right)hn^h_{ti}\right\}-\delta^I\left(1-\frac{\sum_{h'}n^{h'}_{tI}}{\bar{n}}\right)l,$$

where s^h denotes the stock of patients with severity h, n^h is the number of visited patients with severity h. The law of motion of the stock of patients with severity h is

$$s^h_{ti}=(1-p)\left(s^h_{ti-1}-n^h_{ti-1}\right)+f^h_{ti}+\varepsilon^h_{it},\qquad s^h_{t,-1}=s^h_{t-1,I},\qquad s^h_{t,I+1}=s^h_{t+1,0}.$$

The first order condition for $n^h_{ti}, i<I$ when the DM has leisure motives, is very similar to the one analyzed in the previous section, with the difference that by increasing n^h_{ti} the decision maker decreases the distance between stocks and number of visited patients only for patients with severity h. Moreover h now affects how much the DM cares about the patient's health.

$$s^h_{ti}-n^h_{ti}+\sum_{j>i}[\delta(1-p)]^{j-i}\mathbb{E}_{\varepsilon^h_{tj}}\left[s^h_{tj}-n^h_{tj}\right]+\sum_{w>t,j\leq I}[\delta(1-p)]^{I-i+(w-t-1)I+j}\mathbb{E}_{\varepsilon^h_{wj}}\left[s^h_{wj}-n^h_{wj}\right]=0. \tag{5}$$

The first order condition for $n^h_{ti}, i=I$ is

$$s^h_{tI}-n^h_{tI}+\sum_{w>t,j\leq I}[\delta(1-p)]^{I-i+(w-t-1)I+j}\mathbb{E}_{\varepsilon^h_{wj}}\left[s^h_{wj}-n^h_{wj}\right]-(1-h)\frac{1}{\bar{n}}l=0. \tag{6}$$

The first order condition for n^h_{ti} when the DM suffers increasing fatigue is

$$\begin{aligned}s^h_{tI}-n^h_{tI}+\sum_{j>i}[\delta(1-p)]^{j-i}\mathbb{E}_{\varepsilon^h_{tj}}\left[s^h_{tj}-n^h_{tj}\right]+\sum_{w>t,j\leq I}[\delta(1-p)]^{I-i+(w-t-1)I+j}\mathbb{E}_{\varepsilon^h_{wj}}\left[s^h_{wj}-n^h_{wj}\right] \\ -(1-h)hi\sum_{h',j<i}h' n^{h'}_{tj}-(1-h)h\sum_{h',j>i}\delta^{j-i}jh'\mathbb{E}_{\varepsilon^{h'}_{tj}}\left(n^{h'}_{tj}\right)=0.\end{aligned} \tag{7}$$

With respect to equation (4), in this first order condition there is a term $h(1-h)$ which implies

[11] Moreover we implicitly assume that the probability of working overtime does not depend on the severity of patients visited in the last interval of the shift. Indeed, if there is any relation between the two, the direction would be unclear, because on one hand more severe patients probably require more time to be visited, on the other hand taking care of severe patients can involve more time than one entire shift, which needs physicians to pass their patients to physicians in the following shift, and the probability of passing patients is likely to increase with the severity of patients.

that the DM cares more about more severe patients, who are also more costly to visit, in terms of effort. Moreover, the negative marginal effect of visiting more patients with severity h depends on the composition of severity of the number of visited patients in the previous intervals. Analyzing the first order conditions of the leisure and fatigue motives, with these additional assumptions we can state the following.

Proposition 3. *If the decision maker has either leisure motives or suffers from fatigue, she visits in expected terms a number of patients equal to the in-flow of patients in the ED when they have high severity:* $\mathbb{E}\left[n_{ti}^{1*}\right]/f_{ti}^{1}=1$.

Indeed, when patients are extremely severe, the decision maker only values the health of patients without considering either leisure or fatigue. Therefore, the rate of the number of visited over inflow of high severity patients should not change from the end of a shift to the beginning of the following one, indipendently on leisure or fatigue motives. This implies that these two motives are observationally equivalent when looking at high severity patients.

Proposition 4. *For sufficiently low* δ *and sufficiently large p, if the decision maker has only leisure motives, the difference between the rate* $\frac{\mathbb{E}_{\varepsilon_{t+1,0}^{h}}\left[n^{h_{t+1,0}*}\right]}{f_{t+1,0}^{h}}$ *of the number of visited patients over inflow of patients in the first interval of the following shift and the rate* $\frac{\mathbb{E}_{\varepsilon_{tI}^{h}}\left[n^{h_{tI}*}\right]}{f_{tI}^{h}}$ *of the number of visited patients over inflow of patients in the last interval of her shift decreases with the severity of patients entering the ED in the last interval of her shift.*

When the decision maker has only leisure motives, in the last interval she would like to visit less patients to reduce the probability of working overtime. However, if the patients are more severe, she cares more about their health and she visits them more.

Proposition 5. *For sufficiently low* δ *and sufficiently large p, if the decision maker only suffers from fatigue, the difference between the rate of patients visited over inflow in the last interval of her shift and the rate of patients visited over inflow in the first interval of following shift is non-monotonic with respect to the severity of patients entering the ED in the last interval of the shift,* $\mathbb{E}\left[n_{t+1,0}^{\underline{h}*}\right]/f_{t+1,0}^{\underline{h}}-\mathbb{E}\left[n_{tI}^{\underline{h}*}\right]/f_{tI}^{\underline{h}}<\mathbb{E}\left[n_{t+1,0}^{\frac{1}{2}*}\right]/f_{t+1,0}^{\frac{1}{2}}-\mathbb{E}\left[n_{tI}^{\frac{1}{2}*}\right]/f_{tI}^{\frac{1}{2}}>\mathbb{E}\left[n_{t+1,0}^{1*}\right]/f_{t+1,0}^{1}-\mathbb{E}\left[n_{tI}^{1*}\right]/f_{tI}^{1}$.

When the decision maker suffers from increasing fatigue, when patients entering the ED have the lowest severity, she gives low value to visiting them because they are not in a life threatening situation, but visiting them implies little effort. If $\underline{h}=0$, visiting them would be costless. When instead they are extremely severe, she would like to visit them immediately, even if visiting them

is very costly in terms of effort. In both these cases, the number of visited patients is larger with respect to the scenario of patients with intermediate severity. Indeed these patients are costly to visit in terms of effort, and the value of visiting these patients is not too high.

Finally, while with a leisure motive the severity of the patients that entered the ED in the previous intervals does not affect the number of visited patients in the last interval of the shift, with a fatigue motive the more severe the patients that have entered the ED during the shift, the more tired becomes the medical team, which affects the willingness to visit patients in the last interval of the shift. Therefore the following holds.

Proposition 6. *For sufficiently low δ and sufficiently large p, if the decision maker has leisure motives, the number of medium and high severity patients that entered the ED before the last interval of the shift does not affect the number of visited patients in the last interval of the shift. If the decision maker suffers from increasing fatigue, the number of low and medium severity visited patients in the last interval of the shift decreases with the number of medium and high severity patients that entered the ED in the previous intervals of the shift: $\frac{\partial n_{tI}^{h}}{\partial f_{ti}^{h'}} < 0$, $i < I$, $h < 1, h' \geq 1/2$.*

Propositions (4), (5) and (6) deliver the following testable implications:

1. If the difference of visited over inflow of patients between the beginning of a shift and the end of the previous shift is monotonically decreasing with the severity of the patients entering the ED, the decision maker has leisure motives. If this difference is non-monotonic (inverse U-shaped) with respect to the severity of patients entering the ED, the decision maker suffers from increasing fatigue. This implies that looking at the relationship between severity and difference in rate of visited over inflow of patients makes it possible to disentangle leisure from fatigue.

2. If the number of visited patients with low or medium severity at the end of a shift depends negatively on the inflow of medium and high severity patients in the ED before the end of the shift, the decision maker suffers from increasing fatigue. Therefore, looking at the heterogeneity in this number, considering situations with a large versus low in-flow of medium and high severity patients in the ED before the end of a shift makes it possible to understand whether fatigue motives are present.

The next step of this work is to take advantage of these testable implications to disentangle leisure and fatigue motives in the incentives of care providers, using data from the Emergency Department.

4 Data

The dataset includes administrative records of all accesses to the ED in Liguria between January 2013 and December 2015. The information is collected at the patient level. Patients are identified by a pseudo-anonymized identifier that allows patients to be followed over time and across hospitals, and enable linkage between ED and inpatient records.[12] Our main variables of interest are the triage assigned to each patient, the point in time in which the patient is accepted for the medical visit, and the duration of the visit by type of triage. The timing of each step within the ED is collected by a software manager that automatically records the time access, of visit, and of discharge of the patient from the ED. The presence of this software mitigates concerns about biases and measurement errors that regards the performance indicator of interest: the number of patients visited in each point in time.

4.1 Sample construction and data descriptives

Our empirical analysis focuses on a sample of emergency patients treated in EDs that provides 24-hours services. Table 1 illustrates how we restricted the sample adopted in our analysis. We excluded from the raw data the visits that refer to walk-in clinics, that are opened only in day time. Patients treated by these units typically suffer from minor conditions, and this excludes 2.6% of emergency visits. We also exclude outlier values (above 99th percentile of the distribution) for time spent in the waiting room before being visited and the total time (wait time and duration of treatment) spent in the ED. These observations refer to patients that in total spent over 27 hours in the ED. Finally, we drop 39 observations that corresponds to individuals that were already dead when accessed the ED. The final sample includes 1,296,299 observations.

[12] The use of these data was previously authorized under a data use agreement with the health regional authority.

Table 1: Sample Definition

Sample contruction	Observations
Raw data	1,357,774
Emergency Departments opened 24/7	1,322,422
Drop waiting times >99th percentile	1,309,481
Drop treatment times >99th percentile	1,296,338
Drop if already dead when accessed the Emergency Department	1,296,299

Table 2 reports descriptive statistics of the data. The duration of a visit is expressed in minutes and is computed as the difference between the time in which the visit ended and the begin. On average, a medical visit lasts for 137 minutes, ranging between an average of 50 and of 259 minutes, depending on the triage code (h). White codes account for 10% of all accesses, green codes for 67%, yellow codes for 21%, and red codes for 2%. In each bin (i) of 1 minute size, on average 1,163 patients accessed the ED (f) and 1,128 are visited (n).

Table 2: Descriptive Statistics

Variables	Mean	Std. Dev.	25th perc	75th perc
Duration of the medical visit in the ED	137	215	22	155
Average number of accesses per minute of day	1,163	395	1013	1407
Average number of patients visited per minute of day	1,128	351	986	1408
Share of patients assigned with triage "white" (%)	10	30		
Share of patients assigned with triage "green" (%)	67	47		
Share of patients assigned with triage "yellow" (%)	21	41		
Share of patients assigned with triage "red" (%)	2	14		
Duration of visit for white codes	50	97	2	65
Duration of visit for green codes	108	169	17	129
Duration of visit for yellow codes	259	302	92	277
Duration of visit for red codes	253	316	81	264
Number of observations	1,296,251			
Number of bins with 1 minute size	1,440			
Number of bins with 1 minute size by triage	5,759			

5 Graphical analysis

The figures in this section further describe the data by illustrating how the number of patients visited (n_i) evolves during the day; the bin size is of one minute. The red vertical lines indicate the change of shift (at the end of $i = I + 1$). In the graphs, we show the number of accesses (f_i) over time (i). We also report some data elaborations: the ratio of patients visited over the number of accesses and the trend in the stock of patients waiting to be visited. At the beginning of each point in time (i), the stock (s_i) is calculated as $s_i = (s_{i-1} - n_{i-1}) + f_{i-1}$.[13] We also show how the amount of time devoted to the patients changes over time. In the first set of graphs presented, the data are aggregated. In the sets that follow, the sample is split according to the triage code (h) (white, green, yellow and red).

Reading clockwise, the panels reported in Figure 1 show how the flow of patients that entered the emergency department (f_i), the flow of patient visited (n_i), the ratio of number of patient visited over the number of accesses, and the trend in the stock (s_i) evolve over time (i).[14] From this figure, we see that the flow-in is continuous around the cut offs, while the flow-out shows a marked discontinuity. Coherently, in the bottom panels we observe that at the cutoff the ratio between outs and ins is smaller than 1, and that the stock increases. We split the sample by triage codes in order to see how these trends evolve across each group.

[13] Given that we do not observe the effective number of patients that are in the waiting room in a given moment, we pin down a specific trajectory of stock by assigning 0 to the lowest value of the stock: $\min s_i := 0$. For this reason, we look at the trend of the stock, but not at the absolute values.

[14] The choice of a one-minute bin size is to illustrate the pattern of interest trading off between precision and noise.

Figure 1: Flow (f_i) of patients that accessed the Emergency Department (top left panel), number (n_i) of patients visited by a physician (top right), number of patients visited per access (bottom left), and trend of the stock (s_i) of patients in the waiting room (bottom right).

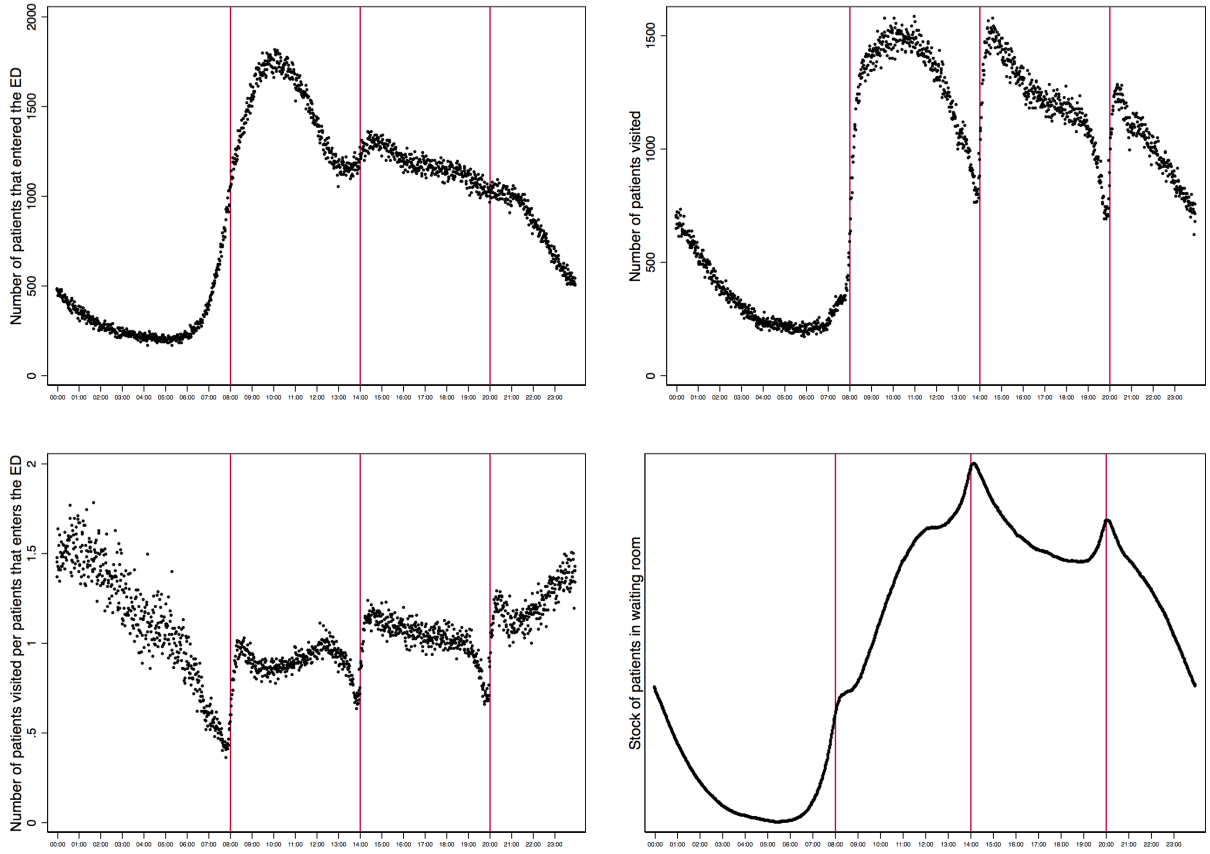
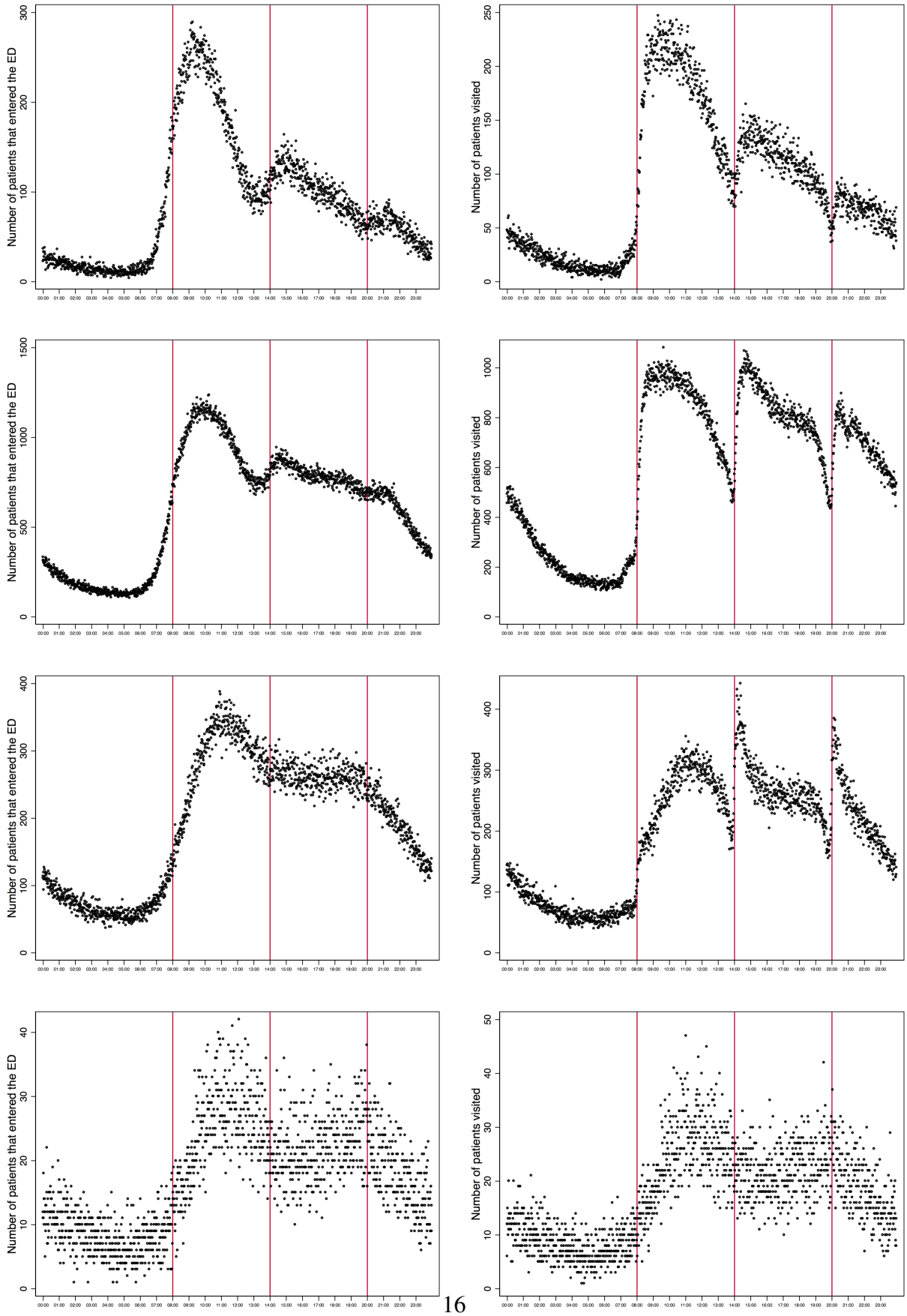


Figure 2 shows the flow (f_i) of patients that entered the emergency department (left column) and that are visited by a physician (n_i) (right column) by triage (h) assigned to the patient at the time of access (from minimum to maximum priority respectively). In these graphs there is a preliminary evidence of heterogeneity in productivity across different triages. The discontinuity at the cutoff is more pronounced for the two central degrees of priority (green and yellow) against lower (white codes, on the top) and highest priorities (red codes, on the bottom). This is more evident when the flow of patients that enters the emergency department is more stable across the change in shift, i.e. around 2pm and around 8pm.

Figure 2: Flow (f_i) of patients that accessed the Emergency Department (left column) and number (n_i) of patients visited by a physician (right column) by degree of priority (h) assigned to the patient at the time of access (from minimum - top row - to maximum - bottom row).



A clearer insight can be gained by looking at Figure 3, which shows how the providers' productivity – expressed as the ratio between flows (n_i/f_i) – changes during the day. This helps to understand something more about the possible factors that affect productivity. As predicted by the theoretical model, by looking at the aggregate data we are not able to distinguish the effect of fatigue from the slacking-off behavior. By splitting the data across triages, we see that the changes in productivity near the EOS show evidence of increasing fatigue, as predicted by the theoretical model. In particular, highest severity patients are always visited immediately, as there is no discontinuity in the number of visits performed before the cutoff, presenting a stable ratio of 1. We instead observe a sharp discontinuity for medium degrees of severity (the central panels), meaning that these patients are queued from the ending shift to the one that will start afterwards. For them, the ratio falls to about 0.5 near the end of the shift. On the right column of figure we see that also the stock trends present remarkable kinks for this types of triage, coherently with the fact that fewer of these patients are visited near EOS. Finally, lower severity patients are only marginally affected by the approaching of the end of the shift. In this case, the ratio is at least 1 during shorter shifts (between 8am-2pm and 2pm-8pm), and it decreases constantly at the end of longest shifts (8pm-8am), further supporting the idea that fatigue might be the mediating factor driving performances.

Figure 3: Number of patients visited per patient that accessed the emergency department (left column) and trend of the stock of patients in the waiting room (right column) by degree of priority assigned to the patient at the time of access (from minimum - top row - to maximum - bottom row).

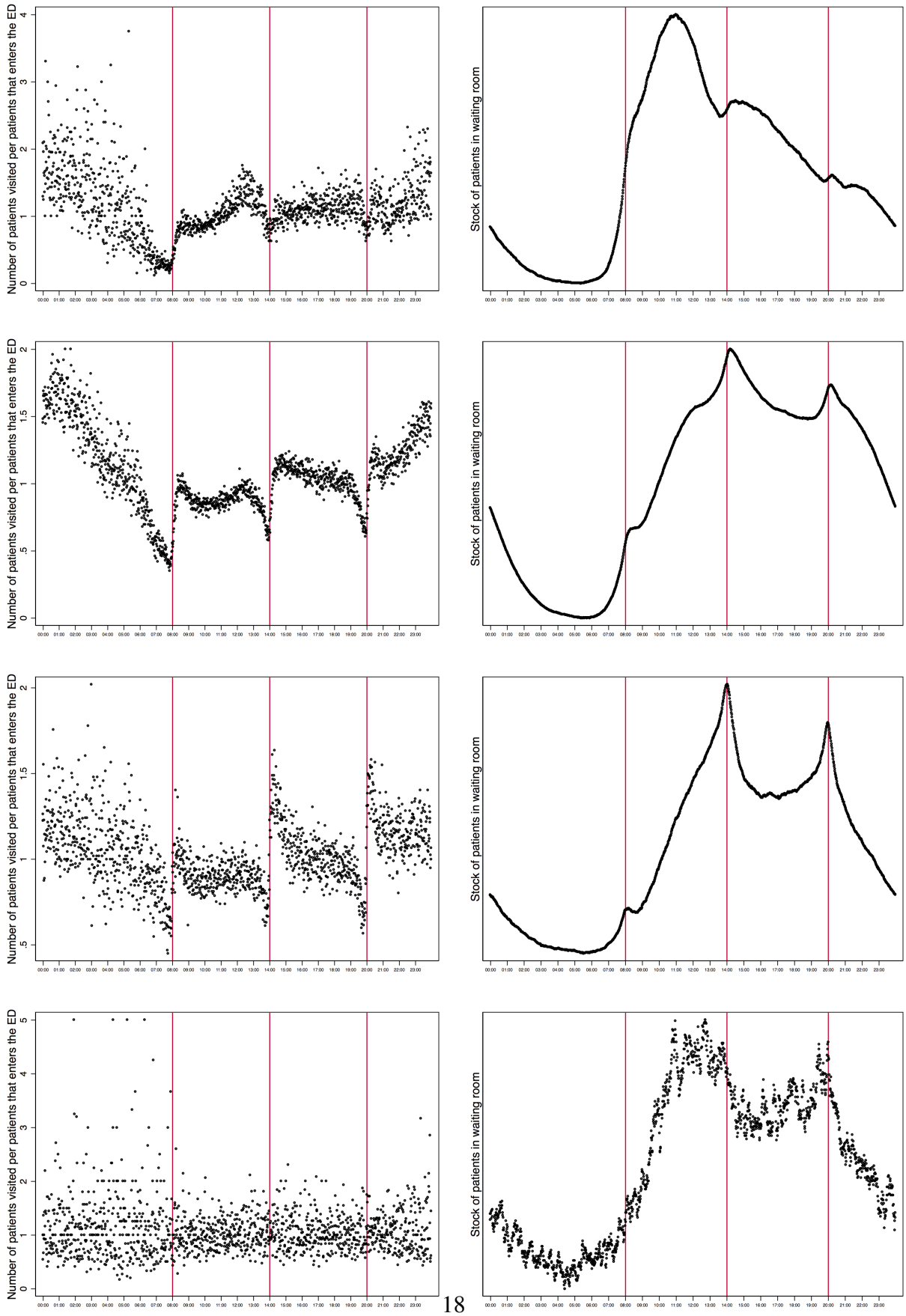
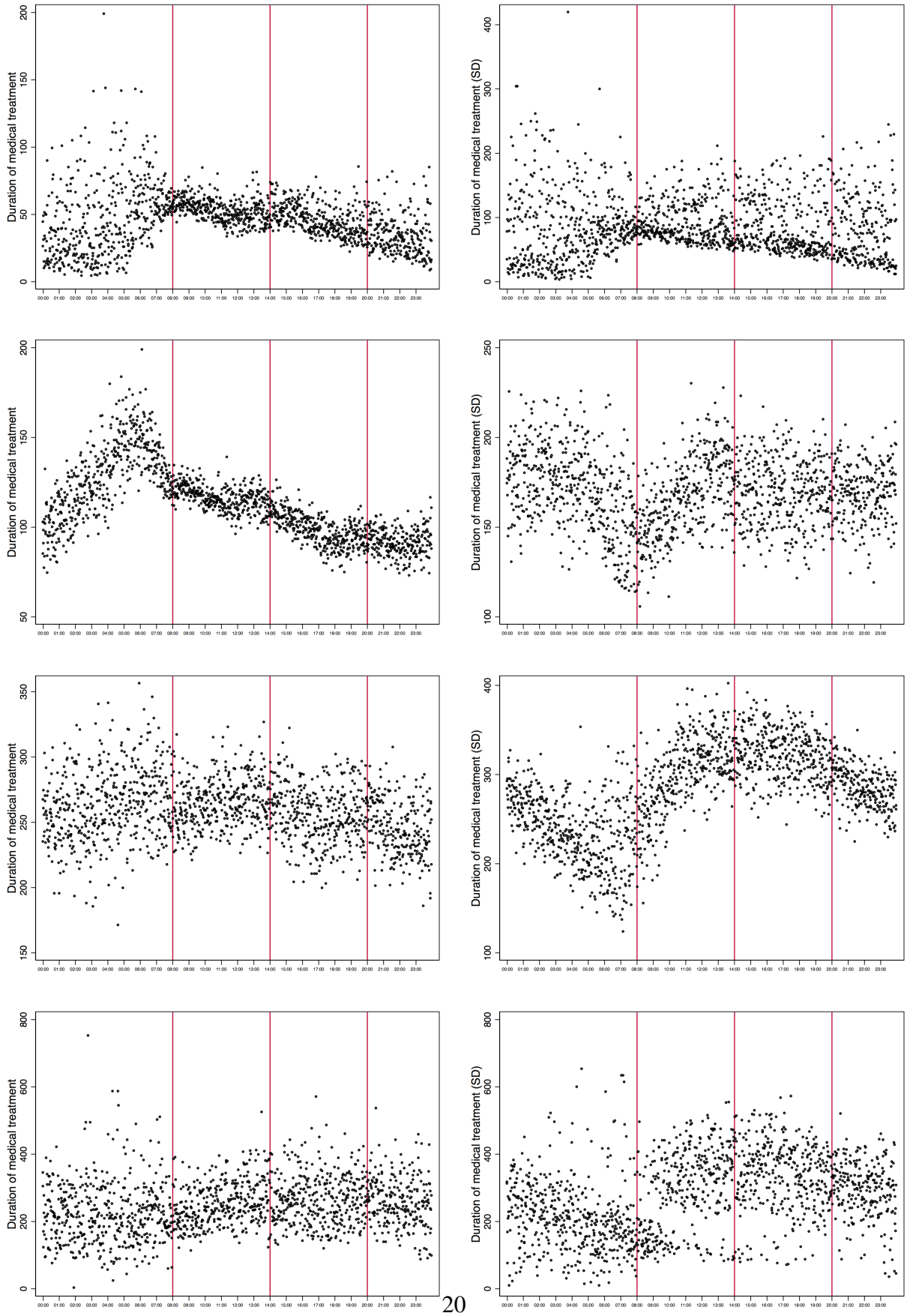


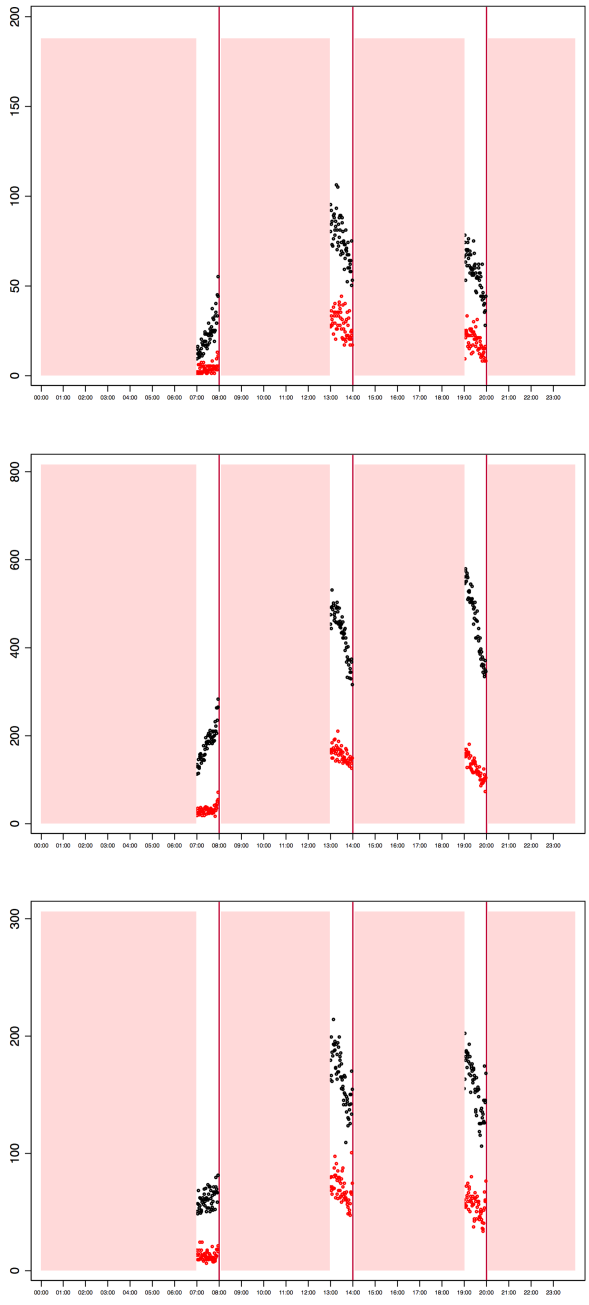
Figure 4 shows the duration of medical treatment (left panel) and its standard deviation (right panel) by degree of priority. The duration of the visit does not present discontinuities around the cutoff. The standard deviation of the duration of the visit is much larger for medium and high triages, against the lower ones. The absence of discontinuities suggests that, once the patient is accepted for the visit, the amount of time spent with the doctor is not affected by the fact that EOS is approaching.

Figure 4: Average duration of the medical treatment (left column) and standard deviation of the duration of the medical treatment (right column) by degree of priority assigned to the patient at the time of access (from minimum - top row - to maximum - bottom row).



Finally, we compute the median amount of high priority codes (red codes) visited up to one hour before the end of shift, by Emergency Department (ED) and shift (8am-2pm; 2pm-8pm; 8pm-8am). The time span up to one hour before the end of shift is the red area in Figure 5, and it represents the time during which care providers are treated with a given number of red code accesses. We are interested to see what is the effect of red code accesses on the performance of providers during the last hour of the shift, when such accesses are either above the median or below. The times in which providers faced a larger number of high priority patients is represented in figure by red dots, while the times with a lower number of high priority accesses is illustrated in black. The graphs in figure report the effect of this treatment on the number of visited patients (n_i) of each of the other priority codes, from white (lowest) to yellow. We observe that the times in which more red codes accessed the ED, fewer patients with lower degree of priority are visited one hour before the end of shift. This is coherent with the prediction of the theoretical model, where such result is attributed to the greater tiredness accumulated by health providers by visiting high severity patients, a type of subject that requires a greater level of effort if compared with other patients.

Figure 5: Number (n_i) of patients visited one hour before the end of shift by days in which the number of high priority codes exceed the median at the shift and emergency department level, against the days in which the number of high priority codes is at the median or below. By degree of priority (h): white, green and yellow, respectively.



To conclude, in general it appears that near EOS there is a discontinuity in the number of patients visited. When EOS in approaching, the productivity of physicians decreases and the most affected patients are green and yellow codes, i.e. the ones for which the risk of death is not imminent, but it requires an higher effort visiting them if compared with white codes. Consistently

with our theoretical framework, we do not observe a discontinuity at the cutoff for patients in life threatening conditions. In addition, we see that the number of patients of any degree of priority visited one hour before EOS is lower when providers visited an higher number of high priority patients during the shift. According with the predictions of the theoretical model, this suggests that the decrease in productivity near EOS is given – at least in part – to fatigue.

6 Final remarks and conclusion

Existing literature discusses the role of fatigue and the preference of workers for enjoying leisure time over working overtime as determinants for worker's performance near end of shift (EOS). The emergency medical setting is an appropriate setting to perform this type of analysis, as workers are intrinsically motivated, contracts are incomplete, and work is scheduled.

In this paper, we propose a theoretical model with testable implications to distinguish between the effect of fatigue and leisure on workers' performance near EOS. We report empirical evidence showing that the number of patients visited near EOS is not remarkably different to the number of patients visited in other points in time during the shift for two types of patients: the ones in life threatening conditions and the lowest urgent ones. The greater drop in the performance level is given by the number of medium degree patients (namely, green and yellow codes) visited when EOS is approaching. We exploit this non-monotonic change in performance to show that the behavior of care providers is determined also by fatigue. Finally, we show that the drop in the performance level is greater when providers visited a larger number of high priority patients during the shift.

Is important to distinguish between fatigue and leisure motives, as the resulting policy implications might differ substantially. The next step of this work is to deepen the empirical analysis.

References

Brachet, T., David, G., and Drechsler, A. M. (2012). The effect of shift structure on performance. *American Economic Journal: Applied Economics*, 4(2):219–246.

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

Hart, O. and Holmstrom, B. (1987). *The Theory of Contracts*. Cambridge University Press.

Tirole, J. (1986). Hierarchies and bureaucracies: On the role of collusion in organizations. *JL Econ. & Org.*, 2:181.

A Appendix

Proof of Proposition (1)
Substituting $n_{ti} = f_{ti} + \varepsilon_{ti}$ solves equation (1), for every t and i. ■
Proof of Proposition (2)
We prove the Proposition for $\delta = 0$ and $p = 1$. By the linearity of the first order conditions the Proposition will then hold for sufficiently low δ and sufficiently large p. Substituting $n_{ti} = s_{ti}$ in equation (2) solves the equation. Substituting $n_{tI} = s_{tI} - (1-k)\frac{1}{\bar{n}}l$ solves equation (3), which implies that in the leisure motive the decision maker visits the following number of patients: $n_{ti} = f_{ti} + \varepsilon_{ti}$, for $1 < i < I$, $n_{tI} = f_{tI} + \varepsilon_{tI} - (1-k)\frac{1}{\bar{n}}l$, and $n_{t0} = f_{t0} + \varepsilon_{t0}$, for every t, which implies the Proposition. Substituting $n_{t0} = s_{t0}$ in equation (4) for $i = 0$ solves the equation. Substituting $n_{ti} = s_{ti} - (1-k)\lambda i \sum_{j=0}^{i-1} n_{tj}$ solves equation (4) for $i > 0$. Therefore the following holds: $\mathbb{E}(n_{t1})/f_{t1} = 1$, $n_{t2} = f_{t2} + \varepsilon_{t2} - (1-k)\lambda n_{t1}$, therefore $\mathbb{E}(n_{t2})/f_{t2} < \mathbb{E}(n_{t1})/f_{t1} = 1$, $n_{t3} = f_{t3} + \varepsilon_{t3} - (1-k)2\lambda(n_{t1} - f_{t2} - \varepsilon_{t2}) + (1-k)^2 2\lambda^2 n_{t1}$, therefore $\mathbb{E}(n_{t3})/f_{t3} < 1$. These inequalities are satisfied for every $i \leq I$: $\mathbb{E}(n_{ti})/f_{ti} < 1$. Therefore the Proposition follows. ■
Proof of Proposition (3)
If $h = 1$, the solution to the first order conditions (5) and (6) is $n^1_{ti} = f^1_{ti} + \varepsilon^1_{ti}$, which proves the Proposition for the leisure motive. If $h = 1$, the solution to the first order condition (7) is $n^1_{ti} = f^1_{ti} + \varepsilon^1_{ti}$, which proves the Proposition for the fatigue motive. ■
Proof of Proposition (4)
We prove the Proposition for $\delta = 0$ and $p = 1$. By the linearity of the first order conditions the Proposition will then hold for sufficiently low δ and sufficiently large p. Substituting $n^h_{ti} = f^h_{ti} + \varepsilon^h_{ti}$ in equation (5) solves the equation. Substituting $n^h_{tI} = f^h_{tI} + \varepsilon^h_{tI} - (1-k)\frac{1}{\bar{n}}l$ solves equation (6). While $\mathbb{E}(n^h_{ti})/f^h_{it}$ is equal to 1, $\mathbb{E}\left[n^h_{tI}\right]/f^h_{tI}$ increases with h. ■
Proof of Proposition (5)
We prove the Proposition for $\delta = 0$ and $p = 1$. By the linearity of the first order conditions the Proposition will then hold for sufficiently low δ and sufficiently large p. In the first interval of shift $t+1$, the decision maker implements $n^h_{t+1,0} = f^h_{t+1,0} + \varepsilon^h_{t+1,0}$, which implies that the rate $\mathbb{E}\left[n^{h*}_{t+1,0}\right]/f^h_{t+1,0} = 1$. In the last interval of shift t, the decision maker implements $n^h_{tI} = f^h_{tI} + \varepsilon^h_{tI} - (1-h)hI\sum_{h',i<I} h' n^{h'}_{ti}$, which implies $\mathbb{E}[n^h_{tI}] = f^h_{tI} - (1-h)hI\sum_{h',i<I} h' n^{h'}_{ti}$ The derivative of the latter expression with respect to h is $(-1+2h)I\sum_{h',i<I} h' n^{h'}_{ti}$, which is negative for $h < \frac{1}{2}$ and positive otherwise. ■
Proof of Proposition (6)
We prove the Proposition for $\delta = 0$ and $p = 1$. By the linearity of the first order conditions the Proposition will then hold for sufficiently low δ and sufficiently large p. When the DM cares about leisure, the solutions $n^h_{tI} = f^h_{tI} + \varepsilon^h_{tI} - (1-k)\frac{1}{\bar{n}}l$ and $n^h_{tI} = f^h_{tI} + \varepsilon^h_{tI}$ do not depend on the inflow in previous intervals. When the DM suffers from increasing fatigue, the solution $n^h_{tI} = f^h_{tI} + \varepsilon^h_{tI} - (1-h)hI\sum_{h',i<I} h' n^{h'}_{ti}$ depends negatively on $n^{h'}_{ti}$ if $h' > 0$ and $h < 1$, which itself depends positively on the inflow of patients $f^{h'}_{ti}$: $n^h_{ti} = f^h_{tI} + \varepsilon_{ti} - (1-h)hi\sum_{h',j<i} h' n^{h'}_{tj}$. ■