INSEAD The Business School for the World®

Working Paper

2023/52/TOM

(Revised version of 2022/47/TOM)

Can Predictive Technology Help Improve Acute Care Operations? Investigating the Impact of Virtual Triage Adoption

Jiatao Ding
INSEAD, jiatao.ding@insead.edu

Michael Freeman
INSEAD, michael.freeman@insead.edu

Sameer Hasija
INSEAD, sameer.hasija@insead.edu

This paper investigates the operational implications and policy impacts of virtual triage adoption within the acute care service setting. A central problem in this context is patients' (in)ability to self-triage accurately, a notable contributor to emergency department overcrowding and treatment delays. While traditional triage solutions, such as phone services, can mitigate these issues to an extent, they struggle with accessibility and accuracy problems. However, recent advances in predictive technology have led to the development and deployment of virtual triage tools, which offer immediate, cost-effective, and potentially more accurate triage recommendations. Despite their potential benefits and increasing adoption, the impact of virtual triage tools on acute care systems remains poorly understood. This paper therefore develops a queueing game model to examine how virtual triage influences patient behavior and system performance, and explores policy actions that maximize the operational advantages of these tools. The analysis uncovers an inherent trade-off between informativeness and volume with respect to patient's compliance with virtual triage recommendations. For system performance, we demonstrate potential drawbacks of off-the-shelf virtual triage solutions, and explore the ways in which these technologies can be customized to specific contexts in order to unlock their full potential. Our findings underline the pressing need for effective regulation and thorough assessment of operational consequences to harness the full potential of virtual triage in improving the delivery of acute care.

Keywords: Healthcare; Acute Care; Predictive Technology; Virtual Triage; Off-the-Shelf vs Custom; Queueing Game
History: Last updated on September 29, 2023.

Can Predictive Technology Help Improve Acute Care Operations? Investigating the Impact of Virtual Triage Adoption

Jiatao Ding, Michael Freeman, Sameer Hasija
INSEAD, 1 Ayer Rajah Avenue, Singapore 138676
jiatao.ding@insead.edu, michael.freeman@insead.edu, sameer.hasija@insead.edu

This paper investigates the operational implications and policy impacts of virtual triage adoption within the acute care service setting. A central problem in this context is patients' (in)ability to self-triage accurately, a notable contributor to emergency department overcrowding and treatment delays. While traditional triage solutions, such as phone services, can mitigate these issues to an extent, they struggle with accessibility and accuracy problems. However, recent advances in predictive technology have led to the development and deployment of virtual triage tools, which offer immediate, cost-effective, and potentially more accurate triage recommendations. Despite their potential benefits and increasing adoption, the impact of virtual triage tools on acute care systems remains poorly understood. This paper therefore develops a queueing game model to examine how virtual triage influences patient behavior and system performance, and explores policy actions that maximize the operational advantages of these tools. The analysis uncovers an inherent trade-off between informativeness and volume with respect to patient's compliance with virtual triage recommendations. For system performance, we demonstrate potential drawbacks of off-the-shelf virtual triage solutions, and explore the ways in which these technologies can be customized to specific contexts in order to unlock their full potential. Our findings underline the pressing need for effective regulation and thorough assessment of operational consequences to harness the full potential of virtual triage in improving the delivery of acute care.

Key words: healthcare, acute care, predictive technology, virtual triage, off-the-shelf vs custom, queueing game

History: Last updated on September 29, 2023.

1. Introduction

The rapidly growing and aging population is placing an unprecedented burden on acute care services worldwide. These services face increasing demand from patients presenting with urgent medical emergencies, acute complications from chronic conditions, and common illnesses necessitating immediate attention (Hirshon et al. 2013). Propelled by this booming demand, the global acute care market is forecasted to expand from USD 2.4 trillion in 2018 to USD 4.0 trillion by 2026, reflecting a compound annual growth rate of 6.7% (Grand View Research 2019). Rising revenue streams, coupled with the growing volume and diversity of demand, have fostered the expansion and comprehensiveness of acute care services. As a result, patients now have access to a wide array of options to meet their acute care needs, including same-day primary care appointments,

urgent care centers, hospital-based emergency departments (EDs), and freestanding EDs (Kocher and Ayanian 2016).

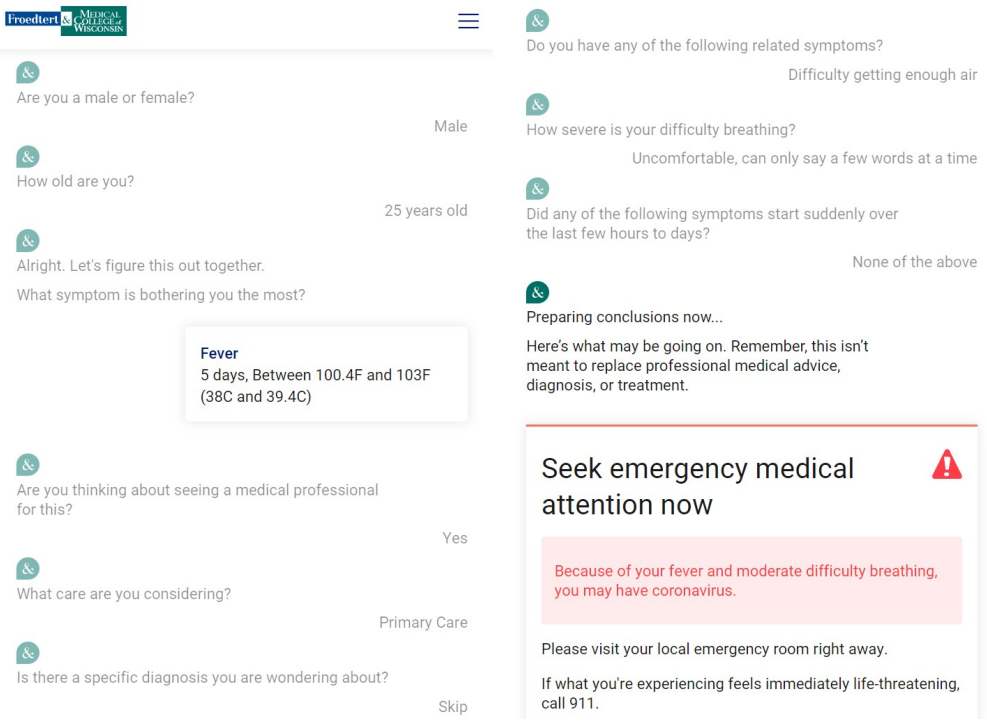
This complex assortment of acute care options presents both opportunities and challenges for patients attempting to determine the appropriate level and location of acute care. Due to a general lack of professional medical knowledge, patients often find it difficult to make clear-cut decisions about which healthcare setting is most suitable for their acute care needs. This decision-making process, termed self-triage, critically depends on the patients' ability to assess the urgency and complexity of their health conditions. This task is made even more challenging by the fact that patients must make these decisions while potentially experiencing discomfort and heightened emotional distress. As a result, patients may end up making inaccurate choices about their acute care needs (Trivedi et al. 2017).

Inaccurate self-triage can lead to a mismatch between the supply and demand of acute care resources. For instance, when patients who require primary care opt for care at an ED, they not only incur unnecessary costs but also exacerbate the *overcrowding* problem at the ED.[1] Conversely, patients who require emergency care but incorrectly choose to seek care in a primary care setting may end up being referred to secondary care, such as an ED or an outpatient specialist, if the general practitioner (GP) is unable to diagnose or treat the patient. This scenario not only leads to unnecessary costs at the primary care level but also results in *treatment delays*.

Given these potential inefficiencies, improving the accuracy of patients' initial self-triage decisions could significantly enhance the efficient allocation of acute care resources. Traditional methods for assisting patients with self-triage, such as phone triage services, have been in use since the 1970s (Coons and DuMoulin 2000). In a phone triage service, patients call in to describe their symptoms to a triage nurse who then provides a recommendation. However, such services have long been plagued by accessibility issues due to the inherent trade-off between responsiveness and cost. High service demand and limited service capacity often result in long waiting times, discouraging patients from using these services. This issue was especially pronounced during the recent COVID-19 pandemic when, e.g., patients in the UK with COVID-19 symptoms struggled to get through to the National Health Service (NHS) 111 phone triage service, with many reporting being put on hold for up to three hours or being disconnected entirely (Dalton 2020).

The advent of predictive analytics, including machine learning and artificial intelligence (AI), offers an alternative way to augment patients' self-triage accuracy instantly and at limited additional cost. Healthtech and technology companies worldwide have been harnessing these advancements to develop and implement "virtual triage" tools in the form of websites and mobile applications. These tools, using a series of questions about a patient's personal information and symptoms, can offer immediate triage recommendations. Because these recommendations are made by

Figure 1 Sample triage recommendation from Buoy Health's AI-powered virtual triage tool on Froedtert's website. Source: https://froedtert.buoyhealth.com/symptom-checker/



pre-trained classification algorithms, virtual triage provides a significant cost advantage over traditional phone triage services: it is highly scalable, with low marginal operating costs, and provides instantaneous triage recommendations without long delays. Additionally, virtual triage can continuously learn and improve its accuracy over time, with more training data and better classification algorithms.

Recognizing these benefits, major health providers have partnered with virtual triage firms to increase the technology's adoption. For example, in 2017, Babylon Health in the UK collaborated with the NHS to offer a virtual triage service. This service typically requires 12 text messages and about one and a half minutes to complete (Lovett 2018). Similarly, in the US, Buoy Health provided an AI-powered virtual triage tool to Froedtert & the Medical College of Wisconsin via its website (see Figure 1 for a sample virtual triage recommendation). Early empirical evidence has shown that these tools are effective in modifying users' care-seeking behavior. For example, a recent study of Buoy Health's virtual triage chatbot found that 32% of its users reduced their intended level of care (Winn et al. 2019).

However, despite the widespread development and deployment of virtual triage tools, little is known about their overall impact on the acute care systems. Previous attempts to deploy medical predictive technologies have found that they can underperform in real-world settings due to a

lack of understanding of specific clinical constraints and operational challenges, even when these technologies demonstrate high accuracy in laboratory settings (Heaven 2020).

With these considerations in mind, this paper develops a queueing game model to investigate the operational impact and policy implications of virtual triage adoption. Our analysis takes into account several relevant and interrelated issues. First, we explore how virtual triage influences patients' care-seeking behavior: Given the decentralized nature of the technology, patients may not always follow the recommendations of virtual triage, particularly when these contradict their initial self-triage decisions.

Second, we examine the impact of virtual triage on the system performance of acute care services. In particular, we analyze two important metrics - social cost and triage safety. Since there is a lack of evidence on how acute care system parameters guide the decision on virtual triage accuracy, we begin by examining the implication of adopting an "off-the-shelf" virtual triage tool with predetermined accuracy. In this setting, we would like to understand whether the introduction of an informative virtual triage solution invariably leads to better outcomes? Moreover, given its unique ability to learn and improve its accuracy over time, we are also interested in whether more accurate virtual triage is always beneficial? Related, we consider whether regulators should only authorize the current version of a virtual triage tool after evaluating its effectiveness, or if they can also authorize subsequent more accurate versions without re-evaluation (Babic et al. 2019)?

From a policy perspective, this paper also analyzes the regulatory actions that optimize the operational benefits of a "custom" virtual triage. Specifically, given that the accuracy of a virtual triage tool can be adjusted along the receiver operating characteristic (ROC) curve, how should the accuracy be determined to optimize system performance? And how should the accuracy change as patient composition changes or as the algorithm's triage capability improves? Moreover, how does the optimal accuracy differ across different objective functions (i.e., minimizing social cost vs maximizing triage safety)? This paper explores and seeks to answer these questions.

Overall, this paper highlights the potential of virtual triage tools to enhance the performance of acute care systems, provided that their implementation is properly managed and their operational implications are fully understood.

2. Related Literature

Our work contributes to multiple healthcare and operations management literature streams given its focus on acute care, triage, information, and technology adoption.

ED Overcrowding and Triage. The ED overcrowding problem has attracted considerable interest within the operations and healthcare management literature. Several papers have explored potential triage mechanisms to improve ED operational efficiency and responsiveness (Saghafian

et al. 2012, Huang et al. 2012, Zayas-Cabán et al. 2014, Kamali et al. 2019). Despite the use of ED triage to prioritize treatment of patients requiring emergency care, patients who require only primary care still arrive in the ED, wasting costly resources. ED resources are further stretched as nurses and physicians must provide triage for these patients, while diverting resources away from direct patient care in this way may act to reduce the overall quality of care provided (Corl 2019). In contrast with this traditional ED triage, virtual triage seeks to prevent patients who only require primary care from making unnecessary ED visits in the first place, thus preserving expensive resources for patients with the highest need and reducing the costs associated with providing care that exceeds patients' needs. Our research therefore contributes to the body of work on triage processes by studying the impact of an upstream decentralized virtual triage service on the ED overcrowding problem.

Two-Tier Services. An important feature of acute care systems is that patients can typically choose whether to be treated at a GP (tier 1) or an ED (tier 2). One stream of literature relating to two-tier service considers a system where a tier 1 server (e.g., a generalist) acts as a gatekeeper for a downstream tier 2 server (e.g., a specialist). In these settings, all customers must first be assessed by the tier 1 gatekeeper, who decides whether to serve the customer themselves or, if the customer's service request is too complex, to refer the customer to a downstream specialist (Shumsky and Pinker 2003, Hasija et al. 2005, Lee et al. 2012, Freeman et al. 2017, Freeman et al. 2020). This existing literature focuses on the strategic behavior of service providers, while assuming customers are nonstrategic in the sense that they all initially arrive at a tier 1 server. However, in settings like acute care, depending on patients' self-assessment of their healthcare requirements, they can choose to visit a GP first or an ED directly, at their discretion. One recent paper that studies customers' strategic behavior in such a two-tier service setting is Sharma et al. (2019). By modeling patients' choice problem as a network queueing game, they analytically characterize the equilibrium outcomes and design novel incentive mechanisms to align equilibrium patient flow to the social optimum. Our paper contributes to this stream of literature by introducing an additional informative virtual triage signal and assessing its impact on patient behavior and system performance. Moreover, we explore a unique information control policy to optimize system performance in the virtual triage context.

Information in Decentralized Systems. Our paper is closely related to the stream of literature on information in decentralized systems. In the queueing literature, existing work has analyzed how customers' queue-joining behavior depends on their private information about service quality (Veeraraghavan and Debo 2009, Debo et al. 2012), service rate (Cui and Veeraraghavan 2016), and real-time delay (Hu et al. 2018). Studies within the social learning literature have analyzed firms' pricing (Papanastasiou et al. 2015) and information provision (Papanastasiou et al. 2018) decisions

in a context where customers observe available reviews and update their product quality beliefs. These existing studies focus on *system information* such as product quality, service rate, or queue length, which is not affected by customers' unique characteristics. Insights from this literature generally underscore the value of information obfuscation: Due to agents' self-interested behavior and the impact of (negative) information externality, full information could lead to suboptimal outcomes, and therefore the optimum is achieved with less information or less accurate information. However, in acute care services, due to the necessity for patients to self-triage before seeking care and to their lack of medical knowledge, customer-specific *personal information* gives rise to major uncertainty. Our findings therefore differ from those of previous studies and are twofold. On one hand, contrary to the existing literature, we find that in our model, full information is strictly preferred because it perfectly reveals a patient's type. On the other hand, an additional informative signal can either improve or degrade system performance, while a more accurate signal can lead to either better or worse outcomes.

Learning of Personal Information. In diagnostic services, customers are heterogeneous and belong to one of a given set of types. To determine a customer's type, the service provider performs a sequence of imperfect tests. Multiple studies have considered such scenarios (Alizamir et al. 2013, Sun et al. 2018, Levi et al. 2019). In this stream of literature, learning of personal information is centralized, i.e., a single service provider conducts diagnostic tests for arriving customers, and therefore providers face the same information-delay trade-off: While running additional tests could improve diagnosis accuracy, it also delays service for other customers in the system. In such cases, it is clear that more information could be detrimental. However, in our setting, learning of personal information is decentralized, i.e., patients use virtual triage to receive an informative signal about the type of their healthcare needs before seeking care. More importantly, as virtual triage is provided by algorithms, additional information is obtained instantaneously. Hence, in our model, the information-delay trade-off no longer exists. Yet, as we show in later sections, when learning of personal information is decentralized and costless, more information can in fact still degrade system performance. Our paper also differs from existing work in terms of the information control policy. In the existing literature, because of the information-delay trade-off, the objective of service providers is to determine the optimal number of tests to perform. Meanwhile, diagnostic accuracy is assumed to be exogenous. By contrast, in our paper, the diagnostic accuracy can be endogenized subject to a given ROC curve.

Telemedicine. Our focus on virtual triage is also related to the literature on telemedicine (Rajan et al. 2019, Bavafa et al. 2018, Bavafa et al. 2019). Recent work by Liu et al. (2018) and Savin et al. (2019) analyzes the delivery of telemedicine through on-demand healthcare service platforms. Most

similar to our work in this stream of literature is Çakıcı and Mills (2020), who analyze the impact of traditional teletriage provided by nurse-staffed phone lines on healthcare demand management.

Given its focus on virtual triage powered by predictive technology, our study differs from the aforementioned studies in multiple dimensions. First, virtual triage differs from telemedicine in that the technology does not remotely deliver care to patients. Instead, its main purpose is to assist patients in triaging to the appropriate level of care. Second, virtual triage has a significant cost advantage over traditional nurse line triage. With virtual triage, recommendations are provided by algorithms and no medical professional is required. Third, for patients, virtual triage is more convenient as they can get instantaneous triage recommendations with no delay, while with traditional nurse lines, long waiting times are possible given high service demand and limited service capacity. Fourth, the accuracy of virtual triage can be endogenized along the ROC curve and improved over time with more training data and better classification algorithms, while the accuracy of nurse line triage is typically fixed given the training and clinical guidelines the nurses receive. We explore these unique characteristics of virtual triage in this paper.

Virtual Triage. To the best of our knowledge, the only paper in the literature that also studies virtual triage analytically is Singh et al. (2020). The authors propose an integral approach where the classifier of virtual triage and the queueing system at an ED are jointly optimized to minimize expected waiting cost in the ED. In related studies, Dai and Singh (2020, 2021) analyze physicians' decisions on the adoption of AI in clinical practice. Our paper instead studies virtual triage as a decision support tool for patients who must choose to seek care from a GP or an ED.

Outside of the operations management literature, there is a growing interest from the medical community in empirically evaluating the accuracy and effect of virtual triage on users' care-seeking behavior (Verzantvoort et al. 2018, Meyer et al. 2020, Semigran et al. 2015, Chambers et al. 2019). They find that triage advice from these tools generally encourages users to seek emergency care. However, patient compliance with virtual triage in this case is limited: While there is generally good agreement between virtual triage recommendation and patients' intended actions, those who the system advises to go to an ED are more likely to seek care from a GP first. This in fact leads to delayed emergency care seeking and a decrease in ED visits. Our paper provides an analytical explanation that reconciles and rationalizes these two seemingly conflicting empirical observations, and this explanation is driven by the inherent informativeness-volume trade-off with respect to patient's compliance with virtual triage recommendations.

3. Model Formulation and Preliminaries

We consider an acute care system consisting of GPs and an ED serving a patient base. Within the set of patients seeking acute care, some are non-strategic. In particular, patients in highly acute

situations or requiring immediate life-saving interventions (e.g., cardiac arrest, major trauma) will visit the ED with certainty, often arriving by ambulance, and will receive prioritized care in the ED. Meanwhile, those patients experiencing a more chronic illness (e.g., hypertension, diabetes) will, depending on the complexity of their condition, almost certainly visit either a GP or a specialist. These patients typically do not require a same-day acute care appointment and can instead wait for an available appointment on some future date. For these two patient types, we can thus assume that the choices of care location remain unaffected by the introduction of virtual triage technology. Meanwhile, given GPs' limited operating hours, the ED could be the only resource available for patients seeking acute care at night or on weekends. These patients do not have an option, and their choices of care locations are unchanged by the adoption of virtual triage as well.

On the other hand, many patients experiencing moderately acute symptoms (e.g., stomach pain, shortness of breath) and seeking care during the day are potentially strategic: they have uncertainty regarding the complexity of their illness and are therefore unsure whether they should schedule a same-day GP appointment or go directly to the ED.[2] In this case, an additional signal from the virtual triage tool about the appropriate location of care for their condition can help reduce their level of self-triage uncertainty and potentially change their care-seeking behavior. The focus of our analysis in this paper is thus on this subset of strategic patients.[3] (Henceforth, we use the terms "strategic patients" and "patients" interchangeably.)

3.1. Patient Arrivals and Self-Triage

We assume that strategic patients are either GP-type, denoted by L, or ED-type, denoted by H. We denote the arrival rate of L patients by λ_L and the arrival rate of H patients by λ_H, with $\lambda = \lambda_L + \lambda_H$. While L patients can get treated at either a GP (at a lower cost) or the ED (at a higher cost), H patients require emergency care resources and can only be treated effectively at the ED. Hence, when H patients visit a GP first, they must subsequently be referred to the ED.

In deciding whether to visit a GP first or the ED directly, strategic patients need to self-triage and determine their type based on their symptoms and medical knowledge. We denote self-triaged GP-type patients by $\hat{L}$ and self-triaged ED-type patients by $\hat{H}$, with an associated self under-triage probability $\hat{\alpha} = Prob(\hat{L}|H)$ and self over-triage probability $\hat{\beta} = Prob(\hat{H}|L)$.[4] Following the extant literature on diagnostics and triage (Alizamir et al. 2013, Çakıcı and Mills 2020), we assume that strategic patients update their belief of being H upon self-triage according to Bayes' rule. Hence, $\hat{L}$ patients' belief of being H is $b_{\hat{L}} = \frac{\hat{\alpha}\lambda_H}{\hat{\alpha}\lambda_H + (1-\hat{\beta})\lambda_L}$, with an arrival rate $\lambda_{\hat{L}} = \hat{\alpha}\lambda_H + (1-\hat{\beta})\lambda_L$; $\hat{H}$ patients' belief of being H is $b_{\hat{H}} = \frac{(1-\hat{\alpha})\lambda_H}{(1-\hat{\alpha})\lambda_H + \hat{\beta}\lambda_L}$, with an arrival rate $\lambda_{\hat{H}} = (1-\hat{\alpha})\lambda_H + \hat{\beta}\lambda_L$. Without loss of generality, we assume that $\hat{\alpha} + \hat{\beta} \leq 1$, and therefore we have $b_{\hat{L}} \leq b_{\hat{H}}$, i.e., $\hat{H}$ patients are more likely to be H than $\hat{L}$ patients.

3.2. Virtual Triage

In addition to self-triage, strategic patients in our setting receive an additional signal from the virtual triage tool. The virtual triage classification algorithm can label the user to be either a GP-type patient, denoted by $\tilde{L}$, or an ED-type patient, denoted by $\tilde{H}$, and recommend that they visit a GP or an ED accordingly. Specifically, the output of the underlying classification algorithm is a probability s, i.e., the predicted probability of the user being H, given the information about the user. We assume that s is unbiased.[5] A user with a probability s below (above) a chosen threshold will then be virtual triaged as $\tilde{L}$ ($\tilde{H}$) and advised to visit a GP (an ED) (Baker et al. 2020).

To characterize the efficacy of the classification algorithm of a given virtual triage tool, let $g(s)$ denote the probability density distribution of the predicted probabilities of all the virtual triage users, with $\int_0^1 sg(s)\,ds = \frac{\lambda_H}{\lambda}$, i.e., the fraction of H patients in the patient base. We assume $g(s)$ is continuous in $s \in [0,1]$. It is then the virtual triage provider's decision to choose a discrimination threshold probability $\bar{s} \in [0,1]$, such that when $s > \bar{s}$, the patient is virtual triaged as $\tilde{H}$ and advised to visit an ED, and when $s \leq \bar{s}$, the patient is virtual triaged as $\tilde{L}$ and advised to visit a GP. This threshold $\bar{s}$ is typically chosen with the objective of maximizing or minimizing some scoring function (Gneiting 2011). For a given $g(s)$, any $\bar{s}$ has an associated virtual under-triage probability $\tilde{\alpha}(\bar{s}) = Prob(\tilde{L}|H) = \frac{\int_0^{\bar{s}} sg(s)ds}{\int_0^1 sg(s)ds}$ and virtual over-triage probability $\tilde{\beta}(\bar{s}) = Prob(\tilde{H}|L) = \frac{\int_{\bar{s}}^1 (1-s)g(s)ds}{\int_0^1 (1-s)g(s)ds}$. As $\bar{s}$ varies, $\tilde{\alpha}(\bar{s})$ and $\tilde{\beta}(\bar{s})$ vary accordingly.[6] We characterize the implicit dependence of $\tilde{\alpha}$ on $\tilde{\beta}$ by the following lemma.

LEMMA 1. *The virtual under-triage probability $\tilde{\alpha}$ is a decreasing and convex function in the virtual over-triage probability $\tilde{\beta}$, denoted by $\tilde{\alpha} = r(\tilde{\beta})$, with $r(0) = 1, r(1) = 0$. Hence, $\tilde{\alpha} + \tilde{\beta} \leq 1$.*

Lemma 1 highlights the underlying accuracy trade-off faced by the virtual triage provider.[7] As $\bar{s}$ increases, H patients are more likely to be virtual under-triaged as $\tilde{L}$, while L patients are less likely to be virtual over-triaged as $\tilde{H}$, resulting in larger $\tilde{\alpha}$ and smaller $\tilde{\beta}$. For binary classification models, this accuracy trade-off is commonly captured by a receiver operating characteristic (ROC) curve, which is defined by plotting sensitivity, i.e., $1-\tilde{\alpha}$, against $1-$specificity, i.e., $\tilde{\beta}$, at various thresholds $\bar{s} \in [0,1]$. To facilitate the exposition, we introduce the function $\tilde{\alpha} = r(\tilde{\beta})$ in Lemma 1 and refer to it as the inverted receiver operating characteristic (IROC) curve.

One characteristic of virtual triage, particularly those powered by learning algorithms, is its capability to improve accuracy over time with more training data and better classification algorithms. Specifically, $g(s)$ will be distributed more towards $s = 0$ and $s = 1$ over time, achieving a higher triage capability. We formalize this learning effect of virtual triage with the following lemma.

LEMMA 2. *Let $g_1(s)$ and $g_2(s)$ denote two probability density distributions of the predicted probabilities of being H for all the users of the virtual triage tool. Suppose $\forall\ \bar{s}_1, \bar{s}_2 \in [0,1]$ s.t. $\int_{\bar{s}_1}^1 (1-$*

$s)g_1(s)ds = \int_{\bar{s}_2}^{1}(1-s)g_2(s)ds$, we have $\int_0^{\bar{s}_1} sg_1(s)ds \geq \int_0^{\bar{s}_2} sg_2(s)ds$. Let $r_1(\tilde{\beta})$ and $r_2(\tilde{\beta})$ be the associated IROC curves for $g_1(s)$ and $g_2(s)$. We then have $r_1(\tilde{\beta}) \geq r_2(\tilde{\beta}),\ \forall\ \tilde{\beta} \in [0,1]$.

Lemma 2 shows that as the virtual triage tool improves its triage capability, we could have a new IROC curve that lies below the original. Consequently, we can achieve higher virtual triage accuracy with lower virtual under-triage and over-triage probabilities.

Now, in the presence of virtual triage, there are four types of patients: $\hat{L}\tilde{L}$, $\hat{L}\tilde{H}$, $\hat{H}\tilde{L}$, and $\hat{H}\tilde{H}$, where the first letter denotes the patient's self-triage decision and the second denotes the virtual triage recommendation. For a given virtual triage accuracy $\tilde{\alpha}$ and $\tilde{\beta}$, we assume it is conditionally independent of self-triage and therefore patients' posterior probabilities of being H are $b_{\hat{T}\tilde{L}} = \frac{\tilde{\alpha} b_{\hat{T}}}{\tilde{\alpha} b_{\hat{T}} + (1-\tilde{\beta})(1-b_{\hat{T}})}$ and $b_{\hat{T}\tilde{H}} = \frac{(1-\tilde{\alpha}) b_{\hat{T}}}{(1-\tilde{\alpha}) b_{\hat{T}} + \tilde{\beta}(1-b_{\hat{T}})}$, and the associated arrival rates of each type of patients are $\lambda_{\hat{T}\tilde{L}} = [\tilde{\alpha} b_{\hat{T}} + (1-\tilde{\beta})(1-b_{\hat{T}})]\lambda_{\hat{T}}$ and $\lambda_{\hat{T}\tilde{H}} = [(1-\tilde{\alpha}) b_{\hat{T}} + \tilde{\beta}(1-b_{\hat{T}})]\lambda_{\hat{T}}$, where $\hat{T} \in \{\hat{L}, \hat{H}\}$.[8] Note that we assume the virtual triage service is provided free of charge, which reflects the current practice. This assumption allows us to focus on understanding the informational effect of virtual triage on patients' care-seeking behavior and acute care system performance.

3.3. Cost Parameters

Disutility of Waiting. We denote strategic patients' disutility of waiting per unit time by w_G at a GP and w_E at the ED.[9] Consistent with much of the literature on tiered healthcare systems, we denote the expected waiting time at a GP by a constant Q_G, independent of the arrival rate of strategic patients at GPs, λ_G (Zorc et al. 2023, Çakıcı and Mills 2020).[10] On the other hand, we denote the expected ED waiting time of strategic patients by $Q_E(\lambda_E)$, where $Q_E(\lambda_E)$ is assumed to be strictly increasing and convex in the arrival rate of strategic patients to the ED, λ_E.[11] Unlike a GP, the ED specializes in emergency medicine and is dedicated to acute care. Hence, strategic patients' experiences at the ED, particularly the expected waiting time, critically depend on the care-seeking behaviors of others. From a modeling perspective, the monotonicity and convexity of $Q_E(\lambda_E)$ are satisfied by common queueing models, including $M/M/c$ and $M/G/1$ under a first-in-first-out discipline. Practically, the convexity assumption captures the stochasticity of both the patient arrival process (ED visits are unscheduled, without prior appointments) and the treatment process (patients with different characteristics follow different care pathways) at the ED.

Acute Care System Service Cost. We assume the expected rates of GP and ED service costs caused by the arrivals of strategic patients, denoted by $S_G(\lambda_G)$ and $S_E(\lambda_E)$, are increasing and linear in λ_G and λ_E, with $S_G(\lambda_G) = a_G\lambda_G$ and $S_E(\lambda_E) = a_E\lambda_E$. Hence, a_G and a_E, respectively, denote the expected marginal service cost per strategic patient arrival to the GP and the ED. Clearly, it is less costly to have an H patient visit the ED directly than visit a GP first, as H

patients can only get treated at the ED. Meanwhile, we assume the ED service cost is large enough such that $a_E > a_G + w_G Q_G - w_E Q_E(\lambda_H)$. This ensures it is less costly to have an L patient visit a GP than visit the ED, which reflects the reality. Otherwise, it would be optimal to have all patients visit the ED directly regardless of their types, making the problem of self-triage and virtual triage irrelevant.

3.4. Equilibrium Characterization

Choice of Care. After self-triage, or after both self-triage and virtual triage in the presence of a virtual triage tool, strategic patients compare the expected patient cost (i.e., the sum of monetary payment and disutility of waiting) of visiting a GP first with the expected patient cost of going to the ED directly, and they choose the option with a lower cost. In particular, patients incur a monetary payment every time they visit a GP or the ED, denoted by the expected GP co-payment p_G and expected ED co-payment p_E. Hence, a patient with probability b of being H decides to visit a GP first if $p_G + w_G Q_G + b[p_E + w_E Q_E(\lambda_E)] \le p_E + w_E Q_E(\lambda_E)$ holds, or they visit the ED directly otherwise. We assume that the ED co-payment is large enough such that $p_E > p_G + w_G Q_G - w_E Q_E(\lambda_H)$, as otherwise all patients would visit the ED directly regardless of the accuracy of self-triage and virtual triage, which clearly does not reflect reality.

Let $\hat{\mathbf{f}}^{\mathbf{e}}(\hat{\alpha}, \hat{\beta}) = (f^e_{\hat{L}}, f^e_{\hat{H}})$ characterize the nonatomic Nash equilibrium (Schmeidler 1973)[12] patient flow in the absence of virtual triage, where $f^e_{\hat{T}} \in [0,1]$ is the probability of $\hat{T}$ patients visiting the ED directly in equilibrium, $\hat{T} \in \{\hat{L}, \hat{H}\}$. To simplify the exposition, we denote $f^e_{\hat{T}} = m$ when $f^e_{\hat{T}} \in (0,1)$, i.e., $\hat{T}$ patients adopt a mixed strategy in equilibrium. Similarly, $\tilde{\mathbf{f}}^{\mathbf{e}}(\tilde{\alpha}, \tilde{\beta}; \hat{\alpha}, \hat{\beta}) = (f^e_{\hat{L}\tilde{L}}, f^e_{\hat{L}\tilde{H}}, f^e_{\hat{H}\tilde{L}}, f^e_{\hat{H}\tilde{H}})$ denotes the equilibrium patient flow in the presence of virtual triage.[13]

Equilibrium Outcomes. Social cost and triage safety are the two main system performance metrics that we consider in this study. We define social cost $C_s(\cdot)$ as the sum of strategic patients' disutility of waiting and the service costs of GP and ED operations,[14] and we have the equilibrium social cost

$$C_s(\hat{\mathbf{f}}^{\mathbf{e}}) = \sum_{l \in \{G,E\}} \lambda_l(\hat{\mathbf{f}}^{\mathbf{e}}) w_l Q_l(\lambda_l(\hat{\mathbf{f}}^{\mathbf{e}})) + S_l(\hat{\mathbf{f}}^{\mathbf{e}})), \tag{1}$$

in the absence of virtual triage, where $\hat{\mathbf{f}}^{\mathbf{e}} \in [0,1]^2$, $\lambda_G(\hat{\mathbf{f}}^{\mathbf{e}}) = \sum_{\hat{T} \in \{\hat{L}, \hat{H}\}} (1 - f^e_{\hat{T}}) \lambda_{\hat{T}}$, and $\lambda_E(\hat{\mathbf{f}}^{\mathbf{e}}) = \sum_{\hat{T} \in \{\hat{L}, \hat{H}\}} (1 - f^e_{\hat{T}}) b_{\hat{T}} \lambda_{\hat{T}} + f^e_{\hat{T}} \lambda_{\hat{T}}$. Similarly, we have the equilibrium social cost

$$C_s(\tilde{\mathbf{f}}^{\mathbf{e}}) = \sum_{l \in \{G,E\}} \lambda_l(\tilde{\mathbf{f}}^{\mathbf{e}}) w_l Q_l(\lambda_l(\tilde{\mathbf{f}}^{\mathbf{e}})) + S_l(\tilde{\mathbf{f}}^{\mathbf{e}})), \tag{2}$$

in the presence of virtual triage, where $\tilde{\mathbf{f}}^{\mathbf{e}} \in [0,1]^4$, $\lambda_G(\tilde{\mathbf{f}}^{\mathbf{e}}) = \sum_{\hat{T}\tilde{T} \in \{\hat{L}, \hat{H}\} \times \{\tilde{L}, \tilde{H}\}} (1 - f^e_{\hat{T}\tilde{T}}) \lambda_{\hat{T}\tilde{T}}$, and $\lambda_E(\tilde{\mathbf{f}}^{\mathbf{e}}) = \sum_{\hat{T}\tilde{T} \in \{\hat{L}, \hat{H}\} \times \{\tilde{L}, \tilde{H}\}} (1 - f^e_{\hat{T}\tilde{T}}) b_{\hat{T}\tilde{T}} \lambda_{\hat{T}\tilde{T}} + f^e_{\hat{T}\tilde{T}} \lambda_{\hat{T}\tilde{T}}$.

Meanwhile, evidence from practice shows that virtual triage tools tend to heavily emphasize triage safety and, when in doubt, often recommend that patients seek emergency care (Semigran et al. 2015, Chambers et al. 2019). This approach aims to ensure that patients who require emergency care resources are not delayed in receiving any necessary treatment. Given this, we include a second performance metric in our analysis that is related to triage safety. Specifically, we focus on the total treatment delay associated with the fraction of ED-type patients that visit a GP first, with equilibrium treatment delay

$$D_s(\hat{\mathbf{f}}^{\mathbf{e}}) = \sum_{\hat{T}\in\{\hat{L},\hat{H}\}} (1 - f^e_{\hat{T}}) b_{\hat{T}} \lambda_{\hat{T}}, \tag{3}$$

in the absence of virtual triage. Similarly, we have the equilibrium treatment delay

$$D_s(\tilde{\mathbf{f}}^{\mathbf{e}}) = \sum_{\hat{T}\tilde{T}\in\{\hat{L},\hat{H}\}\times\{\tilde{L},\tilde{H}\}} (1 - f^e_{\hat{T}\tilde{T}}) b_{\hat{T}\tilde{T}} \lambda_{\hat{T}\tilde{T}}, \tag{4}$$

in the presence of virtual triage.

3.5. Model Extensions and Robustness

We note that our model formulation in Section 3 relies on a set of simplifying assumptions. To demonstrate the robustness, we extend the model by (1) allowing heterogeneity of patient self-triage accuracy, (2) including belief-dependent disutility of waiting, (3) incorporating other costs and disutilities for patients, and (4) relaxing the assumptions on full information and rationality of patients. Detailed discussions are presented in EC.5 of the e-companion. We show that the structural insights continue to hold, and therefore establish the robustness of our findings.

4. Impact of Virtual Triage on Patient Behavior

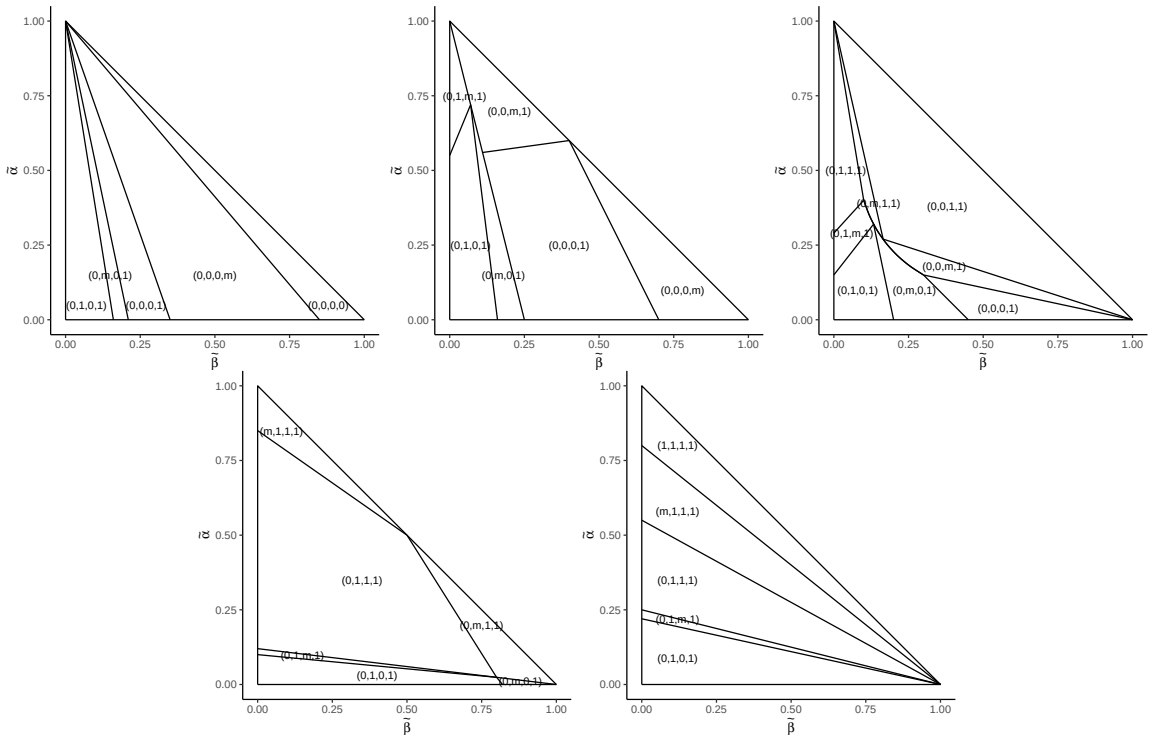
While virtual triage provides recommendations, patients still make the final judgements and decide whether to follow virtual triage recommendations or not (Agrawal et al. 2022). Hence, we first examine how the adoption of virtual triage may or may not change patient care-seeking behavior.

4.1. Patient Care-Seeking Behavior in Equilibrium

The next proposition establishes the uniqueness of equilibrium patient flow without and with virtual triage. It also characterizes the equilibrium regions under varying self-triage and virtual triage accuracy levels.

PROPOSITION 1. *$\forall\ \hat{\alpha} + \hat{\beta} \le 1$, there exists a unique equilibrium patient flow $\hat{\mathbf{f}}^{\mathbf{e}} = (f^e_{\hat{L}}, f^e_{\hat{H}})$ in the absence of virtual triage; in addition, $\forall\ \tilde{\alpha} + \tilde{\beta} \le 1$, there exists a unique equilibrium patient flow $\tilde{\mathbf{f}}^{\mathbf{e}} = (f^e_{\hat{L}\tilde{L}}, f^e_{\hat{L}\tilde{H}}, f^e_{\hat{H}\tilde{L}}, f^e_{\hat{H}\tilde{H}})$ in the presence of virtual triage. The different equilibrium regions in the presence of virtual triage are summarized in Table 1. Moreover, depending on the values of $\hat{\mathbf{f}}^{\mathbf{e}}$, the different subsets of equilibrium regions for $\tilde{\mathbf{f}}^{\mathbf{e}}$ in the presence of virtual triage are as follows:*

Figure 2 Equilibrium regions of patient flow in the presence of virtual triage when $\hat{\mathbf{f}}^{\mathbf{e}} = (0,0)$ (top left), $\hat{\mathbf{f}}^{\mathbf{e}} = (0,m)$ (top middle), $\hat{\mathbf{f}}^{\mathbf{e}} = (0,1)$ (top right), $\hat{\mathbf{f}}^{\mathbf{e}} = (m,1)$ (bottom left), $\hat{\mathbf{f}}^{\mathbf{e}} = (1,1)$ (bottom right).[15]



(i) *When* $\hat{\mathbf{f}}^{\mathbf{e}} = (0,0)$, *we have* $\tilde{\mathbf{f}}^{\mathbf{e}} \in \{(0,0,0,0),(0,0,0,m),(0,0,0,1),(0,m,0,1),(0,1,0,1)\}$.

(ii) *When* $\hat{\mathbf{f}}^{\mathbf{e}} = (0,m)$, *we have* $\tilde{\mathbf{f}}^{\mathbf{e}} \in \{(0,0,0,m),(0,0,0,1),(0,0,m,1),(0,m,0,1),(0,1,0,1),(0,1,m,1)\}$.

(iii) *When* $\hat{\mathbf{f}}^{\mathbf{e}} = (0,1)$, *we have* $\tilde{\mathbf{f}}^{\mathbf{e}} \in \{(0,0,0,1),(0,0,m,1),(0,0,1,1),(0,m,1,1),(0,1,1,1),(0,1,m,1),(0,1,0,1),(0,m,0,1)\}$.

(iv) *When* $\hat{\mathbf{f}}^{\mathbf{e}} = (m,1)$, *we have* $\tilde{\mathbf{f}}^{\mathbf{e}} \in \{(m,1,1,1),(0,1,1,1),(0,m,1,1),(0,1,m,1),(0,1,0,1),(0,m,0,1)\}$.

(v) *When* $\hat{\mathbf{f}}^{\mathbf{e}} = (1,1)$, *we have* $\tilde{\mathbf{f}}^{\mathbf{e}} \in \{(1,1,1,1),(m,1,1,1),(0,1,1,1),(0,1,m,1),(0,1,0,1)\}$.

The relative position of each region is shown in Figure 2.

$R_{p,\nsim}$	$R_{p,\sim}$	$R_{m,\tilde{L}}$	$R_{m,\tilde{H}}$
(0, 0, 0, 0)	(0, 0, 0, 1)	(0, 0, m, 1)	(0, 0, 0, m)
(0, 0, 1, 1)	(0, 1, 0, 1)	(0, 1, m, 1)	(0, m, 0, 1)
(1, 1, 1, 1)	(0, 1, 1, 1)	(m, 1, 1, 1)	(0, m, 1, 1)

Table 1 Different equilibrium regions in the presence of virtual triage. $R_{p,\nsim}$ and $R_{p,\sim}$ includes pure strategy equilibrium regions, while $R_{m,\tilde{L}}$ and $R_{m,\tilde{H}}$ includes mixed strategy equilibrium regions.

Proposition 1 characterizes the uniqueness of equilibrium patient flow and all possible equilibrium regions without and with virtual triage. In particular, 12 different equilibrium regions could emerge after the introduction of virtual triage, and we categorize them into four types, as

shown in Table 1. $R_{p,\asymp} = \{(0,0,0,0),(0,0,1,1),(1,1,1,1)\}$ includes three pure strategy equilibrium regions where the adoption of virtual triage has no impact on patients' care-seeking behavior; $R_{p,\sim} = \{(0,0,0,1),(0,1,0,1),(0,1,1,1)\}$ includes three pure strategy equilibrium regions where the adoption of virtual triage affects patients' care-seeking behavior and patients follow only $\tilde{L}$ recommendations, or only $\tilde{H}$ recommendations, or all recommendations from virtual triage with certainty. On the other hand, $R_{m,\tilde{L}} = \{(0,0,m,1),(0,1,m,1),(m,1,1,1)\}$ includes three mixed strategy equilibrium regions where $\hat{L}\tilde{L}$ or $\hat{H}\tilde{L}$ patients adopt a mixed strategy in equilibrium; $R_{m,\tilde{H}} = \{(0,0,0,m),(0,m,0,1),(0,m,1,1)\}$ includes three mixed strategy equilibrium regions where $\hat{L}\tilde{H}$ or $\hat{H}\tilde{H}$ patients adopt a mixed strategy in equilibrium.

Moreover, Proposition 1 shows that depending on patient self-triage accuracy and therefore patient care-seeking behavior in the absence of virtual triage, different subsets of equilibrium regions emerge in the presence of virtual triage. When $\hat{\mathbf{f}}^{\mathbf{e}} = (0,0)$, all patients go to a GP first in the absence of virtual triage. After introducing virtual triage, patients receiving a signal $\tilde{L}$ will follow this recommendation and go to a GP first, regardless of the virtual triage accuracy. This is because an $\tilde{L}$ recommendation only serves to further reduce their probability of being H. Hence, we have $f^e_{\hat{L}\tilde{L}} = f^e_{\hat{H}\tilde{L}} = 0$ in this case, resulting in Proposition 1 (i). Following a similar argument, when $\hat{\mathbf{f}}^{\mathbf{e}} = (1,1)$, we have $f^e_{\hat{L}\tilde{H}} = f^e_{\hat{H}\tilde{H}} = 1$, resulting in Proposition 1 (v). When $\hat{\mathbf{f}}^{\mathbf{e}} = (0,m)$, $\hat{H}$ patients adopt a mixed strategy in equilibrium in the absence of virtual triage. Hence, it is not possible to simultaneously have $f^e_{\hat{H}\tilde{L}} = f^e_{\hat{H}\tilde{H}} = 0$, or $f^e_{\hat{H}\tilde{L}} = f^e_{\hat{H}\tilde{H}} = 1$, resulting in Proposition 1 (ii). Following a similar argument, when $\hat{\mathbf{f}}^{\mathbf{e}} = (m,1)$, we cannot simultaneously have $f^e_{\hat{L}\tilde{L}} = f^e_{\hat{L}\tilde{H}} = 0$, or $f^e_{\hat{H}\tilde{L}} = f^e_{\hat{H}\tilde{H}} = 1$, resulting in Proposition 1 (iv). Meanwhile, when $\hat{\mathbf{f}}^{\mathbf{e}} = (0,1)$, patient self under-triage and self over-triage probabilities, $\hat{\alpha}$ and $\hat{\beta}$, do not differ by much, such that patients follow their self-triage decisions in the absence of virtual triage. In this case, when virtual triage recommendations confirm self-triage decisions, patients always follow virtual triage recommendations in equilibrium regardless of their accuracy, i.e., $f^e_{\hat{L}\tilde{L}} = 0$ and $f^e_{\hat{H}\tilde{H}} = 1$. This results in Proposition 1 (iii).

Finally, the relative positions of the equilibrium regions in the presence of virtual triage for each of these five cases (as shown in Figure 2) can be explained as follows. As $\tilde{\alpha}$ decreases, GP recommendations become more informative and $\tilde{L}$ patients are more likely to visit a GP ($\hat{L}\tilde{L}$ patients first followed by $\hat{H}\tilde{L}$ patients, since $b_{\hat{L}\tilde{L}} < b_{\hat{H}\tilde{L}}$). Similarly, as $\tilde{\beta}$ decreases, ED recommendations become more informative and $\tilde{H}$ patients are more likely to visit the ED directly ($\hat{H}\tilde{H}$ patients first followed by $\hat{L}\tilde{H}$ patients). A full discussion of the relative positions of the equilibrium regions in each of the five cases is provided in EC.1 of the e-companion.

4.2. Patient Compliance with Virtual Triage Recommendations

The medical community is increasingly interested in empirically evaluating the accuracy of virtual triage and patients' compliance with virtual triage recommendations. Semigran et al. (2015) and

Chambers et al. (2019) have conducted extensive studies and found that virtual triage recommendations tend to encourage patients to seek emergency care. As an example, Baker et al. (2020) also show that Babylon Health's virtual triage tools tend to recommend higher intensity of care than doctors on average, leading to a higher chance of over-triage. This problem has prompted widespread concern that the adoption of virtual triage could lead to an increase in ED visits by the so-called "worried well," thereby worsening the ED overcrowding problem (Turbitt and Freed 2015). However, Chambers et al. (2019) found that while there is generally good agreement between virtual triage recommendations and patients' intended actions, patients who are advised to go to an ED are more likely to seek primary care. This tendency in fact leads to delayed emergency care seeking and a decrease in ED visits. Our model provides a potential explanation that could reconcile and rationalize these two seemingly conflicting empirical findings.

PROPOSITION 2. *Suppose patients adopt a pure strategy in equilibrium in the absence of virtual triage.*

(i) *When* $\hat{\mathbf{f}}^{\mathbf{e}} \in \{(0,1),(1,1)\}$*, there exists virtual triage such that* $\exists\ \bar{s}_u \in (0,1)$ *s.t.* $\forall\ \bar{s} \in (0,\bar{s}_u)$*, we have* $\lambda_G(\hat{\mathbf{f}}^{\mathbf{e}}) < \lambda_G(\tilde{\mathbf{f}}^{\mathbf{e}})$ *and* $\lambda_E(\hat{\mathbf{f}}^{\mathbf{e}}) > \lambda_E(\tilde{\mathbf{f}}^{\mathbf{e}})$.

(ii) *When* $\hat{\mathbf{f}}^{\mathbf{e}} \in \{(0,0),(0,1)\}$*, there exists virtual triage such that* $\exists\ \bar{s}_l \in (0,1)$ *s.t.* $\forall\ \bar{s} \in (\bar{s}_l,1)$*, we have* $\lambda_G(\hat{\mathbf{f}}^{\mathbf{e}}) > \lambda_G(\tilde{\mathbf{f}}^{\mathbf{e}})$ *and* $\lambda_E(\hat{\mathbf{f}}^{\mathbf{e}}) < \lambda_E(\tilde{\mathbf{f}}^{\mathbf{e}})$.

Proposition 2 shows that when virtual triage excessively recommends emergency (primary) care, i.e., $\bar{s}$ is small (large), it may in fact lead to a decrease in ED (GP) visits. These findings highlight a significant way in which virtual triage differs from traditional ED triage: Due to its decentralized nature, patients may not necessarily follow virtual triage recommendations. In particular, when virtual triage excessively (but not unconditionally) recommends emergency care to a high volume of patients, an ED recommendation made by virtual triage carries little information. On the other hand, a recommendation to see a GP in this case is highly informative, though a lower volume of patients will receive this signal. Hence, there is an underlying *informativeness-volume trade-off* of virtual triage recommendations subjective to a given IROC curve. Patients who are advised to seek primary care will then tend to follow the virtual triage recommendations and visit a GP; meanwhile, patients who are advised to visit an ED will tend to ignore the recommendations and continue to follow their prior self-triage decisions. As a result, the net effect is such that the adoption of virtual triage that errs on the side of caution by excessively recommending emergency care could lead to a decrease in ED visits. Similar arguments hold when virtual triage excessively recommends primary care, which could lead to a decrease in GP visits.

By contrast, when patient self-triage accuracy is poor, patients may still follow virtual triage recommendations even when they excessively recommend primary or emergency care. This occurs

because virtual triage recommendations are still more informative than patient self-triage decisions. Hence, the first order intuition that excessively recommending emergency (primary) care will lead to an increase in ED (GP) visits (Turbitt and Freed 2015) may still hold, but not always. Our results thus provide a more holistic understanding of this problem. In particular, acute care systems and virtual triage technology operators should be aware of the potential for unexpected changes in patient care-seeking behavior arising from the informativeness-volume trade-off of virtual triage.

5. Impact of Off-the-shelf Virtual Triage on System Performance

While virtual triage tools are being increasingly deployed, considerable knowledge gaps persist regarding their potential impact on existing acute care systems. Furthermore, it remains unclear how these systems should accommodate the technology and what role system factors should play in guiding scoring function choices and accuracy calibration. Therefore, in this section, we examine the potential consequences of deploying a virtual triage tool with preset accuracy, reflecting the current trend of utilizing "off-the-shelf" virtual triage solutions in practice.

5.1. Impact of Off-the-Shelf Virtual Triage on Social Cost

We now analyze the effect of higher virtual triage accuracy on equilibrium social cost, where the discrimination threshold $\bar{s}$ is chosen with some scoring function that is uninformed by acute care system parameters, resulting in an off-the-shelf virtual triage with an exogenous accuracy $\tilde{\alpha}$ and $\tilde{\beta}$. We first characterize the effect of higher virtual triage accuracy on equilibrium social cost when all patients adopt pure strategies in equilibrium.[16]

Lemma 3.

(i) *When* $\tilde{\mathbf{f}}^{\mathbf{e}} \in R_{p,\nsim}$, *we have* $\frac{dC_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\alpha}} = \frac{dC_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\beta}} = 0$.

(ii) *When* $\tilde{\mathbf{f}}^{\mathbf{e}} \in R_{p,\sim}$, *we have* $\frac{dC_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\alpha}} > 0$ *and* $\frac{dC_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\beta}} > 0$.

Lemma 3 (i) directly follows from the fact that virtual triage does not change patient care-seeking behavior when $\tilde{\mathbf{f}}^{\mathbf{e}} \in R_{p,\nsim}$. For Lemma 3 (ii), when $\tilde{\mathbf{f}}^{\mathbf{e}} \in R_{p,\sim}$, we have certain types of patients follow virtual triage recommendations that are different from their self-triage decisions, and all patients adopt pure strategies in equilibrium. As a result, higher virtual triage accuracy does not change the behavior of each patient type, but only the composition of each patient type. In this case, lower $\tilde{\alpha}$ reduces $\lambda_G(\tilde{\mathbf{f}}^{\mathbf{e}})$, while $\lambda_E(\tilde{\mathbf{f}}^{\mathbf{e}})$ is independent of $\tilde{\alpha}$ (as patients who are virtual under-triaged will go to the ED regardless, either directly or referred by a GP). Hence $C_s(\tilde{\mathbf{f}}^{\mathbf{e}})$ decreases with lower $\tilde{\alpha}$. Meanwhile, if $\tilde{\beta}$ is lower, fewer patients will be virtual over-triaged to go to the ED directly. This will reduce the total ED arrival rate and increase the GP arrival rate by the same amount. Since a patient visit to a GP is less costly than a visit to the ED, $C_s(\tilde{\mathbf{f}}^{\mathbf{e}})$ decreases with lower $\tilde{\beta}$.

We then characterize the effect of higher virtual triage accuracy on equilibrium social cost when certain patient type adopts a mixed strategy in equilibrium after the introduction of virtual triage.

Lemma 4.

(i) *When* $\tilde{\mathbf{f}}^{\mathbf{e}} \in R_{m,\tilde{H}}$, $\exists\ a_E^u, a_E^o$ *s.t.*
   (a) $\frac{dC_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\alpha}} < 0$ *if and only if* $a_E > a_E^u$;
   (b) $\frac{dC_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\beta}} < 0$ *if and only if* $a_E > a_E^o$.

(ii) *When* $\tilde{\mathbf{f}}^{\mathbf{e}} \in R_{m,\tilde{L}}$, $\exists\ a_E^u, a_E^o$ *s.t.*
   (a) $\frac{dC_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\alpha}} < 0$ *if and only if* $a_E < a_E^u$;
   (b) $\frac{dC_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\beta}} < 0$ *if and only if* $a_E < a_E^o$.

The effect of virtual triage accuracy on equilibrium social cost for mixed strategy equilibrium regions is more complicated: a higher virtual triage accuracy not only changes equilibrium social cost via the changes in composition of each patient type (*direct informational effect*), it also changes the behavior of those who adopt mixed strategies, which in turn changes the equilibrium social cost (*indirect behavioral effect*). This is in contrast to the cases under which all patient types adopt pure strategy in equilibrium, for which only the direct informational effect exists, as characterized by Lemma 3. In particular, when $\tilde{H}$ patients adopt mixed strategies, both lower $\tilde{\alpha}$ and lower $\tilde{\beta}$ lead to mixed strategy $\tilde{H}$ patients visiting the ED directly with a higher probability. In the case of lower $\tilde{\alpha}$, while the ED arrival rate and therefore the expected waiting time at the ED do not directly depend on $\tilde{\alpha}$, lower $\tilde{\alpha}$ leads to higher posterior belief of being H for mixed strategy $\tilde{H}$ patients. Hence, mixed strategy $\tilde{H}$ patients will visit the ED directly with a higher probability as $\tilde{\alpha}$ decreases. As a result, the net effect of lower $\tilde{\alpha}$ on equilibrium social cost when $\tilde{H}$ patients adopt mixed strategies is as follows (Lemma 4 (i)(a)):

- When $a_E < a_E^u$, lower $\tilde{\alpha}$ reduces the equilibrium social cost via the direct informational effect, and the direct informational effect dominates over the indirect behavioral effect. As a result, the net effect is such that equilibrium social cost decreases with lower $\tilde{\alpha}$.
- When $a_E > a_E^u$, lower $\tilde{\alpha}$ reduces the equilibrium social cost via the direct informational effect. Meanwhile, the indirect behavioral effect increases the equilibrium social cost and is strong: lower $\tilde{\alpha}$ leads to mixed strategy $\tilde{H}$ patients visiting the ED directly with a higher probability, which leads to large negative cost externality at the ED due to large a_E. As a result, when a_E is sufficiently large, the indirect behavioral effect dominates over the direct informational effect, and the net effect is such that the equilibrium social cost increases with lower $\tilde{\alpha}$.

On the other hand, lower $\tilde{\beta}$ not only leads to a higher posterior belief of being H for mixed strategy $\tilde{H}$ patients, but it also reduces the ED arrival rate and therefore the expected waiting time at the ED. Hence, lower $\tilde{\beta}$ also leads to mixed strategy $\tilde{H}$ patients visiting the ED directly with a higher probability. Meanwhile, the net effect of lower $\tilde{\beta}$ on equilibrium social cost when $\tilde{H}$ patients adopt mixed strategies follows the same argument (Lemma 4 (i)(b)).

Conversely, Lemma 4 (ii) shows that when $\tilde{L}$ patients instead adopt mixed strategies, lower $\tilde{\alpha}$ ($\tilde{\beta}$) leads to a higher equilibrium social cost when the ED service cost is relatively small, i.e., $a_E < a_E^u$ ($a_E < a_E^o$). This is because lower $\tilde{\alpha}$ or $\tilde{\beta}$ will lead to a lower posterior belief of being H for $\tilde{L}$ patients who adopt mixed strategies, and these patients tend to visit a GP first with a higher probability. In this case, these patients generate large negative cost externality at GPs. As a result, the indirect behavioral effect dominates over the direct informational effect if a_E is sufficiently small, and the equilibrium social cost increases with lower $\tilde{\alpha}$ or $\tilde{\beta}$.

Having characterized the different equilibrium regions (Proposition 1) and the potential non-monotonicity of equilibrium social cost in off-the-shelf virtual triage accuracy (Lemma 4), we show that for arbitrary patient self-triage accuracy (or equivalently, arbitrary equilibrium outcome in the absence of virtual triage), the adoption of an informative off-the-shelf virtual triage tool could worsen equilibrium social cost.

PROPOSITION 3. $\forall\ \hat{\alpha}+\hat{\beta} \leq 1$, *or equivalently,* $\forall\ \hat{\mathbf{f}}^{\mathbf{e}} \in [0,1]^2$, $\exists\ \tilde{\alpha}+\tilde{\beta} < 1$ *and* a_E *s.t.* $C_s(\hat{\mathbf{f}}^{\mathbf{e}}) < C_s(\tilde{\mathbf{f}}^{\mathbf{e}})$.

Proposition 3 shows that the adoption of informative off-the-shelf virtual triage could worsen equilibrium social cost even with arbitrarily poor patient self-triage accuracy. The existence of such outcomes is driven by the relative accuracy of virtual triage to self-triage. On the one hand, when virtual triage accuracy is much higher than self-triage accuracy, the adoption of virtual triage is likely to help. Specifically, virtual triage recommendations not only change patient care-seeking behavior with certainty, but also the probability of virtual over-triage or virtual under-triage is small; therefore, the social cost that arises from the mismatch of acute care resources is reduced. On the other hand, when virtual triage accuracy is much lower than self-triage accuracy, the adoption of virtual triage is likely to have no impact: Its informativeness is limited and therefore virtual triage recommendations do not change patient care-seeking behavior. However, when virtual triage accuracy is similar to or marginally more accurate than self-triage accuracy, neither patient self-triage decisions nor virtual triage recommendations dominate. This scenario can result in mixed strategy equilibria, creating potential for higher equilibrium social cost after adopting virtual triage. Thus, when evaluating the impact of off-the-shelf virtual triage adoption on social cost, patient self-triage accuracy and acute care system parameters should be taken into account.

5.2. Impact of Off-the-Shelf Virtual Triage on Triage Safety

We next analyze the effect of higher virtual triage accuracy on equilibrium treatment delay and the implications for triage safety in an off-the-shelf setting.

LEMMA 5.

(i) *When* $\tilde{\mathbf{f}}^{\mathbf{e}} \in R_{p,\infty}$, *we have* $\frac{dD_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\alpha}} = \frac{dD_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\beta}} = 0$.

(ii) *When* $\tilde{\mathbf{f}}^{\mathbf{e}} \in R_{p,\sim}$, *we have* $\frac{dD_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\alpha}} > 0$ *and* $\frac{dD_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\beta}} = 0$.

As in Lemma 3 (i), Lemma 5 (i) follows from the fact that virtual triage does not change patient care-seeking behavior when $\tilde{\mathbf{f}}^{\mathbf{e}} \in R_{p,\sim}$. For Lemma 5 (ii), when $\tilde{\mathbf{f}}^{\mathbf{e}} \in R_{p,\sim}$, virtual triage does change patient behavior and patients follow either their self-triage decisions or virtual triage recommendations with certainty. In this case, for patients that follow virtual triage recommendations, lower $\tilde{\alpha}$ reduces under-triaged H patients and therefore improves triage safety. On the other hand, $\tilde{\beta}$ is unrelated to under-triage and therefore has no impact on triage safety.

We now characterize the effect of higher virtual triage accuracy on equilibrium treatment delay when certain patient type adopts a mixed strategy after the introduction of virtual triage.

Lemma 6.

(i) *When* $\tilde{\mathbf{f}}^{\mathbf{e}} \in R_{m,\tilde{H}}$, *we have*
   (a) $\frac{dD_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\alpha}} > 0$;
   (b) $\frac{dD_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\beta}} > 0$.

(ii) *When* $\tilde{\mathbf{f}}^{\mathbf{e}} \in R_{m,\tilde{L}}$, $\exists\ p_G^u, p_G^o$ *s.t.*
   (a) $\frac{dD_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\alpha}} < 0$ *if and only if* $p_G > p_G^u$;
   (b) $\frac{dD_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\beta}} < 0$ *if and only if* $p_G < p_G^o$.

Lemma 6 (i) shows that when $\tilde{H}$ patients adopt mixed strategy, both lower $\tilde{\alpha}$ and lower $\tilde{\beta}$ reduce the equilibrium treatment delay. In particular, while lower $\tilde{\beta}$ has no direct informational effect on the equilibrium treatment delay (as shown by Lemma 5 (ii)), it has an indirect behavioral effect via changes in the probability of mixed strategy patients visiting the ED directly. With lower $\tilde{\beta}$, mixed strategy patients have a higher posterior belief of being H patients. As a result, they will visit the ED directly with a higher probability, reducing the overall equilibrium treatment delay.

Lemma 6 (ii) shows that instead when $\tilde{L}$ patients adopt mixed strategy, the effect of higher virtual triage accuracy on the equilibrium treatment delay depends on GP co-payment. While lower $\tilde{\alpha}$ reduces the equilibrium treatment delay through the direct informational effect, it also reduces the posterior belief of mixed strategy patients being H patients, which in turn has an indirect behavioral effect that reduces their probability of going to the ED directly. This will increase the equilibrium treatment delay. When the GP co-payment is large, the posterior belief of mixed strategy patients being H patients is highly sensitive to $\tilde{\alpha}$ and reduces significantly with lower $\tilde{\alpha}$. Therefore, mixed strategy patients will visit a GP first with a significantly higher probability. As a result, the indirect behavioral effect dominates over the direct informational effect, and the net effect is such that lower $\tilde{\alpha}$ can increase the equilibrium treatment delay.

On the other hand, while lower $\tilde{\beta}$ has no direct effect on the equilibrium treatment delay, it reduces the posterior belief of mixed strategy patients being H patients, which in turn has an

indirect behavioral effect that reduces their probability of going to the ED directly. Further, lower $\tilde{\beta}$ also reduces the ED arrival rate of pure strategy patients and therefore the ED overcrowding problem. As a result, it also has an indirect behavioral effect that increases mixed strategy patients' probability of going to the ED directly. When the GP co-payment is small, the former indirect effect dominates the later indirect effect, and therefore the net effect is such that lower $\tilde{\beta}$ reduces mixed strategy patients' probability of going to the ED directly and therefore increase the overall equilibrium treatment delay.

5.3. Practical Implications

These findings shed light on the practical and regulatory challenges surrounding the adoption of medical predictive technologies with the prevailing off-the-shelf approach. We show that in an unregulated environment, the adoption of an informative off-the-shelf virtual triage tool with reasonably high accuracy could lead to worse system performance. Furthermore, even after adoption, as the precision of virtual triage improves it might inadvertently lead to a deterioration in both social cost and treatment delay. This suggests that upgrading a virtual triage algorithm for enhanced accuracy – without comprehensive reassessment and supplementary regulatory clearance – can potentially negate any formerly demonstrated benefits. Given this, periodic re-certification for subsequent versions of these tools becomes essential to ensure their real-world efficacy.

Moreover, we characterize how acute care system parameters moderate the impact of off-the-shelf virtual triage adoption, and how they differ across performance metrics. In particular, while the impact on social cost of an off-the-shelf virtual triage tool depends on the ED service cost, the GP co-payment plays the key role for the impact on triage safety. As a result, when opting for an off-the-shelf virtual triage technology, it is important to ensure that the tool's accuracy aligns with the cost dynamics of the acute care systems and the selected performance criteria.

6. Unlocking the Operational Benefits with Custom Virtual Triage

In the previous section, we assessed the consequences and challenges associated with deploying off-the-shelf virtual triage technology. As detailed in Section 5, we identified a key inefficiency in equilibrium outcomes: the potential misalignment of the chosen scoring function and the performance metric of acute care systems, given a specific IROC curve. This misalignment can lead to increased social costs and reduced triage safety when the virtual triage tool is not tailored to the specificities of its context. To address this, we now investigate the capabilities of custom virtual triage. By optimizing the discrimination threshold of virtual triage in accordance with a particular performance metric and system parameters, it presents the potential to unlock operational advantages for acute care systems.

6.1. Impact of Custom Virtual Triage on Social Cost

We characterize the optimal virtual triage accuracy for minimizing equilibrium social cost, $(\tilde{\alpha}^*, \tilde{\beta}^*)$, subject to the constraint that $\tilde{\alpha} = r(\tilde{\beta}), \tilde{\beta} \in [0,1]$ under existing cost parameters and patient co-payments. Clearly, compared with the equilibrium social cost in the absence of virtual triage, equilibrium social cost after the adoption of virtual triage will not be increased by endogenizing $\tilde{\beta}$: $\bar{s}$ can be set to either 0 or 1, in which case we have $r(\tilde{\beta}) + \tilde{\beta} = 1$, and patients' posteriors remain the same as their priors. However, the introduction of virtual triage in this case has no impact, and therefore this does not occur in practice. Instead, an acute care system might adopt virtual triage with the goal of minimizing the equilibrium social cost. This objective is equivalent to using the equilibrium social cost as the scoring function to determine the optimal virtual over-triage probability $\tilde{\beta}^*$ and virtual under-triage probability $\tilde{\alpha}^* = r(\tilde{\beta}^*)$ for a given IROC curve.

The minimization of equilibrium social cost is complicated by the fact that the IROC curve can intersect with different subsets of equilibrium regions whose relative positions in the presence of virtual triage depend on patients' care-seeking behavior in the absence of virtual triage. This makes a complete analytical characterization of $r(\tilde{\beta}^*)$ and $\tilde{\beta}^*$ infeasible. Nevertheless, we can characterize the optimal virtual triage accuracy if it leads to a pure strategy equilibrium. Specifically, if all patients adopt pure strategies in equilibrium regions $R_{p,\sim}$ under $r(\tilde{\beta}^*)$ and $\tilde{\beta}^*$, then the adoption of virtual triage has no impact. Furthermore, if all patients adopt pure strategies in equilibrium regions $R_{p,\sim}$ under $r(\tilde{\beta}^*)$ and $\tilde{\beta}^*$, we can show that $\tilde{\beta}^*$ is unique and given by the following proposition.

PROPOSITION 4. *For a given virtual triage tool, let $(r(\tilde{\beta}^*), \tilde{\beta}^*)$ denote the optimal virtual triage accuracy that minimizes equilibrium social cost subject to the IROC curve $r(\tilde{\beta})$. Suppose $\tilde{\mathbf{f}}^{\mathbf{e}}(r(\tilde{\beta}^*), \tilde{\beta}^*)$ is in the interior of $R_{p,\sim}$.*[17] *Then, there exists a unique $\tilde{\beta}^*$ given by the solution to*

$$(a_G + w_G Q_G)[\lambda_o - r'(\tilde{\beta}^*)\lambda_u] = [a_E + w_E Q_E(\lambda_E) + \lambda_E w_E \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}]\lambda_o, \tag{5}$$

where we have

(i) $\lambda_o = (1 - b_{\hat{H}})\lambda_{\hat{H}}$ *and* $\lambda_u = b_{\hat{H}}\lambda_{\hat{H}}$ *if* $\tilde{\mathbf{f}}^{\mathbf{e}}(r(\tilde{\beta}^*), \tilde{\beta}^*) = (0,0,0,1)$;

(ii) $\lambda_o = (1 - b_{\hat{L}})\lambda_{\hat{L}}$ *and* $\lambda_u = b_{\hat{L}}\lambda_{\hat{L}}$ *if* $\tilde{\mathbf{f}}^{\mathbf{e}}(r(\tilde{\beta}^*), \tilde{\beta}^*) = (0,1,1,1)$;

(iii) $\lambda_o = \lambda_L$ *and* $\lambda_u = \lambda_H$ *if* $\tilde{\mathbf{f}}^{\mathbf{e}}(r(\tilde{\beta}^*), \tilde{\beta}^*) = (0,1,0,1)$.

Proposition 4 captures the trade-off faced by the virtual triage provider when endogenizing $\tilde{\beta}$ in pure strategy equilibrium regions $R_{p,\sim}$. While a higher $\tilde{\beta}$ increases the ED arrival rate, it also reduces the GP arrival rate, and leads to a lower $\tilde{\alpha} = r(\tilde{\beta})$ which further reduces the GP arrival rate. Meanwhile, in determining the optimal virtual triage accuracy, the GP arrival rate reduction is weighted by the cost externality of a GP arrival, while the ED arrival rate increase is weighted by the cost externality of an ED arrival. We can show that in pure strategy regions,

the equilibrium social cost is concave in $\tilde{\beta}$. Therefore, the optimal $r(\tilde{\beta}^*)$ and $\tilde{\beta}^*$ are achieved when the marginal equilibrium social cost increase due to higher patient volume at the ED equals the marginal equilibrium social cost reduction arising from lower patient volume at GPs. In particular, Proposition 4 shows that:

- When $\tilde{\mathbf{f}}^{\mathbf{e}}(r(\tilde{\beta}^*), \tilde{\beta}^*) = (0, 0, 0, 1)$, virtual triage recommendations can only modify the care-seeking behavior of $\hat{H}$ patients: with a higher $\tilde{\beta}$ and a lower $\tilde{\alpha} = r(\tilde{\beta})$, L patients in $\hat{H}$ patients (with an arrival rate $(1 - b_{\hat{H}})\lambda_{\hat{H}}$) are more likely to be virtual over-triaged, while H patients in $\hat{H}$ patients (with an arrival rate $b_{\hat{H}}\lambda_{\hat{H}}$) are less likely to be virtual under-triaged.
- When $\tilde{\mathbf{f}}^{\mathbf{e}}(r(\tilde{\beta}^*), \tilde{\beta}^*) = (0, 1, 1, 1)$, virtual triage recommendations can only modify the care-seeking behavior of $\hat{L}$ patients: with a higher $\tilde{\beta}$ and a lower $\tilde{\alpha} = r(\tilde{\beta})$, L patients in $\hat{L}$ patients (with an arrival rate $(1 - b_{\hat{L}})\lambda_{\hat{L}}$) are more likely to be virtual over-triaged, while H patients in $\hat{L}$ patients (with an arrival rate $b_{\hat{L}}\lambda_{\hat{L}}$) are less likely to be virtual under-triaged.
- When $\tilde{\mathbf{f}}^{\mathbf{e}}(r(\tilde{\beta}^*), \tilde{\beta}^*) = (0, 1, 0, 1)$, all patients follow virtual triage recommendations: with a higher $\tilde{\beta}$ and a lower $\tilde{\alpha} = r(\tilde{\beta})$, all L patients (with an arrival rate λ_L) are more likely to be virtual over-triaged, while all H patients (with an arrival rate λ_H) are less likely to be virtual under-triaged.

While the above holds only for pure strategy regions, we show later using extensive numerical analysis (see Section 7.3) that patients rarely adopt mixed strategies in equilibrium under $r(\tilde{\beta}^*)$ and $\tilde{\beta}^*$, due to the inefficiency as characterized by Lemma 4. Next, we analyze how changes in system parameters affect $r(\tilde{\beta}^*)$ and $\tilde{\beta}^*$. In particular, we examine the impact of system parameters from two perspectives: *patient* composition and the triage capabilities of *virtual triage*.

Patient composition. We first characterize how changes in the base rate of H patients affect the optimal virtual triage accuracy in pure strategy equilibrium regions by the following proposition, where we denote the fraction of H patients by $h \in [0, 1]$, where $\lambda_H = h\lambda$.

PROPOSITION 5. *Suppose* $\tilde{\mathbf{f}}^{\mathbf{e}}(r(\tilde{\beta}^*), \tilde{\beta}^*)$ *is in the interior of* $R_{p,\sim}$. $\exists\ a_E^h$ *s.t. we have* $\frac{\partial \tilde{\beta}^*}{\partial h} > 0$ *if and only if* $a_E > a_E^h$.

Proposition 5 shows that in order to minimize the equilibrium social cost in pure strategy equilibrium regions, when the fraction of H patients in the patient base increases, the virtual triage provider would need to send more patients to the ED directly if and only if the ED services are sufficiently costly. This interesting effect occurs because the optimal virtual triage accuracy $(r(\tilde{\beta}^*), \tilde{\beta}^*)$ is achieved when the marginal equilibrium social cost increase at the ED (due to larger $\tilde{\beta}$) equals the marginal equilibrium social cost reduction at GPs (due to larger $\tilde{\beta}$ and lower $\tilde{\alpha} = r(\tilde{\beta})$). In this case, the lower fraction of L patients in the patient base means that there are also fewer L patients who can be virtual over-triaged when $\tilde{\beta}$ gets larger, while H patients will visit

the ED anyway. If ED services are sufficiently costly, this leads to a significantly lower marginal equilibrium social cost increase at the ED when $\tilde{\beta}$ gets larger. To restore the first-order conditions (FOCs) characterized by Proposition 4, the virtual triage provider will thus need to increase $\tilde{\beta}^*$ and send more patients to the ED directly, resulting in a higher marginal equilibrium social cost increase at the ED due to larger $\tilde{\beta}$. Meanwhile, we notice that the threshold a_E^h can be negative: if $r'(\tilde{\beta}^*)$ is small, the FOCs can be restored by the marginal equilibrium social cost reduction at GPs, with a higher $\tilde{\beta}^*$ and a large reduction in $\tilde{\alpha}^* = r(\tilde{\beta}^*)$, even when the ED service cost is small.

Triage capability. Meanwhile, Proposition 6 characterizes how the optimal virtual triage accuracy changes in pure strategy equilibrium regions as triage capability improves.

PROPOSITION 6. *Let $r_1(\tilde{\beta})$ and $r_2(\tilde{\beta})$ denote the IROC curves of two virtual triage tools such that $r_2(\tilde{\beta}) < r_1(\tilde{\beta})$, $\forall\ \tilde{\beta} \in (0,1)$. Suppose both $(r_1(\tilde{\beta}_1^*), \tilde{\beta}_1^*)$ and $(r_2(\tilde{\beta}_2^*), \tilde{\beta}_2^*)$ result in the same pure-strategy equilibrium which is in the interior of $R_{p,\sim}$. Define $\tilde{\beta}^c = r_2^{-1}(r_1(\tilde{\beta}_1^*))$. We have $C_s(\tilde{\mathbf{f}}^{\mathbf{e}}(r_1(\tilde{\beta}_1^*), \tilde{\beta}_1^*)) > C_s(\tilde{\mathbf{f}}^{\mathbf{e}}(r_2(\tilde{\beta}_2^*), \tilde{\beta}_2^*))$. Moreover, $\exists\ r_2^d < 0$ s.t.*

(i) *If $r_1'(\tilde{\beta}_1^*) > r_2'(\tilde{\beta}_1^*)$, we have $\tilde{\beta}_1^* < \tilde{\beta}_2^*$ and $r_1(\tilde{\beta}_1^*) > r_2(\tilde{\beta}_2^*)$.*

(ii) *If $r_1'(\tilde{\beta}_1^*) < r_2'(\tilde{\beta}_1^*)$ and $r_2'(\tilde{\beta}^c) > r_2^d$, we have $\tilde{\beta}_1^* > \tilde{\beta}_2^*$ and $r_1(\tilde{\beta}_1^*) < r_2(\tilde{\beta}_2^*)$.*

(iii) *If $r_1'(\tilde{\beta}_1^*) < r_2'(\tilde{\beta}_1^*)$ and $r_2'(\tilde{\beta}^c) < r_2^d$, we have $\tilde{\beta}_1^* > \tilde{\beta}_2^*$ and $r_1(\tilde{\beta}_1^*) > r_2(\tilde{\beta}_2^*)$.*

Importantly, we show that optimal virtual triage accuracy for minimizing equilibrium social cost may not be monotone in triage capability, despite the fact that both $r(\tilde{\beta}^*)$ and $\tilde{\beta}^*$ can be jointly reduced as the triage capability improves. In particular, either $r(\tilde{\beta}^*)$ or $\tilde{\beta}^*$ could increase as the triage capability improves. Note that this is the case in pure strategy equilibrium regions, and the non-monotonicity of optimal virtual triage accuracy here is not caused by the potential non-monotonicity of equilibrium social cost in virtual triage accuracy under mixed strategy equilibrium regions (Lemma 4).

To understand the intuition behind Proposition 6, consider two IROC curves where $r_2(\tilde{\beta})$ has a higher triage capability than $r_1(\tilde{\beta})$, and suppose the associated optimal virtual triage accuracies for minimizing social cost, denoted by $(r_2(\tilde{\beta}_2^*), \tilde{\beta}_2^*)$ and $(r_1(\tilde{\beta}_1^*), \tilde{\beta}_1^*)$, respectively, lead to the same pure strategy equilibrium regions. The conditions under which optimal virtual triage accuracy may or may not decrease monotonically depend on the first order derivatives of $r_2(\tilde{\beta})$ evaluated at $\tilde{\beta} = \tilde{\beta}_1^*$ and $\tilde{\alpha} = r_1(\tilde{\beta}_1^*)$, or equivalently, at $\tilde{\beta} = \tilde{\beta}_1^*$ and $\tilde{\beta} = r_2^{-1}(r_1(\tilde{\beta}_1^*))$. Then Proposition 6 shows that:

- If the first order derivative of $r_2(\tilde{\beta})$ w.r.t. $\tilde{\beta}$ at $\tilde{\beta} = \tilde{\beta}_1^*$ is very small, a marginal increase in $\tilde{\beta}$ can reduce $\tilde{\alpha} = r_2(\tilde{\beta})$ significantly. That is, by slightly worsening the ED overcrowding problem, we can significantly improve the treatment delay problem at GPs. As a result, we have the optimal accuracy along $r_2(\tilde{\beta})$ with larger virtual over-triage probability and smaller virtual under-triage probability than $r_1(\tilde{\beta})$. This typically happens with small $\tilde{\beta}_1^*$ and large $\tilde{\alpha}_1^*$, where the first order derivative is small due to the decreasing and convex nature of the IROC curve.

- If the first order derivative of $r_2(\tilde{\beta})$ w.r.t. $\tilde{\beta}$ at $\tilde{\beta} = r_2^{-1}(r_1(\tilde{\beta}_1^*))$ is very large, a marginal increase in $\tilde{\alpha}$ can reduce $\tilde{\beta} = r_2^{-1}(\tilde{\alpha})$ significantly. In other words, by slightly worsening the treatment delay problem at GPs, we can significantly improve the ED overcrowding problem. As a result, we have the optimal accuracy along $r_2(\tilde{\beta})$ with smaller virtual over-triage probability and larger virtual under-triage probability than $r_1(\tilde{\beta})$. This typically happens with large $\tilde{\beta}_1^*$ and small $\tilde{\alpha}_1^*$, where the first order derivative is large.
- Lastly, we could have $\tilde{\beta}_2^*$ lying between $\tilde{\beta}_1^*$ and $r_2^{-1}(r_1(\tilde{\beta}_1^*))$. In this case, the optimal accuracy along $r_2(\tilde{\beta})$ will have both smaller virtual over-triage probability and smaller virtual under-triage probability than $r_1(\tilde{\beta})$. In other words, at its optimal accuracy, the higher virtual triage capability further alleviates both the treatment delay problem at GPs and the ED overcrowding problem. This typically happens with small $\tilde{\beta}_1^*$ and small $\tilde{\alpha}_1^*$.

Note that when the triage capability improves along with higher $\tilde{\beta}^*$, there are more ED arrivals in order to minimize the equilibrium social cost. Hence, while the equilibrium social cost is reduced, the ED is more congested and patients who visit the ED, either H patients or L patients who visit the ED directly, are worse off. Similarly, when the triage capability improves along with higher $\tilde{\alpha}^*$, more H patients will visit a GP first in order to minimize the equilibrium social cost. Hence, while the equilibrium social cost is reduced, there are more H patients who are worse off and receive delayed treatment by unnecessarily visiting a GP first. As a result, we caution that while an improvement in virtual triage capability under optimal accuracy does reduce the overall equilibrium social cost, it can make certain subsets of patients worse off with higher patient costs.

6.2. Impact of Custom Virtual Triage on Triage Safety

We now characterize the optimal virtual triage accuracy for maximizing triage safety, or equivalently, minimizing treatment delay, $(\tilde{\alpha}^\star, \tilde{\beta}^\star)$, subject to the constraint that $\tilde{\alpha} = r(\tilde{\beta})$, $\tilde{\beta} \in [0,1]$ under existing cost parameters and patient co-payments.

PROPOSITION 7. *For a given virtual triage tool, let $(r(\tilde{\beta}^\star), \tilde{\beta}^\star)$ denote the optimal virtual triage accuracy that minimizes equilibrium treatment delay subject to the IROC curve $r(\tilde{\beta})$. $\tilde{\mathbf{f}}^{\mathbf{e}}(r(\tilde{\beta}^\star), \tilde{\beta}^\star)$ is not in the interior of $R_{p,\sim}$.*

Proposition 7 shows that optimizing the virtual triage accuracy to minimize equilibrium treatment delay cannot result in a pure strategy equilibrium where all patient types strictly prefer their choices of care locations. This stands in contrast to the case aiming to minimize equilibrium social cost, in which optimal accuracy more often results in a pure strategy equilibrium, as detailed in Section 7.3. Consequently, the common approach in practice of prioritizing patient safety may not align with efforts to minimize the equilibrium social cost. Indeed, these two goals might necessitate distinct optimal virtual triage accuracy settings, as the optimal accuracies are likely situated

in different equilibrium regions. This observation underscores a potential misalignment in setting optimal virtual triage accuracy across different scoring functions. Thus, it becomes essential to clarify the main objective during the adoption of custom virtual triage.

6.3. Practical Implications

Our analysis highlights the problem of adopting a universal operating discrimination threshold. Such an approach fails to consider the heterogeneity inherent in the operational efficiency of different acute care systems, differences across patient conditions, and differences across performance metrics. Our results instead indicate that for healthcare providers to fully harness the operational advantages of virtual triage, they should steer clear of a one-size-fits-all discrimination threshold, opting instead to adjust the threshold based on specific contextual considerations. Yet, this tailored approach is seldom seen in practice. For instance, Babylon Health's virtual triage chatbot (elaborated further in Section 7) employs a fixed discrimination threshold, consistently applied to every user of their platform.

Furthermore, we demonstrate that the optimal virtual triage accuracy, constrained by a particular IROC curve, both depends upon and is non-monotonic in its triage capability. The optimal virtual under-triage or over-triage probability might increase as the triage capability improves. Therefore, if a virtual triage system improves triage efficiency by leveraging more expansive training data and advanced classification algorithms, providers should refrain from hastily rolling out a refined, more precise version of the software. Instead, the optimal operating discrimination threshold must undergo careful reconsideration, considering the modified IROC curve.

7. Numerical Analysis

In this section, we perform extensive numerical analysis to evaluate the choice of optimal virtual triage accuracy, calibrating both the IROC curve and acute care costs. This analysis has two main aims. First, we would like to understand the sensitivity of our conclusions to different parameter values, as our characterizations of optimal virtual triage accuracy for minimizing equilibrium social cost in Section 6.1 only hold for pure strategy equilibrium regions. Second, we want to understand the conditions under which the optimal virtual triage accuracies under two performance metrics are more aligned or misaligned. In particular, we do so within the concrete context of the UK acute care system and using Babylon Health's (abbreviated henceforth as Babylon's) virtual triage tool.

7.1. Characterization of the IROC Curve: Babylon's Virtual Triage

Baker et al. (2020) present the precision-recall curve for Babylon's virtual triage chatbot. From their data, the actual IROC curve for the tool cannot be recovered. However, the authors report the

average recall (80.0%), average precision (44.4%), and safety (97.0%) of Babylon's virtual triage recommendations for the operating discrimination threshold $\bar{s}$ that is used in practice. Using these data, we can then calculate the virtual under-triage probability $\tilde{\alpha} = 0.20$ and virtual over-triage probability $\tilde{\beta} = 0.18$ corresponding to this discrimination threshold.

To proxy the IROC curve of Babylon's virtual triage, $\tilde{\alpha} = r_B(\tilde{\beta})$, we assume it takes the implicit functional form $(1-\tilde{\alpha})(1-\tilde{\beta})2^{-k} = \tilde{\alpha}\tilde{\beta}$, $k \in [0, \infty)$. This functional form leads to a balanced IROC curve, which is consistent with recent evidence on the triage capability of virtual triage tools (Thomas et al. 2021, Chang et al. 2022). For this functional form, it is easy to verify that $\tilde{\alpha}$ is a decreasing and convex function of $\tilde{\beta}$, with $r_B(0) = 1$ and $r_B(1) = 0$, while the parameter k effectively captures the triage capability of the virtual triage tool. Higher virtual triage capability is reflected by a larger k, as shown in Figure 3 (a). Using this, the IROC curve for Babylon's virtual triage tool can be proxied by setting $k = 4.2$.

7.2. Parameters Relating to Acute Care Provision

To demonstrate the robustness of our results, in the numerical analysis we use a wide range of plausible parameter values relating to acute care provision. Parameter values are drawn from the UK context, where Babylon's virtual triage tool was first deployed. We list below the key parameters and their values, and we provide justification in EC.2 of the e-companion.

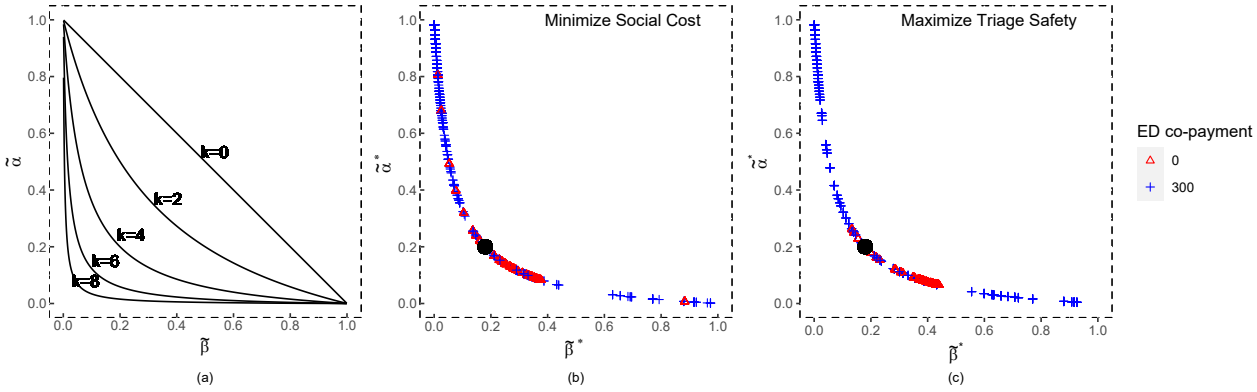
- Expected service cost $a_G = 40$GBP/patient, $a_E \in \{100, 200, 300, 400\}$GBP/patient
- Arrival rate of strategic patients $\lambda \in \{200, 300, 400, 500\}$patients/hr
- Fraction of ED-type patients $h \in \{0.1, 0.2, 0.3, 0.4\}$
- Expected waiting time $Q_G = 4$hrs, $Q_E(\lambda_E) = c\lambda_E^2$ with $Q_E(\lambda) = 4$hrs
- Disutility of waiting per unit time $w_G = 3$GBP/hr, $w_E = 15$GBP/hr
- Patient co-payment $p_G = 0$GBP/patient, $p_E \in \{0, 100, 200, 300\}$GBP/patient
- Patient self-triage accuracy $(\hat{\alpha}, \hat{\beta}) \in \{(0.1, 0.2), (0.1, 0.4), (0.1, 0.6), (0.1, 0.8), (0.3, 0.2), (0.3, 0.4), (0.3, 0.6), (0.5, 0.2), (0.5, 0.4), (0.7, 0.2)\}$
- Triage capability of virtual triage $k \in \{0, 0.2, 0.4, ..., 8\}$

7.3. Numerical Characterization of Optimal Virtual Triage Accuracy

We numerically solve for the optimal virtual triage accuracy subjective to a given IROC curve for the combinations of the aforementioned parameter values, and for both objective functions (i.e., minimizing equilibrium social cost and maximizing equilibrium triage safety), which leads to 209,920 instances in total.

We first examine the sensitivity of the results in Section 6.1 to different parameter values. Of the 104,960 instances when minimizing equilibrium social cost, 1,077 have an optimal virtual triage

Figure 3 (a) IROC curves associated with virtual triage algorithms with different triage capabilities. The corresponding optimal virtual triage accuracy for Babylon's virtual triage tool under different parameter combinations with objective function of (b) minimizing social cost and (c) maximizing triage safety, where the black dot highlights their commonly used operating accuracy.



accuracy that leads to mixed strategy equilibrium patient flows. Further analysis reveals that 560 of these 1,077 instances occur when $k = 0$, i.e., when virtual triage is uninformative and has no impact. Of the remaining 517 instances that arise when $k > 0$, 501 (97%) occur when $p_E = 0$, i.e., when ED usage is free. This is because patients are much more likely to visit the ED directly when $p_E = 0$. In this case, patients' disutility of waiting at the ED, which depends on other patients' behavior, plays a more important role in the total patient cost of an ED visit, thus making the existence of mixed strategy equilibria more likely.[18] Meanwhile, of the 2,560 instances with $k = 4.2$ (i.e., corresponding to Babylon's virtual triage tool), no instances have an optimal virtual triage accuracy that leads to mixed strategy equilibrium patient flows. These numerical results thus indicate that under optimal virtual triage accuracy for minimizing equilibrium social cost, the equilibrium patient flow is rarely in mixed strategy form.

Figure 3 (b) shows the optimal virtual triage accuracy minimizing equilibrium social cost for Babylon's tool under the parameter combinations specified in Section 7.2, with ED co-payment $p_E \in \{0, 300\}$. Figure 3 (c) shows the same results when maximizing equilibrium triage safety. The black dot in Figures 3 (b) and (c) corresponds to the operating accuracy that is used in practice for Babylon's virtual triage chatbot. First, we can see that the optimal accuracy spreads along the IROC curve as the underlying system parameters change and that the actual operating accuracy used by Babylon can be far from the optimal. This is because despite being informative, Babylon's virtual triage tool has limited predictive power. Hence, a balanced accuracy may not be able to alleviate the ED overcrowding problem and the treatment delay problem simultaneously, making the choice of balanced accuracy suboptimal.

Moreover, Figures 3 (b) and (c) show the ED co-payment's effect on optimal virtual triage accuracy. In particular, when ED access is free, the optimal virtual triage accuracy tends to be more

robust to both various parameter combinations and different objective functions: it tends to have an optimal accuracy with a relatively small virtual under-triage probability and a relatively large virtual over-triage probability. This combination makes GP recommendations more informative; hence, the adoption of virtual triage under optimal accuracy leads to lower equilibrium social cost by reducing visits to the ED. This makes sense as with free ED access, many GP-type patients tend to visit the ED in the absence of virtual triage. This leads to ED overcrowding, high ED costs, and therefore high system inefficiency, which the virtual triage provider can help alleviate with a small virtual under-triage probability. However, the optimal virtual under-triage probability cannot be too small due to the intrinsic informativeness-volume trade-off as discussed in Section 4.2.

Conversely, when there is a fee associated with ED access, the optimal accuracy tends to shift more noticeably as other system parameters change. Therefore, in scenarios with high ED co-payments, identifying specific conditions for virtual triage implementation might hold more merit than universally applying the technology across the board. This strategy could be advantageous both for healthcare systems and virtual triage providers. Moreover, generally a larger virtual under-triage probability paired with a smaller virtual over-triage probability is preferred to the opposite. The rationale behind this is that ED resources typically come at a premium. Hence, over-triaging patients virtually is more costly than under-triaging them, making a small virtual over-triage probability the more desirable outcome in most instances.

7.4. Misalignment of Optimal Virtual Triage Accuracies under the Two Objective Functions

We also offer a detailed characterization of the scenarios where the two objective of minimizing social cost and maximizing triage safety either align more closely or deviate significantly. To achieve this, we conduct a linear regression, comparing the differences in optimal virtual under-triage probability between the two objective functions across various parameter combinations. This analysis reveals that discrepancies in the optimal virtual triage accuracy are more pronounced ($p < 0.001$) under scenarios of low patient volume, high ED co-payment, a low fraction of ED-type patients, high patient self-triage accuracy, or low virtual triage capability. Given these findings, acute care systems confronted with these scenarios should deliberate more carefully on their primary objective prior to implementing custom virtual triage.

8. Conclusions and Practical Implications

This paper has presented a comprehensive analysis of the operational impact and policy implications of virtual triage adoption in acute care systems. Our findings highlight the potential of virtual triage tools to enhance system performance through more accurate triage recommendations, provided their implementation and accuracy calibration are properly managed.

From a technical perspective, we characterize patients' equilibrium care-seeking behavior and demonstrate the uniqueness of equilibrium outcomes both with and without virtual triage. Importantly, we reveal an underlying informativeness-volume trade-off that could lead to unintended changes in patient behavior when virtual triage excessively (but not unconditionally) recommends one level of care. We also establish that the introduction of an off-the-shelf virtual triage solution may increase equilibrium social cost and reduce equilibrium triage safety compared to a setting without virtual triage, especially when accuracy is moderately high.

The potential adverse outcomes associated with an off-the-shelf virtual triage solution motivates us to explore policy actions to optimize the operational advantages of custom virtual triage. We derive analytical conditions for optimal virtual triage accuracy minimizing equilibrium social cost or maximizing triage safety subject to a given IROC curve. We find potential non-monotonicity, whereby the optimal accuracy may entail a higher probability of virtual under-triage or over-triage even as overall capability improves. We also characterize how optimal accuracy depends on and adapts to changes in patient composition and improvements in the virtual triage tool's capability.

For practitioners, our findings underline the importance of customizing virtual triage tools to specific acute care contexts, rather than utilizing a universal discrimination threshold. The optimal calibration of accuracy relies heavily on associated care costs, patient demographics, the tool's capability, and the primary performance objective. Furthermore, we underscore the risk that upgrading tools for enhanced accuracy without comprehensive reassessment can potentially undo benefits. As such, regulators should mandate a recertification process whenever a new version of the tool is introduced. We also recommend that technology providers prioritize collaboration with healthcare partners to identify contexts that will yield more robust improvements in system performance. Overall, by providing guidance on effective integration, our research can assist practitioners in leveraging virtual triage tools to deliver improved and more equitable acute care.

While our model aims to capture the key trade-offs involved in virtual triage adoption, there are several avenues that provide opportunities for future work. First, we assume a binary classification of patients into GP-type or ED-type. Extending the model to account for care pathways with more nuanced triage levels could offer additional insights. Second, we assume conditional independence between self-triage accuracy and virtual triage accuracy. While we do not expect changes in structural insights, future work could look into this dimension. Moreover, as a first step to understand the operational implications of virtual triage adoption, we do not consider the potential agency problems, such as the potential malicious representation behavior of the virtual triage providers given the trade-off between increasing immediate adoption versus long-term sustainability of the business. We leave this problem to future research.

In closing, this paper offers a comprehensive analytical exploration of the nuanced effects of virtual triage adoption on patient care-seeking behavior and acute care system performance. We provide important technical insights and practical guidance to assist healthcare practitioners in effectively integrating these emerging tools, unlocking their potential to enhance triage accuracy, improve resource allocation, and deliver higher quality care. More broadly, our findings emphasize the significance of incorporating decentralized behavior and implementation details when assessing new technologies. By aligning incentives and optimizing solutions to context, unintended consequences can be avoided and intended benefits can be fully realized. This study demonstrates the vital role of operations management research in responsibly steering the adoption of transformative innovations across application domains.

Endnotes

1. While ED triage (i.e., the prioritization of treatment for patients requiring urgent care) has been used as a tool to counteract this issue (Iserson and Moskop 2007), it still results in the wasteful use of costly emergency care resources when patients could have been adequately treated in a primary care setting.
2. The same analysis could be applicable to an acute care system with urgent care facilities instead of EDs, e.g., in countries like the US or UK, where the number of urgent care facilities has grown.
3. It should be noted that the focus on strategic patients does not lead to a loss of generality. Patients affect other patients through the changes in expected waiting time and therefore disutility of waiting. Given our generic formulation without specific assumption on queuing discipline in Section 3.3, our model allows for an arbitrary arrival rate of non-strategic patients and patients seeking care at night or on weekends. Since such patients do not change their care-seeking behavior with the adoption of virtual triage and their arrival rate is a constant system primitive, their impact has been implicitly captured by the expected waiting time when the arrival rate of strategic patients is zero. Hence, explicitly including these patients in the model formulation does not make any difference at the margin of the expected waiting time and therefore does not change any result.
4. The homogeneous self-triage accuracy assumption is without loss of generality, as shown in EC.5.1 of the e-companion.
5. We can show that the structural results remain if bias is small.
6. Note that there is a one-to-one mapping between $\bar{s}$ and the pair of $\tilde{\alpha}$ and $\tilde{\beta}$. To simplify the exposition, we may omit the dependence of $\tilde{\alpha}$ and $\tilde{\beta}$ on $\bar{s}$ for the rest of the paper.
7. We note that Webb and Mills (2019) capture a similar trade-off in a centralized setting; in this paper, we explore the effect of such an accuracy trade-off on patients' care-seeking behavior in a decentralized setting.
8. We assume patients are informed of the virtual triage accuracy. Instead, patients could also learn about virtual triage accuracy from repeated interactions with the technology, which leads to the same equilibrium outcomes. We also discuss the relaxation of full information assumption concerning patients in EC.5.4 of the e-companion. We show that the structural results persist as long as patients possess a general sense of system parameters.

9. We examine the scenario where patients' disutility of waiting per unit time also depends on their belief of being ED-type in EC.5.2 of the e-companion. All the results continue to hold.

10. This follows from the observation that strategic patients seeking acute care only account for a small fraction of all patients accessing GP services. In particular, GPs manage various types of illness, including the delivery of chronic disease care, treatment of acute non-life-threatening diseases, early detection and referral of patients with urgent serious diseases, health education, and immunization. Thus, to ensure that acute care patients can receive timely, prioritized care despite this varied caseload, GPs typically reserve capacity each day for acute care appointments (Gupta and Wang 2008). They also have the ability to reallocate resources between chronic and acute care services, modify the amount of capacity reserved for acute care appointments, and adjust working hours, thereby ensuring that waiting times are relatively stable. Nevertheless, our results continue to hold when the expected waiting time at GPs is increasing in the arrival rate of strategic patients at GPs, so long as the rate of increase is small.

11. Note that λ_E is lower bounded by λ_H, i.e., $\lambda_E \geq \lambda_H$, since all ED-type patients will visit the ED, either directly or after being referred from a GP.

12. For simplicity, we refer to the nonatomic Nash equilibrium patient flow as the equilibrium patient flow.

13. To simplify the exposition, we may omit the dependence of $\hat{\mathbf{f}}^{\mathbf{e}}$ and $\tilde{\mathbf{f}}^{\mathbf{e}}$ on $\hat{\alpha}, \hat{\beta}, \tilde{\alpha}$ and $\tilde{\beta}$.

14. We omit the social cost of non-strategic patients and out-of-hours care seekers in $C_s(\cdot)$, as their behaviors and social cost aren't influenced by the adoption of virtual triage or strategic patients' actions. Further, research indicates that waiting times for priority patients at EDs aren't significantly affected by less critical patients (Schull et al. 2007, Zane 2007), as they receive priority treatment (ESI level 1 or 2) over patients with less severe conditions (ESI level 3, 4, or 5) (Gilboy et al. 2020). Additionally, chronic patients visiting GPs usually have pre-scheduled appointments and typically aren't allocated to same-day capacity reserved for acute care. Thus, the social cost of non-strategic patients remains constant, regardless of virtual triage adoption or strategic patients' behaviors.

15. Parameter values: $\lambda_L = 30, \lambda_H = 7, w_G = 5, w_E = 10, p_G = 40, p_E = 100, Q_G = 10$. $\hat{\mathbf{f}}^{\mathbf{e}} = (0,0) : Q_E(\lambda_E) = 0.05\lambda_E^2, \hat{\alpha} = 0.1, \hat{\beta} = 0.8$; $\hat{\mathbf{f}}^{\mathbf{e}} = (0,m) : Q_E(\lambda_E) = 0.05\lambda_E^2, \hat{\alpha} = 0.3, \hat{\beta} = 0.3$; $\hat{\mathbf{f}}^{\mathbf{e}} = (0,1) : Q_E(\lambda_E) = 0.05\lambda_E^2, \hat{\alpha} = 0.4, \hat{\beta} = 0.1$; $\hat{\mathbf{f}}^{\mathbf{e}} = (m,1) : Q_E(\lambda_E) = 0.0005\lambda_E^2, \hat{\alpha} = 0.5, \hat{\beta} = 0.1$; $\hat{\mathbf{f}}^{\mathbf{e}} = (1,1) : Q_E(\lambda_E) = 0.0005\lambda_E^2, \hat{\alpha} = 0.8, \hat{\beta} = 0.1$.

16. Note that in practice, different virtual triage providers may follow different predetermined scoring functions to update their accuracy with a better IROC curve. Since we do not assume any specific IROC curve or scoring function here for the sake of the generality of our analysis, we study the updating of off-the-shelf virtual triage accuracy in an implicit manner: in particular, we analyze the separate effect of lower virtual under-triage (over-triage) probability while virtual over-triage (under-triage) probability remains unchanged. With the separate effect along with the functional form of IROC curve and the scoring function, we can then recover the net effect of higher off-the-shelf virtual triage accuracy with a better IROC curve.

17. By "interior", we mean all patient types strictly prefer their choices of care locations for a given pure strategy equilibrium; in other words, it will lead to strictly higher patient costs for any patient that deviates from the equilibrium.

18. However, we note that free access to the ED does not guarantee that mixed strategy equilibria will occur. Out of the 26,240 instances where $p_E = 0$, mixed strategy equilibria only arise in 1,061 (4%) instances.

References

Agrawal A, Gans J, Goldfarb A. 2022. Power and prediction: The disruptive economics of artificial intelligence. Boston: Harvard Business Review Press.

Alizamir S, de Véricourt F, Sun P. 2013. Diagnostic accuracy under congestion. *Management Sci.* 59(1):157–171.

Babic B, Gerke S, Evgeniou T, Cohen IG. 2019. Algorithms on regulatory lockdown in medicine. *Science* 366(6470):1202–1204.

Baker A, Perov Y, Middleton K, et al. 2020. A comparison of artificial intelligence and human doctors for the purpose of triage and diagnosis. *Frontiers in Artificial Intelligence* 3(543405): 1–9.

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

Bavafa H, Savin S, Terwiesch C. 2019. Redesigning primary care delivery: customized office revisit intervals and e-visits. *Working Paper*.

Blue Shield of California. 2020. COVID-19: virtual triage can help patients, ease burden on hospitals. (April 6). https://www.bcbs.com/the-health-of-america/articles/covid-19-virtual-triage-can-help-patients-ease-burden-hospitals

Boyle S. 2011. United Kingdom (England): health system review.

Çakıcı ÖE, Mills AF. 2020. On the role of teletriage in healthcare demand management. *Manufacturing Service Oper. Management*, forthcoming.

Chambers D, Cantrell AJ, Johnson M, Preston L, Baxter SK, Booth A, Turner J. 2019. Digital and online symptom checkers and health assessment/triage services for urgent health problems: systematic review. *BMJ Open* 9:e027743.

Chang YH, Shih HM, Wu JE. 2022. Machine learning–based triage to identify low-severity patients with a short discharge length of stay in emergency department. *BMC Emergency Medicine* 22(88):1–10.

Coons KC, DuMoulin JP. 2000. Telephone triage. Technical report, American College of Physicians–American Society of Internal Medicine, Washington, DC.

Corl K. 2019. Hospitals' new emergency department triage systems boost profits but compromise care. *STAT* (September 5). https://www.statnews.com/2019/09/05/triage-system-boost-profits-compromises-care/.

Cui S, Veeraraghavan S. 2016. Blind queues: the impact of consumer beliefs on revenues and congestion. *Management Sci.* 62(12):3656–3672.

Dai T, Singh S. 2020. Conspicuous by its absence: Diagnostic expert testing under uncertainty. *Marketing Science.* 39(3):540–563.

Dai T, Singh S. 2021. Artificial intelligence on call: The physician's decision of whether to use AI in clinical practice. *Working Paper*.

Dalton J. 2020. Coronavirus: callers to NHS 111 phone line wait hours and get cut off without being able to speak to nurse. *INDEPENDENT* (March 13). https://www.independent.co.uk/news/uk/home-news/coronavirus-uk-symptoms-nhs-111-phone-line-nurse-a9400351.html.

Debo L, Parlour C, Rajan U. 2012. Signaling quality via queues. *Management Sci.* 58(5):876–891.

Eatock J, Cooke M, Young TP. 2017. Performing or not performing: What's in a target? *Future Healthcare Journal* 4(3):167–172.

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

Freeman M, Robinson S, Scholtes S. 2020. Gatekeeping, fast and slow: an empirical study of referral errors in the emergency department. *Management Sci.*, forthcoming.

Gilboy N, Tanabe P, Travers D, Rosenau AM. 2020. Emergency severity index, version 4: implementation handbook. Emergency Nurses Association, Schaumburg, IL.

Gneiting T. 2011. Making and evaluating point forecasts. *Journal of the American Statistical Association* 106(494):746–762.

Grand View Research. 2019. Acute hospital care market growth & trends. https://www.grandviewresearch.com/press-release/global-acute-hospital-care-market

Gupta D, Wang L. 2008. Revenue management for a primary-care clinic in the presence of patient choice. *Oper. Res.* 56(3):576–592.

Hao K. 2020. Doctors are using AI to triage COIVD-19 patients. The tools may be here to stay. *MIT Technology Review* (April 23). https://www.technologyreview.com/2020/04/23/1000410/ai-triage-covid-19-patients-health-care/?truid=cb0787e5baf82e6f6cc19ac58536e5b4&utm_source=the_download&utm_medium=email&utm_campaign=the_download.unpaid.engagement&utm_content=04-24-2020.

Hasija S, Pinker E, Shumsky R. 2005. Staffing and routing in a two-tier call centre. *Internat. J. Oper. Res.* 1(1/2):8–29.

Heaven WD. 2020. Google's medical AI was super accurate in a lab. Real life was a different story. *MIT Technology Review* (April 27). https://www.technologyreview.com/2020/04/27/1000658/google-medical-ai-accurate-lab-real-life-clinic-covid-diabetes-retina-disease/?truid=cb0787e5baf82e6f6cc19ac58536e5b4&utm_source=the_algorithm&utm_medium=email&utm_campaign=the_algorithm.unpaid.engagement&utm_content=05-01-2020

Hirshon JM, Risko N, Calvello EJ, de Ramirez SS, Narayan M, Theodosis C, O'Neill J. 2013. Health systems and services: the role of acute care. *Bulletin of the World Health Organization* 91(5):386–388.

Hu M, Li Y, Wang J. 2018. Efficient ignorance: information heterogeneity in a queue. *Management Sci.* 64(6):2650–2671.

Huang J, Carmeli B, Mandelbaum A. 2012. Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Oper. Res.* 63(4):892–908.

Iserson KV, Moskop JC. 2007. Triage in medicine, part i: concept, history, and types. *Annals of Emergency Medicine* 49(3):275–281.

Kamali MF, Tezcan T, Yildiz O. 2019. When to use provider triage in emergency departments. *Management Sci.* 65(3):1003–1019.

Kocher KE, Ayanian JZ. 2016. A fractured system: where do you go when you suddenly need health care? *The Conversation* (October 31). https://theconversation.com/a-fractured-system-where-do-you-go-when-you-suddenly-need-health-care-66662.

Lee H, Pinker E, Shumsky R. 2012. Outsourcing a two-level service process. *Management Sci.* 58(8):1569–1584.

Lega F, Mengoni A. 2008. Why non-urgent patients choose emergency over primary care services? Empirical evidence and managerial implications. *Health Policy* 88(2):326–338.

Levi R, Magnanti T, Shaposhnik Y. 2019. Scheduling with testing. *Management Sci.* 65(2):776–793.

Liu Y, Wang X, Gilbert S, Lai G. 2018. Pricing, quality and competition at on-demand healthcare service platforms. *Working Paper.*

Lovett L. 2018. AI triage chatbots trekking toward a standard of care despite criticism. *Mobile Health News* (November 2). https://www.mobihealthnews.com/content/ai-triage-chatbots-trekking-toward-standard-care-despite-criticism.

Meyer A, Giardina TD, Spitzmueller C, Shahid U, Scott T, Singh H. 2020. Patient perspectives on the usefulness of an artificial intelligence–assisted symptom checker: cross-sectional survey study. *J Med Internet Res* 22(1):e14679.

Papanastasiou Y, Bakshi N, Savva N. 2015. Scarcity strategies under quasi-Bayesian social learning. *Working Paper.*

Papanastasiou Y, Bimpikis K, Savva N. 2018. Crowdsourcing exploration. *Management Sci.* 64(4):1727–1746.

Rajan B, Tezcan T, Seidmann A. 2019. Service systems with heterogeneous customers: investigating the effect of telemedicine on chronic care. *Management Sci.* 65(3):1236–1267.

Saghafian S, Hopp WJ, Van Oyen MP, Desmond JS, Kronick SL. 2012. Patient streaming as a mechanism for improving responsiveness in emergency departments. *Oper. Res.* 60(5):1080–1097.

Savin S, Xu Y, Zhu L. 2019. Delivering multi-specialty care via online telemedicine platforms. *Working Paper.*

Schmeidler D. 1973. Equilibrium points of nonatomic games. *J. Statist. Phys.* 7(4):295–300.

Schull MJ, Kiss A, Szalai J. 2007. The effect of low-complexity patients on emergency department waiting times. *Annals of Emergency Medicine* 49(3):257–264.

Semigran HL, Linder JA, Gidengil C, Mehrotra A. 2015. Evaluation of symptom checkers for self diagnosis and triage: audit study. *BMJ* 351:h3480.

Sharma S, Xu Y, Gupta MK, Courcoubetis C. 2019. Non-urgent visits and emergency department congestion: patients' choice and incentive mechanisms. *Working Paper.*

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

Singh S, Gurvich I, Van Mieghem JA. 2020. Feature-based design of priority queues: digital triage in healthcare. *Working Paper.*

Sun Z, Argon NT, Ziya S. 2018. Patient triage and prioritization under austere conditions. *Management Sci.* 64(10):4471–4489.

Thomas B, Goodacre S, Lee E, et al. 2021. Prognostic accuracy of emergency department triage tools for adults with suspected COVID-19: the PRIEST observational cohort study. *Emerg Med J* 2021(38):587–593.

Trivedi S, Littmann J, Kapur P, Betz M, Stempien J. 2017. LO09: assessing the ability of emergency department patients to self-triage by using an electronic questionnaire: a pilot study. *CJEM* 19(S1):S30–S30.

Turbitt E, Freed GL. 2015. Use of a telenursing triage service by Victorian parents attending the emergency department for their child's lower urgency condition. *Emerg Med Australas* 27(6):558–562.

Veeraraghavan S, Debo L. 2009. Joining longer queues: information externalities in queue choice. *Manufacturing Service Oper. Management* 11(4):543–562.

Verzantvoort NCM, Teunis T, Verheij TJM, van der Velden A. 2018. Self-triage for acute primary care via a smartphone application: practical, safe and efficient? *PLOS ONE* 13(6):e0199284.

Webb EM, Mills AF. 2019. Incentive-compatible prehospital triage in emergency medical services. *Prod. Oper. Manag.* 28(9):2221–2241.

Winn AN, Somai M, Fergestrom N, Crotty BH. 2019. Association of use of online symptom checkers with patients' plans for seeking care. *JAMA Network Open* 2(12):e1918561.

Zane RD. 2007. Are low-acuity patients clogging up the ED? *NEJM Journal Watch* Reviewing Schull et al. 2007 Ann Emerg Med 2007 Mar.

Zayas-Cabán G, Xie J, Green LV, Lewis ME. 2014. Optimal control of an emergency room triage and treatment process. *Working Paper*.

Zorc S, Chick SE, Hasija S. 2023. Outcomes-based reimbursement policies for chronic care pathways. *Working Paper*.

Online Technical Appendix

In this e-companion we provide detailed discussion of Proposition 1, justifications for the choices of parameter values relating to acute care provision for the numerical analysis, and technical results that are required for our analysis, detailed proofs of all the mathematical results in the paper, as well as a set of model extensions and robustness.

EC.1. Detailed Discussion of the Relative Positions of Equilibrium Regions in Proposition 1

Proposition 1 (iii).

When $\hat{\mathbf{f}}^{\mathbf{e}} = (0, 1)$, patient self under-triage and self over-triage probabilities, $\hat{\alpha}$ and $\hat{\beta}$, do not differ by much, such that patients follow their self-triage decisions in the absence of virtual triage. In this case, when virtual triage recommendations confirm self-triage decisions, patients always follow virtual triage recommendations in equilibrium regardless of their accuracy, i.e., $f^e_{\hat{L}\tilde{L}} = 0$ and $f^e_{\hat{H}\tilde{H}} = 1$. On the other hand, when the virtual triage recommendation contradicts the self-triage decision, patients may follow their self-triage decision, follow the virtual triage recommendation, or adopt a mixed strategy, depending on the values of $\tilde{\alpha}$ and $\tilde{\beta}$.

When the accuracy of virtual triage is low, i.e., both $\tilde{\alpha}$ and $\tilde{\beta}$ are relatively large, patients' posterior beliefs of being ED-type center around their priors. In this case, despite patients' being better informed about their healthcare needs, they still follow their self-triage decisions, resulting in the equilibrium patient flow $\tilde{\mathbf{f}}^{\mathbf{e}} = (0, 0, 1, 1)$. If $\tilde{\alpha}$ remains relatively large but $\tilde{\beta}$ gets smaller, $\hat{L}\tilde{H}$ patients will have a higher posterior belief of being ED-type and start to visit the ED directly with a positive probability, resulting in the equilibrium patient flow $\tilde{\mathbf{f}}^{\mathbf{e}} = (0, m, 1, 1)$. When $\tilde{\beta}$ gets very small, $\hat{L}\tilde{H}$ patients will have a posterior belief close to 1 and therefore all of them will follow the virtual triage recommendations instead of their self-triage decisions, resulting in the equilibrium patient flow $\tilde{\mathbf{f}}^{\mathbf{e}} = (0, 1, 1, 1)$. Meanwhile, if $\tilde{\beta}$ instead remains relatively large but $\tilde{\alpha}$ gets smaller, $\hat{H}\tilde{L}$ patients will have a lower posterior belief of being ED-type and start to visit a GP first with a positive probability, leading to the equilibrium patient flows $\tilde{\mathbf{f}}^{\mathbf{e}} = (0, 0, m, 1)$ and $\tilde{\mathbf{f}}^{\mathbf{e}} = (0, 0, 0, 1)$.

If both $\tilde{\alpha}$ and $\tilde{\beta}$ are close to 0, patients simply follow the virtual triage recommendations in equilibrium regardless of their self-triage decisions. This leads to the equilibrium patient flow $\tilde{\mathbf{f}}^{\mathbf{e}} = (0, 1, 0, 1)$. If $\tilde{\alpha}$ gets larger, $\hat{H}\tilde{L}$ patients are more likely to be virtual under-triaged, and therefore they will start to visit the ED directly with a positive probability. This results in the equilibrium patient flow $\tilde{\mathbf{f}}^{\mathbf{e}} = (0, 1, m, 1)$. If instead $\tilde{\beta}$ gets larger, $\hat{L}\tilde{H}$ patients are more likely to be virtual over-triaged and therefore they will start to visit a GP directly with a positive probability. This results in the equilibrium patient flow $\tilde{\mathbf{f}}^{\mathbf{e}} = (0, m, 0, 1)$.

Proposition 1 (ii) and (iv).

When $\hat{\mathbf{f}}^{\mathbf{e}} = (0, m)$, patient self-triage accuracy is such that $\hat{H}$ is less accurate than $\hat{L}$. Hence, $\hat{H}$ patients may go to a GP first with a positive probability. In this case, $\tilde{L}$ patients' behavior is primarily determined by $\tilde{\alpha}$. If $\tilde{\alpha}$ is large, a GP recommendation from virtual triage carries little information. Hence, $\hat{H}\tilde{L}$ patients continue to adopt a mixed strategy in equilibrium while all $\hat{L}\tilde{L}$ patients visit a GP first, leading to equilibrium patient flow $\tilde{\mathbf{f}}^{\mathbf{e}} \in \{(0,1,m,1),(0,0,m,1)\}$. If $\tilde{\alpha}$ is small, a GP recommendation from virtual triage is highly informative. All $\tilde{L}$ patients will then go to a GP first independent of their self-triage decisions. This leads to equilibrium patient flow $\tilde{\mathbf{f}}^{\mathbf{e}} \in \{(0,1,0,1),(0,m,0,1),(0,0,0,1),(0,0,0,m)\}$. On the other hand, $\tilde{H}$ patients' behavior is primarily determined by $\tilde{\beta}$. When $\tilde{\beta}$ is large, an ED recommendation from virtual triage carries little information, such that even $\hat{H}\tilde{H}$ patients still adopt a mixed strategy in equilibrium, leading to the equilibrium patient flow $\tilde{\mathbf{f}}^{\mathbf{e}} = (0,0,0,m)$. As $\tilde{\beta}$ decreases, an ED recommendation from virtual triage is more informative. Hence, as $\tilde{\beta}$ decreases, $\hat{L}\tilde{H}$ patients will shift from all going to a GP first ($\tilde{\mathbf{f}}^{\mathbf{e}} \in \{(0,0,m,1),(0,0,0,1)\}$), to going to the ED directly with a positive probability ($\tilde{\mathbf{f}}^{\mathbf{e}} = (0,m,0,1)$), to all going to the ED directly ($\tilde{\mathbf{f}}^{\mathbf{e}} \in \{(0,1,m,1),(0,1,0,1)\}$). Similar arguments hold when $\hat{\mathbf{f}}^{\mathbf{e}} = (m,1)$, with the roles of $\tilde{\alpha}$ and $\tilde{\beta}$ being exchanged.

Proposition 1 (i) and (v).

Lastly, when $\hat{\mathbf{f}}^{\mathbf{e}} = (0,0)$, all $\tilde{L}$ patients follow virtual triage recommendations and go to a GP first regardless of virtual triage accuracy: As all patients go to a GP first in the absence of virtual triage, a $\tilde{L}$ recommendation further reduces their beliefs of being ED-type. On the other hand, $\tilde{H}$ patients' care-seeking behavior critically depends on the virtual over-triage probability $\tilde{\beta}$. As $\tilde{\beta}$ decreases, more $\hat{H}\tilde{H}$ patients will start to go to the ED directly ($\tilde{\mathbf{f}}^{\mathbf{e}} \in \{(0,0,0,0),(0,0,0,m),(0,0,0,1)\}$), followed by $\hat{L}\tilde{H}$ patients ($\tilde{\mathbf{f}}^{\mathbf{e}} \in \{(0,m,0,1),(0,1,0,1)\}$). Similar arguments hold when $\hat{\mathbf{f}}^{\mathbf{e}} = (1,1)$ and the equilibrium patient flow in the presence of virtual triage critically depends on the virtual under-triage probability $\tilde{\alpha}$.

EC.2. Parameter Values Relating to Acute Care Provision for Numerical Analysis

The parameter values utilized for numerical analysis are drawn from the UK context, where Babylon Health's virtual triage tool was first deployed.

Cost.

The average cost of a GP appointment in the UK was £39.23 in 2020 (The King's Fund 2022), while the average cost for an ED visit was £193 (NHS 2021). Hence, we use the following parameters in the numerical analysis: $a_G = 40$GBP/patient, $a_E \in \{100, 200, 300, 400\}$GBP/patient.

Patient arrival rate.

The arrival rate of patients seeking acute care varies across different locations and time periods. As a benchmark, average monthly ED arrivals were around 1,400,000 for the EDs of 126 English NHS trusts in 2020, which gives a daily arrival rate of about 370 patients/day per ED (Care Quality Commission 2022). Given that most ED arrivals occur during daytime when GP practices are also available (Eatock et al. 2017), this translates to about 45 patients on average per ED per hour. On the other hand, there are around 7,000 GP practices in England (Bostock 2019) and 308 million appointments, with 44% of these being same-day appointments in 2018 (NHS 2019). This translates to 53 same-day arrivals per GP practice per day, or about 7 patients per GP practice per hour given a GP's limited operating hours (assumed to be 8 hours per day). Assuming an average of 55 ($=7000/126$) GP practices per ED, the average catchment area around an ED will experience an arrival rate of around 430 ($=55\times 7+45$) potentially strategic patients per hour. Note, however, that not all of these are strategic, e.g., approximately 25% of patients arrive by ambulance (O'Keeffe et al. 2018). Thus, to account for the variations in the arrival rate of strategic patients as well as different GP practice/ED ratios, we use $\lambda\in\{200,300,400,500\}$patients/hr for the arrival rate of strategic patients. Meanwhile we assume a fraction of h patients are H, where $h\in\{0.1,0.2,0.3,0.4\}$.

Waiting times.

For the expected waiting time at the ED, we assume a quadratic functional form, $Q_E(\lambda_E)=c\lambda_E^2$, and $Q_E(\lambda)=4$hrs given the four-hour waiting time target in the UK, which is monitored within all types of EDs (The King's Fund 2022). For GPs, as we focus on strategic patients seeking to make same-day appointments, we assume $Q_G=4$hrs in the numerical analysis. We assume $w_E=15$GBP/hr, which is the average full-time hourly wage in the UK (Clark 2021). On the other hand, the disutility of waiting for a GP visit is smaller, as patients could be working or at home while waiting, and we assume $w_E=3$GBP/hr. Moreover, given that medical services are covered free of charge through the NHS in the UK (Boyle 2011), and to examine the role of co-payments, we assume $p_G=0$GBP/patient, and $p_E\in\{0,100,200,300\}$GBP/patient.

Triage accuracy.

Lastly, we include a wide range of patient self-triage accuracy values, $(\hat{\alpha},\hat{\beta})\in\{(0.1,0.2),(0.1,0.4),(0.1,0.6),(0.1,0.8),(0.3,0.2),(0.3,0.4),(0.3,0.6),(0.5,0.2),(0.5,0.4),(0.7,0.2)\}$, and a wide range of virtual triage capabilities, $k\in\{0,0.2,0.4,...,8\}$ in the analysis.

EC.3. Characterization of Equilibrium Patient Flow

In this section, we provide a general characterization of equilibrium accommodating an arbitrary number of patient groups, each defined by their own belief of being ED-type. Suppose that there are n groups of patients seeking acute care. Patients of group i have an arrival rate λ_i, with a belief b_i of being ED-type, $i \in \{1,2,...,n\}$. Without loss of generality, we assume $b_1 < b_2 < ... < b_n$. Suppose the expected GP co-payment per visit is p_G, and the expected ED co-payment per visit is p_E. Let $f_i \in [0,1]$ denote the probability of group i patients visiting the ED directly, with $\mathbf{f} = (f_1, f_2, ..., f_n)$.

We define the following *potential function* (Roughgarden 2007), $\Phi(\mathbf{f})$, for our nonatomic game:

$$\Phi(\mathbf{f}) = \int_0^{\lambda_G(\mathbf{f})} w_G Q_G dx + \int_0^{\lambda_E(\mathbf{f})} w_E Q_E(x) dx + \int_0^{\lambda_G(\mathbf{f})} p_G dx + \int_0^{\lambda_E(\mathbf{f})} p_E dx. \tag{EC.1}$$

In this case, the solution to the following problem:

$$\min_{\mathbf{f} \in [0,1]^n} \Phi(\mathbf{f}), \tag{EC.2}$$

is the equilibrium patient flow, and we denote it by $\mathbf{f^e} = (f_1^e, f_2^e, ..., f_n^e)$. In particular, the first order partial derivative of $\int_0^{\lambda_E(\mathbf{f})} w_E Q_E(x) dx$ w.r.t $\lambda_E(\mathbf{f})$ is

$$\frac{\partial(\int_0^{\lambda_E(\mathbf{f})} w_E Q_E(x) dx)}{\partial \lambda_E(\mathbf{f})} = w_E Q_E(\lambda_E(\mathbf{f})) > 0,$$

while the second order partial derivative is

$$\frac{\partial^2(\int_0^{\lambda_E(\mathbf{f})} w_E Q_E(x) dx)}{\partial \lambda_E(\mathbf{f})^2} = w_E \frac{\partial Q_E(\lambda_E(\mathbf{f}))}{\partial \lambda_E(\mathbf{f})} > 0.$$

Hence $\int_0^{\lambda_E(\mathbf{f})} w_E Q_E(x) dx$ is strictly convex in $\lambda_E(\mathbf{f})$. As $\lambda_E(\mathbf{f})$ is linear in $\mathbf{f}$, by preservation of convexity, $\int_0^{\lambda_E(\mathbf{f})} w_E Q_E(x) dx$ is jointly convex in $\mathbf{f}$. As the remaining terms in (EC.1) are linear in $\mathbf{f}$, $\Phi(\mathbf{f})$ is jointly convex in $\mathbf{f}$.

Hence, for a given equilibrium patient flow $\mathbf{f^e}$ of Problem (EC.2), when we have an interior solution for f_i, i.e., $f_i^e \in (0,1)$, we then have

$$\left.\frac{\partial \Phi(\mathbf{f})}{\partial f_i}\right|_{\mathbf{f^e}} = -(p_G + w_G Q_G)\lambda_i + (1 - b_i)\lambda_i(p_E + w_E Q_E(\lambda_E(\mathbf{f^e}))) = 0. \tag{EC.3}$$

The intuition of (EC.3) is as follows. Group i patients adopt a mixed strategy in equilibrium by visiting the ED directly with a probability $f_i^e \in (0,1)$. In this case, they are indifferent between visiting a GP first and visiting the ED directly in equilibrium as the patient costs (disutilities of waiting + co-payments) for these two options are the same, i.e., $p_G + w_G Q_G + b_i(p_E + w_E Q_E(\lambda_E(\mathbf{f^e}))) = p_E + w_E Q_E(\lambda_E(\mathbf{f^e}))$. Hence, they have no incentive to change their strategy.

Instead, when we have a left corner solution for f_i with $f_i^e = 0$, we then have

$$\left.\frac{\partial \Phi(\mathbf{f})}{\partial f_i}\right|_{\mathbf{f^e}} = -(p_G + w_G Q_G)\lambda_i + (1-b_i)\lambda_i(p_E + w_E Q_E(\lambda_E(\mathbf{f^e}))) \geq 0. \tag{EC.4}$$

This corresponds to the scenario in which group i patients adopt a pure strategy in equilibrium by visiting a GP first with certainty. In this case, there is no patient cost reduction if they deviate from the equilibrium to visit the ED directly with a positive probability, as the patient costs for visiting the ED directly is no lower than visiting a GP first, i.e., $p_G + w_G Q_G + b_i(p_E + w_E Q_E(\lambda_E(\mathbf{f^e}))) \leq p_E + w_E Q_E(\lambda_E(\mathbf{f^e}))$.

Similarly, when we have a right corner solution for f_i with $f_i^e = 1$, we then have

$$\left.\frac{\partial \Phi(\mathbf{f})}{\partial f_i}\right|_{\mathbf{f^e}} = -(p_G + w_G Q_G)\lambda_i + (1-b_i)\lambda_i(p_E + w_E Q_E(\lambda_E(\mathbf{f^e}))) \leq 0. \tag{EC.5}$$

This corresponds to the scenario in which group i patients adopt a pure strategy in equilibrium by visiting the ED directly with certainty. In this case, there is no patient cost reduction if they deviate from the equilibrium, as the patient costs for visiting a GP first is no lower than visiting the ED directly, i.e., $p_G + w_G Q_G + b_i(p_E + w_E Q_E(\lambda_E(\mathbf{f^e}))) \geq p_E + w_E Q_E(\lambda_E(\mathbf{f^e}))$.

With this, we can then characterize the following structural property of $\mathbf{f^e}$.

LEMMA EC.1. *$\mathbf{f^e}$ satisfies the following structural property: $\exists\ i \in \{1,2,...,n\}$ s.t. we have $f_i^e \in [0,1], f_j^e = 0, \forall\ j < i$, and $f_k^e = 1, \forall\ k > i$.*

Proof of Lemma EC.1. Case 1: Suppose $\exists\ i \in \{1,2,...,n\}$ s.t. $f_i^e \in (0,1)$. In this case, we have an interior solution for f_i^e of Problem (EC.2), and therefore we have

$$\left.\frac{\partial \Phi(\mathbf{f})}{\partial f_i}\right|_{\mathbf{f^e}} = -(p_G + w_G Q_G)\lambda_i + (1-b_i)\lambda_i(p_E + w_E Q_E(\lambda_E(\mathbf{f^e}))) = 0. \tag{EC.6}$$

Since $\forall\ j < i$ we have $b_j < b_i$ by assumption, this implies that

$$\begin{aligned}\left.\frac{\partial \Phi(\mathbf{f})}{\partial f_j}\right|_{\mathbf{f^e}} &= -(p_G + w_G Q_G)\lambda_j + (1-b_j)\lambda_j(p_E + w_E Q_E(\lambda_E(\mathbf{f^e}))) > 0 \\ &\Leftrightarrow p_G + w_G Q_G + b_j(p_E + w_E Q_E(\lambda_E(\mathbf{f^e}))) < p_E + w_E Q_E(\lambda_E(\mathbf{f^e})),\end{aligned} \tag{EC.7}$$

given (EC.6). Hence, we have $f_j^e = 0, \forall\ j < i$: If $f_j^e > 0$, group j patients will find they can enjoy lower patient costs by visiting a GP first instead of visiting the ED directly according to (EC.7), which will lead to lower f_j^e. This contradicts to f_j^e being the equilibrium patient flow for group j patients. Similarly, $\forall\ k > i$ we have $b_k > b_i$ by assumption, and therefore we have

$$\begin{aligned}\left.\frac{\partial \Phi(\mathbf{f})}{\partial f_k}\right|_{\mathbf{f^e}} &= -(p_G + w_G Q_G)\lambda_k + (1-b_k)\lambda_k(p_E + w_E Q_E(\lambda_E(\mathbf{f^e}))) < 0 \\ &\Leftrightarrow p_G + w_G Q_G + b_k(p_E + w_E Q_E(\lambda_E(\mathbf{f^e}))) > p_E + w_E Q_E(\lambda_E(\mathbf{f^e})),\end{aligned} \tag{EC.8}$$

given (EC.6). Hence, we have $f_k^e = 1, \forall\ k > i$: If $f_k^e < 1$, group k patients will find they can enjoy lower patient costs by visiting the ED directly instead of visiting a GP first according to (EC.8), which will lead to higher f_k^e. This contradicts to f_k^e being the equilibrium patient flow for group k patients.

Case 2: Suppose $\nexists\ i \in \{1, 2, ..., n\}$ s.t. $f_i^e \in (0, 1)$. $\forall\ i$ s.t. $f_i^e = 0$, we have

$$\left.\frac{\partial \Phi(\mathbf{f})}{\partial f_i}\right|_{\mathbf{f^e}} = -(p_G + w_G Q_G)\lambda_i + (1 - b_i)\lambda_i(p_E + w_E Q_E(\lambda_E(\mathbf{f^e}))) \geq 0,$$

according to (EC.4). Since $\forall\ j < i$, this implies that

$$\left.\frac{\partial \Phi(\mathbf{f})}{\partial f_j}\right|_{\mathbf{f^e}} = -(p_G + w_G Q_G)\lambda_j + (1 - b_j)\lambda_j(p_E + w_E Q_E(\lambda_E(\mathbf{f^e}))) > 0.$$

Hence, we have $f_j^e = 0, \forall\ j < i$. Similarly, $\forall\ i$ s.t. $f_i^e = 1$, we have

$$\left.\frac{\partial \Phi(\mathbf{f})}{\partial f_i}\right|_{\mathbf{f^e}} = -(p_G + w_G Q_G)\lambda_i + (1 - b_i)\lambda_i(p_E + w_E Q_E(\lambda_E(\mathbf{f^e}))) \leq 0,$$

according to (EC.5). Since $\forall\ k > i$ we have $b_k > b_i$ by assumption, this implies that

$$\left.\frac{\partial \Phi(\mathbf{f})}{\partial f_k}\right|_{\mathbf{f^e}} = -(p_G + w_G Q_G)\lambda_k + (1 - b_k)\lambda_k(p_E + w_E Q_E(\lambda_E(\mathbf{f^e}))) < 0.$$

Hence, we have $f_k^e = 1, \forall\ k > i$. □

Lemma EC.1 shows that equilibrium patient flow f_i^e is non-decreasing in i, when we assume patients' belief of being ED-type b_i is increasing in i. Moreover, for any equilibrium patient flow, there is at most one group of patients adopts mixed strategy. In this case, for the rest of the groups of patients, those with lower belief of being ED-type will adopt pure strategy by visiting a GP first, and those with higher belief of being ED-type will adopt pure strategy by visiting the ED directly. Given this structural property of $\mathbf{f^e}$ characterized by Lemma EC.1, we now prove the uniqueness of $\mathbf{f^e}$ with the following lemma.

LEMMA EC.2. *There exists a unique patient flow* $\mathbf{f^e}$ *in equilibrium.*

Proof of Lemma EC.2. Let $\mathbf{f^e}$ denote an equilibrium patient flow.

Case 1: Suppose $\exists\ i \in \{1, 2, ..., n\}$ s.t. $f_i^e \in (0, 1)$. Then we have

$$\left.\frac{\partial \Phi(\mathbf{f})}{\partial f_i}\right|_{\mathbf{f^e}} = -(p_G + w_G Q_G) + (1 - b_i)(p_E + w_E Q_E(\lambda_E(\mathbf{f^e}))) = 0, \tag{EC.9}$$

which determines the unique $f_i^e \in (0, 1)$. We first exclude the possibility of having another mixed strategy equilibrium patient flow $\mathbf{f^{e,a}}$ where we have $f_j^{e,a} \in (0, 1), j < i$. We prove by contradiction.

Suppose we have another mixed strategy equilibrium patient flow $\mathbf{f^{e,a}}$ where $f_j^{e,a} \in (0, 1), j < i$. We then have

$$\left.\frac{\partial \Phi(\mathbf{f})}{\partial f_j}\right|_{\mathbf{f^{e,a}}} = -(p_G + w_G Q_G) + (1 - b_j)(p_E + w_E Q_E(\lambda_E(\mathbf{f^{e,a}}))) = 0.$$

We have $b_j < b_i$ by assumption. Moreover, we have $\lambda_E(\mathbf{f^{e,a}}) > \lambda_E(\mathbf{f^e})$: $j < i$ and Lemma EC.1 imply that there are more patients with higher beliefs of being ED-type visiting the ED directly under $\mathbf{f^{e,a}}$ than under $\mathbf{f^e}$, and this leads to $\lambda_E(\mathbf{f^{e,a}}) > \lambda_E(\mathbf{f^e})$. Together, this implies

$$\left.\frac{\partial \Phi(\mathbf{f})}{\partial f_i}\right|_{\mathbf{f^e}} = -(p_G + w_G Q_G) + (1-b_i)(p_E + w_E Q_E(\lambda_E(\mathbf{f^e}))) < 0,$$

which contradicts to (EC.9). Similarly, we can also exclude the possibility of having another mixed strategy equilibrium patient flow $\mathbf{f^{e,a}}$ where $f_k^{e,a} \in (0,1), k > i$, as well as the possibility of having another pure strategy equilibrium.

Case 2: Suppose $\nexists\ i \in \{1,2,...,n\}$ s.t. $f_i^e \in (0,1)$. Suppose we have $f_i^e = 0$. Then we have

$$\left.\frac{\partial \Phi(\mathbf{f})}{\partial f_i}\right|_{\mathbf{f^e}} = -(p_G + w_G Q_G) + (1-b_i)(p_E + w_E Q_E(\lambda_E(\mathbf{f^e}))) \geq 0. \tag{EC.10}$$

We first exclude the possibility of having another pure strategy equilibrium patient flow $\mathbf{f^{e,a}}$ where $f_j^{e,a} = 1, j < i$. We prove by contradiction.

Suppose we have another pure strategy equilibrium patient flow $\mathbf{f^{e,a}}$ where $f_j^{e,a} = 1, j < i$. We then have

$$\left.\frac{\partial \Phi(\mathbf{f})}{\partial f_j}\right|_{\mathbf{f^{e,a}}} = -(p_G + w_G Q_G) + (1-b_j)(p_E + w_E Q_E(\lambda_E(\mathbf{f^{e,a}}))) \leq 0.$$

Since we have $b_j < b_i$ by assumption, as well as $\lambda_E(\mathbf{f^{e,a}}) > \lambda_E(\mathbf{f^e})$ given $j < i$ and Lemma EC.1, this implies

$$\left.\frac{\partial \Phi(\mathbf{f})}{\partial f_i}\right|_{\mathbf{f^e}} = -(p_G + w_G Q_G) + (1-b_i)(p_E + w_E Q_E(\lambda_E(\mathbf{f^e}))) < 0,$$

which contradicts to (EC.10). Similarly, we can also exclude the possibility of having $f_i^e = 1$ and another pure strategy equilibrium patient flow $\mathbf{f^{e,a}}$ where $f_k^{e,a} = 0, k > i$, as well as the possibility of having another mixed strategy equilibrium. □

EC.4. Proofs

This section provides detailed proofs of all the mathematical results in the paper.

EC.4.1. Proofs for Section 3

Proof of Lemma 1. Let $p_H = \int_0^1 sg(s)ds$ denote the fraction of H patients in the patient base and $p_L = 1 - p_H$ denote the fraction of L patients. We have the virtual under-triage probability defined as

$$\tilde{\alpha}(\bar{s}) = Prob(\tilde{L}|H) = \frac{\int_0^{\bar{s}} sg(s)ds}{p_H}, \tag{EC.11}$$

and the virtual over-triage probability defined as

$$\tilde{\beta}(\bar{s}) = Prob(\tilde{H}|L) = \frac{\int_{\bar{s}}^1 (1-s)g(s)ds}{p_L} = 1 - \frac{\int_0^{\bar{s}} (1-s)g(s)ds}{p_L}. \tag{EC.12}$$

Clearly when $\bar{s}=0$, we have $\tilde{\alpha}(\bar{s})=0, \tilde{\beta}(\bar{s})=1$; when $\bar{s}=1$, we have $\tilde{\alpha}(\bar{s})=1, \tilde{\beta}(\bar{s})=0$. In addition, we have

$$\frac{\partial\tilde{\alpha}}{\partial\tilde{\beta}}=\frac{\partial\tilde{\alpha}}{\partial\bar{s}}\frac{\partial\bar{s}}{\partial\tilde{\beta}}=\frac{\bar{s}g(\bar{s})}{p_H}[-\frac{p_L}{(1-\bar{s})g(\bar{s})}]=-\frac{p_L}{p_H}\frac{\bar{s}}{(1-\bar{s})}\leq 0,$$

and

$$\frac{\partial^2\tilde{\alpha}}{\partial\tilde{\beta}^2}=-\frac{p_L}{p_H}\frac{\partial[\bar{s}/(1-\bar{s})]}{\partial\tilde{\beta}}=-\frac{p_L}{p_H}\frac{\partial[\bar{s}/(1-\bar{s})]}{\partial\bar{s}}\frac{\partial\bar{s}}{\partial\tilde{\beta}}=\frac{p_L^2}{p_H(1-\bar{s})^3g(\bar{s})}\geq 0.$$

Hence, $\tilde{\alpha}=r(\tilde{\beta})$ is a decreasing and convex function in $\tilde{\beta}$, with $r(0)=1, r(1)=0$. Hence, we also have $\tilde{\alpha}+\tilde{\beta}\leq 1$. □

Proof of Lemma 2. $\forall\ \bar{s}_1,\bar{s}_2\in[0,1]$ s.t.

$$\frac{\int_{\bar{s}_1}^1(1-s)g_1(s)ds}{p_L}=\frac{\int_{\bar{s}_2}^1(1-s)g_2(s)ds}{p_L},$$

we have

$$\frac{\int_0^{\bar{s}_1}sg_1(s)ds}{p_H}\geq\frac{\int_0^{\bar{s}_2}sg_2(s)ds}{p_H}.$$

Hence, $\forall\ \bar{s}_1,\bar{s}_2\in[0,1]$ s.t. $\tilde{\beta}(\bar{s}_1)=\tilde{\beta}(\bar{s}_2)$, we have $\tilde{\alpha}(\bar{s}_1)\geq\tilde{\alpha}(\bar{s}_2)$ according to (EC.11) and (EC.12). This implies $r_1(\tilde{\beta})\geq r_2(\tilde{\beta})$, $\forall\ \tilde{\beta}\in[0,1]$. □

EC.4.2. Proofs for Section 4

Proof of Proposition 1. By Lemmas EC.1 and EC.2 with $n=2$, the unique $\hat{\mathbf{f}}^{\mathbf{e}}$ in the absence of virtual triage takes one of the following forms: $\hat{\mathbf{f}}^{\mathbf{e}}\in\{(0,0),(0,m),(0,1),(m,1),(1,1)\}$, as we have $b_{\hat{L}}\leq b_{\hat{H}}$ with $\hat{\alpha}+\hat{\beta}\leq 1$. Note that we have $m\in(0,1)$.

Meanwhile, with the adoption of virtual triage, we have $b_{\hat{L}\tilde{L}}\leq b_{\hat{H}\tilde{L}}, b_{\hat{L}\tilde{H}}\leq b_{\hat{H}\tilde{H}}$ as $\hat{\alpha}+\hat{\beta}\leq 1$, and we have $b_{\hat{L}\tilde{L}}\leq b_{\hat{L}\tilde{H}}, b_{\hat{H}\tilde{L}}\leq b_{\hat{H}\tilde{H}}$ as $\tilde{\alpha}+\tilde{\beta}\leq 1$. Moreover, we may have $b_{\hat{L}\tilde{H}}\leq b_{\hat{H}\tilde{L}}$ or $b_{\hat{L}\tilde{H}}>b_{\hat{H}\tilde{L}}$ depending on the values of $\hat{\alpha},\hat{\beta},\tilde{\alpha}$ and $\tilde{\beta}$.

When we have $b_{\hat{L}\tilde{L}}\leq b_{\hat{L}\tilde{H}}\leq b_{\hat{H}\tilde{L}}\leq b_{\hat{H}\tilde{H}}$, by Lemmas EC.1 and EC.2 with $n=4$, the unique $\tilde{\mathbf{f}}^{\mathbf{e}}$ in the presence of virtual triage takes one of the following forms: $\tilde{\mathbf{f}}^{\mathbf{e}}\in\{(0,0,0,0),(0,0,0,m),(0,0,0,1),(0,0,m,1),(0,0,1,1),(0,m,1,1),(0,1,1,1),(m,1,1,1),(1,1,1,1)\}$. Meanwhile, when we have $b_{\hat{L}\tilde{L}}\leq b_{\hat{H}\tilde{L}}\leq b_{\hat{L}\tilde{H}}\leq b_{\hat{H}\tilde{H}}$, the unique $\tilde{\mathbf{f}}^{\mathbf{e}}$ in the presence of virtual triage takes one of the following forms: $\tilde{\mathbf{f}}^{\mathbf{e}}\in\{(0,0,0,0),(0,0,0,m),(0,0,0,1),(0,m,0,1),(0,1,0,1),(0,1,m,1),(0,1,1,1),(m,1,1,1),(1,1,1,1)\}$. Combining these two cases, we have

$$\begin{aligned}\tilde{\mathbf{f}}^{\mathbf{e}}\in\{&(0,0,0,0),(0,0,0,m),(0,0,0,1),(0,0,m,1),(0,0,1,1),(0,m,1,1),\\&(0,1,1,1),(m,1,1,1),(1,1,1,1),(0,m,0,1),(0,1,0,1),(0,1,m,1)\}.\end{aligned}\tag{EC.13}$$

Let $C_{\hat{T},l}(\hat{\mathbf{f}})$ denote the expected patient costs, i.e., sum of the total disutility of waiting and co-payment, for a patient of type $\hat{T}$ visiting l given a patient flow $\hat{\mathbf{f}}$ in the absence of virtual triage,

where $\hat{T} \in \{\hat{L}, \hat{H}\}$ and $l \in \{G, E\}$ (G denotes GP and E denotes ED). Let $C_{\hat{T}\tilde{T},l}(\tilde{\mathbf{f}})$ denote the expected patient costs for a patient of type $\hat{T}\tilde{T}$ visiting l given a patient flow $\tilde{\mathbf{f}}$ in the presence of virtual triage, where $\tilde{T} \in \{\tilde{L}, \tilde{H}\}$. Specifically, we have

$$\begin{aligned} C_{\hat{T},G}(\hat{\mathbf{f}}) &= p_G + w_G Q_G + b_{\hat{T}}[p_E + w_E Q_E(\lambda_E(\hat{\mathbf{f}}))] \\ C_{\hat{T},E}(\hat{\mathbf{f}}) &= p_E + w_E Q_E(\lambda_E(\hat{\mathbf{f}})), \end{aligned}$$

in the absence of virtual triage. Patients of type $\hat{T}$ then compare $C_{\hat{T},G}(\hat{\mathbf{f}})$ and $C_{\hat{T},E}(\hat{\mathbf{f}})$ and choose the choice of care location (i.e., GP or ED) with lower expected patient costs. Similarly, we have

$$\begin{aligned} C_{\hat{T}\tilde{T},G}(\tilde{\mathbf{f}}) &= p_G + w_G Q_G + b_{\hat{T}\tilde{T}}[p_E + w_E Q_E(\lambda_E(\tilde{\mathbf{f}}))] \\ C_{\hat{T}\tilde{T},E}(\tilde{\mathbf{f}}) &= p_E + w_E Q_E(\lambda_E(\tilde{\mathbf{f}})), \end{aligned}$$

in the presence of virtual triage. Patients of type $\hat{T}\tilde{T}$ then compare $C_{\hat{T}\tilde{T},G}(\hat{\mathbf{f}})$ and $C_{\hat{T}\tilde{T},E}(\hat{\mathbf{f}})$ and choose the choice of care location with lower expected patient costs.

For each of the five cases of $\hat{\mathbf{f}}^{\mathbf{e}}$, we characterize the existence of associated $\tilde{\mathbf{f}}^{\mathbf{e}}$ and the relative positions of each region by characterizing their boundaries.

(i) When $\hat{\mathbf{f}}^{\mathbf{e}} = (0,0)$, both $\hat{L}$ and $\hat{H}$ patients visit a GP first with certainty in equilibrium in the absence of virtual triage, as visiting the ED directly with a positive probability does not lead to lower patient costs for either of them. Hence, we have

$$\begin{aligned} & C_{\hat{H},E}(0,0) \geq C_{\hat{H},G}(0,0) \\ \Leftrightarrow\ & (1-b_{\hat{H}})[p_E + w_E Q_E(b_{\hat{L}}\lambda_{\hat{L}} + b_{\hat{H}}\lambda_{\hat{H}})] \geq p_G + w_G Q_G. \end{aligned} \tag{EC.14}$$

Note that we also have $C_{\hat{L},E}(0,0) \geq C_{\hat{L},G}(0,0)$ in this case, which is implied by (EC.14) since $b_{\hat{L}} \leq b_{\hat{H}}$. In this case, in the presence of virtual triage, $\forall\ \tilde{\alpha} + \tilde{\beta} \leq 1$, we can show that $f^e_{\hat{H}\tilde{L}} = 0$ with proof by contradiction. Let $\tilde{\mathbf{f}}^{\mathbf{e}} = (f^e_{\hat{L}\tilde{L}}, f^e_{\hat{L}\tilde{H}}, f^e_{\hat{H}\tilde{L}}, f^e_{\hat{H}\tilde{H}})$ denote the equilibrium patient flow in the presence of virtual triage. If $f^e_{\hat{H}\tilde{L}} > 0$, we have $\hat{H}\tilde{L}$ patients find lower patient costs by visiting the ED directly with a positive probability under $\tilde{\mathbf{f}} = (f^e_{\hat{L}\tilde{L}}, f^e_{\hat{L}\tilde{H}}, 0, f^e_{\hat{H}\tilde{H}})$. Hence, we have

$$\begin{aligned} & C_{\hat{H}\tilde{L},E}(f^e_{\hat{L}\tilde{L}}, f^e_{\hat{L}\tilde{H}}, 0, f^e_{\hat{H}\tilde{H}}) < C_{\hat{H}\tilde{L},G}(f^e_{\hat{L}\tilde{L}}, f^e_{\hat{L}\tilde{H}}, 0, f^e_{\hat{H}\tilde{H}}) \\ \Leftrightarrow\ & (1-b_{\hat{H}\tilde{L}})[p_E + w_E Q_E(\lambda_E(f^e_{\hat{L}\tilde{L}}, f^e_{\hat{L}\tilde{H}}, 0, f^e_{\hat{H}\tilde{H}}))] < p_G + w_G Q_G \\ \Rightarrow\ & (1-b_{\hat{H}\tilde{L}})[p_E + w_E Q_E(\lambda_E(0,0,0,0))] < p_G + w_G Q_G, \end{aligned}$$

as we have $\lambda_E(f^e_{\hat{L}\tilde{L}}, f^e_{\hat{L}\tilde{H}}, 0, f^e_{\hat{H}\tilde{H}}) > \lambda_H = \lambda_E(0,0,0,0)$. Hence, we have $C_{\hat{H}\tilde{L},E}(0,0,0,0) < C_{\hat{H}\tilde{L},G}(0,0,0,0)$, i.e., $(1-b_{\hat{H}\tilde{L}})[p_E + w_E Q_E(\lambda_H)] < p_G + w_G Q_G$. This implies $(1-b_{\hat{H}})[p_E + w_E Q_E(\lambda_H)] < p_G + w_G Q_G$, i.e., $C_{\hat{H},E}(0,0) < C_{\hat{H},G}(0,0)$, since we have $b_{\hat{H}} \geq b_{\hat{H}\tilde{L}}$. Hence, we have $C_{\hat{H},E}(0,0) < C_{\hat{H},G}(0,0)$, which contradicts (EC.14). This proves that $f^e_{\hat{H}\tilde{L}} = 0$.

With $f^e_{\hat{H}\tilde{L}} = 0$, we also have $f^e_{\hat{L}\tilde{L}} = 0$ by Lemma EC.1 since $b_{\hat{L}\tilde{L}} \leq b_{\hat{H}\tilde{L}}$. Hence, out of the 12 different equilibrium regions of (EC.13) in the presence of virtual triage, we have $\tilde{\mathbf{f}}^{\mathbf{e}} \in$

$\{(0,0,0,0),(0,0,0,m),(0,0,0,1),(0,m,0,1),(0,1,0,1)\}$ when $\hat{\mathbf{f}}^{\mathbf{e}}=(0,0)$. We now characterize each of the five equilibrium regions in the presence of virtual triage in this case as follows.

$\tilde{\mathbf{f}}^{\mathbf{e}}=(0,0,0,0)$: In this case, no patients enjoy lower patient costs by visiting the ED directly when all patients go to a GP first in equilibrium. In particular, $\hat{H}\tilde{H}$ patients do not enjoy lower patient costs by visiting the ED directly with a positive probability when we have $\tilde{\mathbf{f}}=(0,0,0,0)$:

$$\begin{aligned} & C_{\hat{H}\tilde{H},E}(0,0,0,0)\geq C_{\hat{H}\tilde{H},G}(0,0,0,0) \\ \Leftrightarrow\ & (1-b_{\hat{H}\tilde{H}})[p_E+w_EQ_E(b_{\hat{L}}\lambda_{\hat{L}}+b_{\hat{H}}\lambda_{\hat{H}})]\geq p_G+w_GQ_G. \end{aligned} \tag{EC.15}$$

Note that $\hat{L}\tilde{L}$, $\hat{L}\tilde{H}$ and $\hat{H}\tilde{L}$ patients also do not enjoy lower patient costs by visiting the ED directly with a positive probability, and they are implied by (EC.15) since $b_{\hat{L}\tilde{L}}\leq b_{\hat{H}\tilde{H}}, b_{\hat{L}\tilde{H}}\leq b_{\hat{H}\tilde{H}}, b_{\hat{H}\tilde{L}}\leq b_{\hat{H}\tilde{H}}$.

$\tilde{\mathbf{f}}^{\mathbf{e}}=(0,0,0,m)$: In this case, $\hat{H}\tilde{H}$ patients adopt a mixed strategy in equilibrium. This implies that, under patient flow $\tilde{\mathbf{f}}=(0,0,0,0)$, $\hat{H}\tilde{H}$ patients enjoy lower patient costs by visiting the ED directly with a positive probability; meanwhile, under patient flow $\tilde{\mathbf{f}}=(0,0,0,1)$, $\hat{H}\tilde{H}$ patients enjoy lower patient costs by going to a GP first with a positive probability:

$$\begin{aligned} & C_{\hat{H}\tilde{H},E}(0,0,0,0)<C_{\hat{H}\tilde{H},G}(0,0,0,0)\wedge C_{\hat{H}\tilde{H},E}(0,0,0,1)>C_{\hat{H}\tilde{H},G}(0,0,0,1) \\ \Leftrightarrow\ & (1-b_{\hat{H}\tilde{H}})[p_E+w_EQ_E(b_{\hat{L}}\lambda_{\hat{L}}+b_{\hat{H}}\lambda_{\hat{H}})]<p_G+w_GQ_G\wedge \\ & (1-b_{\hat{H}\tilde{H}})[p_E+w_EQ_E(b_{\hat{L}}\lambda_{\hat{L}}+b_{\hat{H}}\lambda_{\hat{H}}+\tilde{\beta}(1-b_{\hat{H}})\lambda_{\hat{H}})]>p_G+w_GQ_G, \end{aligned}$$

where $\lambda_E(0,0,0,0)=b_{\hat{L}}\lambda_{\hat{L}}+b_{\hat{H}}\lambda_{\hat{H}}$ and $\lambda_E(0,0,0,1)=b_{\hat{L}}\lambda_{\hat{L}}+b_{\hat{H}}\lambda_{\hat{H}}+\tilde{\beta}(1-b_{\hat{H}})\lambda_{\hat{H}}$.

$\tilde{\mathbf{f}}^{\mathbf{e}}=(0,0,0,1)$: In this case, under patient flow $\tilde{\mathbf{f}}=(0,0,0,1)$, $\hat{H}\tilde{H}$ patients do not enjoy lower patient costs by going to a GP first with a positive probability, while $\hat{L}\tilde{H}$ and $\hat{H}\tilde{L}$ patients do not enjoy lower patient costs by visiting the ED directly with a positive probability. Note that depending on the virtual triage accuracy, we may have $b_{\hat{L}\tilde{H}}\leq b_{\hat{H}\tilde{L}}$ or $b_{\hat{H}\tilde{L}}\leq b_{\hat{L}\tilde{H}}$. If we have $b_{\hat{L}\tilde{H}}\leq b_{\hat{H}\tilde{L}}$ ($b_{\hat{H}\tilde{L}}\leq b_{\hat{L}\tilde{H}}$), $\hat{L}\tilde{H}$ ($\hat{H}\tilde{L}$) patients do not enjoy lower patient costs by visiting the ED directly with a positive probability is implied by $\hat{H}\tilde{L}$ ($\hat{L}\tilde{H}$) patients do not enjoy lower patient costs by visiting the ED directly with a positive probability. Moreover, either of these two conditions implies that $\hat{L}\tilde{L}$ patients do not enjoy lower patient costs by visiting the ED directly with a positive probability. Putting these conditions together, we have

$$\begin{aligned} & C_{\hat{H}\tilde{H},E}(0,0,0,1)\leq C_{\hat{H}\tilde{H},G}(0,0,0,1)\wedge C_{\hat{H}\tilde{L},E}(0,0,0,1)\geq C_{\hat{H}\tilde{L},G}(0,0,0,1) \\ & \qquad \wedge C_{\hat{L}\tilde{H},E}(0,0,0,1)\geq C_{\hat{L}\tilde{H},G}(0,0,0,1) \\ \Leftrightarrow\ & (1-b_{\hat{H}\tilde{H}})[p_E+w_EQ_E(b_{\hat{L}}\lambda_{\hat{L}}+b_{\hat{H}}\lambda_{\hat{H}}+\tilde{\beta}(1-b_{\hat{H}})\lambda_{\hat{H}})]\leq p_G+w_GQ_G\wedge \\ & (1-b_{\hat{H}\tilde{L}})[p_E+w_EQ_E(b_{\hat{L}}\lambda_{\hat{L}}+b_{\hat{H}}\lambda_{\hat{H}}+\tilde{\beta}(1-b_{\hat{H}})\lambda_{\hat{H}})]\geq p_G+w_GQ_G\wedge \\ & (1-b_{\hat{L}\tilde{H}})[p_E+w_EQ_E(b_{\hat{L}}\lambda_{\hat{L}}+b_{\hat{H}}\lambda_{\hat{H}}+\tilde{\beta}(1-b_{\hat{H}})\lambda_{\hat{H}})]\geq p_G+w_GQ_G, \end{aligned}$$

where $\lambda_E(0,0,0,1)=b_{\hat{L}}\lambda_{\hat{L}}+b_{\hat{H}}\lambda_{\hat{H}}+\tilde{\beta}(1-b_{\hat{H}})\lambda_{\hat{H}}$.

$\tilde{\mathbf{f}}^{\mathbf{e}} = (0, m, 0, 1)$: In this case, we have $b_{\hat{H}\tilde{L}} \leq b_{\hat{L}\tilde{H}}$ by Lemma EC.1, and $\hat{L}\tilde{H}$ patients adopt a mixed strategy in equilibrium. This implies that, under patient flow $\tilde{\mathbf{f}} = (0, 0, 0, 1)$, $\hat{L}\tilde{H}$ patients enjoy lower patient costs by visiting the ED directly with a positive probability; meanwhile, under patient flow $\tilde{\mathbf{f}} = (0, 1, 0, 1)$, $\hat{L}\tilde{H}$ patients enjoy lower patient costs by going to a GP first with a positive probability:

$$\begin{aligned} & C_{\hat{L}\tilde{H},E}(0,0,0,1) < C_{\hat{L}\tilde{H},G}(0,0,0,1) \wedge C_{\hat{L}\tilde{H},E}(0,1,0,1) > C_{\hat{L}\tilde{H},G}(0,1,0,1) \\ \Leftrightarrow\ & (1-b_{\hat{L}\tilde{H}})[p_E + w_E Q_E(b_{\hat{L}}\lambda_{\hat{L}} + b_{\hat{H}}\lambda_{\hat{H}} + \tilde{\beta}(1-b_{\hat{H}})\lambda_{\hat{H}})] < p_G + w_G Q_G \wedge \\ & (1-b_{\hat{L}\tilde{H}})[p_E + w_E Q_E(b_{\hat{L}}\lambda_{\hat{L}} + b_{\hat{H}}\lambda_{\hat{H}} + \tilde{\beta}(1-b_{\hat{L}})\lambda_{\hat{L}} + \tilde{\beta}(1-b_{\hat{H}})\lambda_{\hat{H}})] > p_G + w_G Q_G, \end{aligned}$$

where $\lambda_E(0,0,0,1) = b_{\hat{L}}\lambda_{\hat{L}} + b_{\hat{H}}\lambda_{\hat{H}} + \tilde{\beta}(1-b_{\hat{H}})\lambda_{\hat{H}}$ and $\lambda_E(0,1,0,1) = b_{\hat{L}}\lambda_{\hat{L}} + b_{\hat{H}}\lambda_{\hat{H}} + \tilde{\beta}(1-b_{\hat{L}})\lambda_{\hat{L}} + \tilde{\beta}(1-b_{\hat{H}})\lambda_{\hat{H}}$.

$\tilde{\mathbf{f}}^{\mathbf{e}} = (0, 1, 0, 1)$: In this case, we have $b_{\hat{H}\tilde{L}} \leq b_{\hat{L}\tilde{H}}$ by Lemma EC.1. In addition, under patient flow $\tilde{\mathbf{f}} = (0, 1, 0, 1)$, $\hat{L}\tilde{H}$ patients do not enjoy lower patient costs by going to a GP first with a positive probability, while $\hat{H}\tilde{L}$ patients do not enjoy lower patient costs by visiting the ED directly with a positive probability:

$$\begin{aligned} & C_{\hat{L}\tilde{H},E}(0,1,0,1) \leq C_{\hat{L}\tilde{H},G}(0,1,0,1) \wedge C_{\hat{H}\tilde{L},E}(0,1,0,1) \geq C_{\hat{H}\tilde{L},G}(0,1,0,1) \\ \Leftrightarrow\ & (1-b_{\hat{L}\tilde{H}})[p_E + w_E Q_E(b_{\hat{L}}\lambda_{\hat{L}} + b_{\hat{H}}\lambda_{\hat{H}} + \tilde{\beta}(1-b_{\hat{L}})\lambda_{\hat{L}} + \tilde{\beta}(1-b_{\hat{H}})\lambda_{\hat{H}})] \leq p_G + w_G Q_G \wedge \\ & (1-b_{\hat{H}\tilde{L}})[p_E + w_E Q_E(b_{\hat{L}}\lambda_{\hat{L}} + b_{\hat{H}}\lambda_{\hat{H}} + \tilde{\beta}(1-b_{\hat{L}})\lambda_{\hat{L}} + \tilde{\beta}(1-b_{\hat{H}})\lambda_{\hat{H}})] \geq p_G + w_G Q_G, \end{aligned} \tag{EC.16}$$

where $\lambda_E(0,1,0,1) = b_{\hat{L}}\lambda_{\hat{L}} + b_{\hat{H}}\lambda_{\hat{H}} + \tilde{\beta}(1-b_{\hat{L}})\lambda_{\hat{L}} + \tilde{\beta}(1-b_{\hat{H}})\lambda_{\hat{H}}$. Note that in this case, $\hat{L}\tilde{L}$ patients do not enjoy lower patient costs by visiting the ED directly with a positive probability, and $\hat{H}\tilde{H}$ patients do not enjoy lower patient costs by going to a GP first with a positive probability, and they are implied by (EC.16).

(ii) When $\hat{\mathbf{f}}^{\mathbf{e}} = (0, m)$, $\hat{H}$ patients adopt a mixed strategy in equilibrium in the absence of virtual triage. This implies that, under patient flow $\hat{\mathbf{f}} = (0, 0)$, $\hat{H}$ patients enjoy lower patient costs by visiting the ED directly with a positive probability; meanwhile, under patient flow $\hat{\mathbf{f}} = (0, 1)$, $\hat{H}$ patients enjoy lower patient costs by going to a GP first with a positive probability. Hence, we have

$$\begin{aligned} & C_{\hat{H},E}(0,0) < C_{\hat{H},G}(0,0) \wedge C_{\hat{H},E}(0,1) > C_{\hat{H},G}(0,1) \\ \Leftrightarrow\ & (1-b_{\hat{H}})[p_E + w_E Q_E(b_{\hat{L}}\lambda_{\hat{L}} + b_{\hat{H}}\lambda_{\hat{H}})] < p_G + w_G Q_G \wedge \\ & (1-b_{\hat{H}})[p_E + w_E Q_E(b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}})] > p_G + w_G Q_G, \end{aligned} \tag{EC.17}$$

where $\lambda_E(0,0) = b_{\hat{L}}\lambda_{\hat{L}} + b_{\hat{H}}\lambda_{\hat{H}}$ and $\lambda_E(0,1) = b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}}$. In this case, in the presence of virtual triage, $\forall\ \tilde{\alpha} + \tilde{\beta} \leq 1$, we can show that $f^e_{\hat{L}\tilde{L}} = 0$ with proof by contradiction. If $f^e_{\hat{L}\tilde{L}} > 0$, we have $\hat{L}\tilde{L}$

patients find lower patient costs by visiting the ED directly under $\tilde{\mathbf{f}} = (0, f^e_{\hat{L}\tilde{H}}, f^e_{\hat{H}\tilde{L}}, f^e_{\hat{H}\tilde{H}})$. Hence, we have

$$\begin{aligned} & C_{\hat{L}\tilde{L},E}(0, f^e_{\hat{L}\tilde{H}}, f^e_{\hat{H}\tilde{L}}, f^e_{\hat{H}\tilde{H}}) < C_{\hat{L}\tilde{L},G}(0, f^e_{\hat{L}\tilde{H}}, f^e_{\hat{H}\tilde{L}}, f^e_{\hat{H}\tilde{H}}) \\ \Leftrightarrow\ & (1-b_{\hat{L}\tilde{L}})[p_E + w_E Q_E(\lambda_E(0, f^e_{\hat{L}\tilde{H}}, f^e_{\hat{H}\tilde{L}}, f^e_{\hat{H}\tilde{H}}))] < p_G + w_G Q_G \\ \Rightarrow\ & (1-b_{\hat{L}\tilde{L}})[p_E + w_E Q_E(\lambda_E(0,0,1,1))] < p_G + w_G Q_G, \end{aligned}$$

as we have $\lambda_E(0, f^e_{\hat{L}\tilde{H}}, f^e_{\hat{H}\tilde{L}}, f^e_{\hat{H}\tilde{H}}) > \lambda_E(0,0,1,1)$: if $f^e_{\hat{L}\tilde{L}} > 0$, we have $f^e_{\hat{L}\tilde{H}} = f^e_{\hat{H}\tilde{L}} = f^e_{\hat{H}\tilde{H}} = 1$ by Lemma EC.1, which implies $\lambda_E(0, f^e_{\hat{L}\tilde{H}}, f^e_{\hat{H}\tilde{L}}, f^e_{\hat{H}\tilde{H}}) > \lambda_E(0,0,1,1)$. Hence, we have $C_{\hat{L}\tilde{L},E}(0,0,1,1) < C_{\hat{L}\tilde{L},G}(0,0,1,1)$, i.e., $(1-b_{\hat{L}\tilde{L}})[p_E + w_E Q_E(b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}})] < p_G + w_G Q_G$. This implies $(1-b_{\hat{H}})[p_E + w_E Q_E(b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}})] < p_G + w_G Q_G$, i.e., $C_{\hat{H},E}(0,1) < C_{\hat{H},G}(0,1)$, since we have $b_{\hat{H}} \geq b_{\hat{L}\tilde{L}}$. Hence, we have $C_{\hat{H},E}(0,1) < C_{\hat{H},G}(0,1)$, which contradicts (EC.17). This proves that $f^e_{\hat{L}\tilde{L}} = 0$.

Moreover, for $\forall\ \tilde{\alpha} + \tilde{\beta} \leq 1$, we can show that $f^e_{\hat{H}\tilde{H}} > 0$ with proof by contradiction. If $f^e_{\hat{H}\tilde{H}} = 0$, $\hat{H}\tilde{H}$ patients do not enjoy lower patient costs by visiting the ED directly under $\tilde{\mathbf{f}} = (f^e_{\hat{L}\tilde{L}}, f^e_{\hat{L}\tilde{H}}, f^e_{\hat{H}\tilde{L}}, 0)$, and we have $f^e_{\hat{L}\tilde{L}} = f^e_{\hat{L}\tilde{H}} = f^e_{\hat{H}\tilde{L}} = 0$ here by Lemma EC.1. Hence, we have

$$\begin{aligned} & C_{\hat{H}\tilde{H},E}(0,0,0,0) \geq C_{\hat{H}\tilde{H},G}(0,0,0,0) \\ \Leftrightarrow\ & (1-b_{\hat{H}\tilde{H}})[p_E + w_E Q_E(\lambda_E(0,0,0,0))] \geq p_G + w_G Q_G \\ \Rightarrow\ & (1-b_{\hat{H}})[p_E + w_E Q_E(\lambda_E(0,0,0,0))] \geq p_G + w_G Q_G, \end{aligned}$$

since we have $b_{\hat{H}\tilde{H}} \geq b_{\hat{H}}$. This implies that $C_{\hat{H},E}(0,0) \geq C_{\hat{H},G}(0,0)$, which contradicts (EC.17). This proves that $f^e_{\hat{H}\tilde{H}} > 0$.

Similarly, for $\forall\ \tilde{\alpha} + \tilde{\beta} \leq 1$, we can show that $f^e_{\hat{H}\tilde{L}} < 1$ with proof by contradiction. If $f^e_{\hat{H}\tilde{L}} = 1$, $\hat{H}\tilde{L}$ patients do not enjoy lower patient costs by going to a GP first under $\tilde{\mathbf{f}} = (f^e_{\hat{L}\tilde{L}}, f^e_{\hat{L}\tilde{H}}, 1, f^e_{\hat{H}\tilde{H}})$, and we have $f^e_{\hat{H}\tilde{H}} = 1$ here by Lemma EC.1. Hence, we have

$$\begin{aligned} & C_{\hat{H}\tilde{L},E}(f^e_{\hat{L}\tilde{L}}, f^e_{\hat{L}\tilde{H}}, 1, 1) \leq C_{\hat{H}\tilde{L},G}(f^e_{\hat{L}\tilde{L}}, f^e_{\hat{L}\tilde{H}}, 1, 1) \\ \Leftrightarrow\ & (1-b_{\hat{H}\tilde{L}})[p_E + w_E Q_E(\lambda_E(f^e_{\hat{L}\tilde{L}}, f^e_{\hat{L}\tilde{H}}, 1, 1))] \leq p_G + w_G Q_G \\ \Rightarrow\ & (1-b_{\hat{H}\tilde{L}})[p_E + w_E Q_E(\lambda_E(0,0,1,1))] \leq p_G + w_G Q_G, \end{aligned}$$

since we have $\lambda_E(f^e_{\hat{L}\tilde{L}}, f^e_{\hat{L}\tilde{H}}, 1, 1) \geq \lambda_E(0,0,1,1)$. Hence, we have $C_{\hat{H}\tilde{L},E}(0,0,1,1) \leq C_{\hat{H}\tilde{L},G}(0,0,1,1)$, i.e., $(1-b_{\hat{H}\tilde{L}})[p_E + w_E Q_E(b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}})] \leq p_G + w_G Q_G$. This implies $(1-b_{\hat{H}})[p_E + w_E Q_E(b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}})] \leq p_G + w_G Q_G$, i.e., $C_{\hat{H},E}(0,1) \leq C_{\hat{H},G}(0,1)$, since we have $b_{\hat{H}} \geq b_{\hat{H}\tilde{L}}$. Hence, we have $C_{\hat{H},E}(0,1) \leq C_{\hat{H},G}(0,1)$, which contradicts (EC.17). This proves that $f^e_{\hat{H}\tilde{L}} < 1$.

Hence, out of the 12 different equilibrium regions of (EC.13) in the presence of virtual triage, we have $\tilde{\mathbf{f}}^{\mathbf{e}} \in \{(0,0,0,m), (0,0,0,1), (0,0,m,1), (0,m,0,1), (0,1,0,1), (0,1,m,1)\}$ when $\hat{\mathbf{f}}^{\mathbf{e}} = (0,m)$. The

characterization of the six equilibrium regions in the presence of virtual triage follows the same argument as part (i).

(iii) When $\hat{\mathbf{f}}^{\mathbf{e}} = (0,1)$, $\hat{L}$ patients go to a GP first with certainty in equilibrium, as visiting the ED directly with a positive probability does not lead to lower patient costs for them; meanwhile, $\hat{H}$ patients visit the ED directly with certainty in equilibrium, as going to a GP first with a positive probability does not lead to lower patient costs for them. Hence, we have

$$\begin{aligned} & C_{\hat{H},E}(0,1) \le C_{\hat{H},G}(0,1) \wedge C_{\hat{L},E}(0,1) \ge C_{\hat{L},G}(0,1) \\ \Leftrightarrow\ & (1-b_{\hat{H}})[p_E + w_E Q_E(b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}})] \le p_G + w_G Q_G \wedge \\ & (1-b_{\hat{L}})[p_E + w_E Q_E(b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}})] \ge p_G + w_G Q_G. \end{aligned} \tag{EC.18}$$

In this case, in the presence of virtual triage, $\forall\ \tilde{\alpha} + \tilde{\beta} \le 1$, we can show that $f^e_{\hat{L}\tilde{L}} = 0$ with proof by contradiction. Let $\tilde{\mathbf{f}}^{\mathbf{e}} = (f^e_{\hat{L}\tilde{L}}, f^e_{\hat{L}\tilde{H}}, f^e_{\hat{H}\tilde{L}}, f^e_{\hat{H}\tilde{H}})$ denote the equilibrium patient flow in the presence of virtual triage. If $f^e_{\hat{L}\tilde{L}} > 0$, we have $\hat{L}\tilde{L}$ patients find lower patient costs by visiting the ED directly under $\tilde{\mathbf{f}} = (0, f^e_{\hat{L}\tilde{H}}, f^e_{\hat{H}\tilde{L}}, f^e_{\hat{H}\tilde{H}})$. Hence, we have

$$\begin{aligned} & C_{\hat{L}\tilde{L},E}(0, f^e_{\hat{L}\tilde{H}}, f^e_{\hat{H}\tilde{L}}, f^e_{\hat{H}\tilde{H}}) < C_{\hat{L}\tilde{L},G}(0, f^e_{\hat{L}\tilde{H}}, f^e_{\hat{H}\tilde{L}}, f^e_{\hat{H}\tilde{H}}) \\ \Leftrightarrow\ & (1-b_{\hat{L}\tilde{L}})[p_E + w_E Q_E(\lambda_E(0, f^e_{\hat{L}\tilde{H}}, f^e_{\hat{H}\tilde{L}}, f^e_{\hat{H}\tilde{H}}))] < p_G + w_G Q_G \\ \Rightarrow\ & (1-b_{\hat{L}\tilde{L}})[p_E + w_E Q_E(\lambda_E(0,0,1,1))] < p_G + w_G Q_G, \end{aligned}$$

since we have $\lambda_E(0, f^e_{\hat{L}\tilde{H}}, f^e_{\hat{H}\tilde{L}}, f^e_{\hat{H}\tilde{H}}) > \lambda_E(0,0,1,1)$: if $f^e_{\hat{L}\tilde{L}} > 0$, we have $f^e_{\hat{L}\tilde{H}} = f^e_{\hat{H}\tilde{L}} = f^e_{\hat{H}\tilde{H}} = 1$ by Lemma EC.1, which implies $\lambda_E(0, f^e_{\hat{L}\tilde{H}}, f^e_{\hat{H}\tilde{L}}, f^e_{\hat{H}\tilde{H}}) > \lambda_E(0,0,1,1)$. Hence, we have $C_{\hat{L}\tilde{L},E}(0,0,1,1) < C_{\hat{L}\tilde{L},G}(0,0,1,1)$, i.e., $(1-b_{\hat{L}\tilde{L}})[p_E + w_E Q_E(b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}})] < p_G + w_G Q_G$. This implies $(1-b_{\hat{L}})[p_E + w_E Q_E(b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}})] < p_G + w_G Q_G$, i.e., $C_{\hat{L},E}(0,1) < C_{\hat{L},G}(0,1)$, since we have $b_{\hat{L}} \ge b_{\hat{L}\tilde{L}}$. Hence, we have $C_{\hat{L},E}(0,1) < C_{\hat{L},G}(0,1)$, which contradicts (EC.18). This proves that $f^e_{\hat{L}\tilde{L}} = 0$.

Moreover, for $\forall\ \tilde{\alpha} + \tilde{\beta} \le 1$, we can show that $f^e_{\hat{H}\tilde{H}} = 1$ with proof by contradiction. If $f^e_{\hat{H}\tilde{H}} < 1$, we have $\hat{H}\tilde{H}$ patients find lower patient costs by going to a GP first under $\tilde{\mathbf{f}} = (f^e_{\hat{L}\tilde{L}}, f^e_{\hat{L}\tilde{H}}, f^e_{\hat{H}\tilde{L}}, 1)$. Hence, we have

$$\begin{aligned} & C_{\hat{H}\tilde{H},E}(f^e_{\hat{L}\tilde{L}}, f^e_{\hat{L}\tilde{H}}, f^e_{\hat{H}\tilde{L}}, 1) > C_{\hat{H}\tilde{H},G}(f^e_{\hat{L}\tilde{L}}, f^e_{\hat{L}\tilde{H}}, f^e_{\hat{H}\tilde{L}}, 1) \\ \Leftrightarrow\ & (1-b_{\hat{H}\tilde{H}})[p_E + w_E Q_E(\lambda_E(f^e_{\hat{L}\tilde{L}}, f^e_{\hat{L}\tilde{H}}, f^e_{\hat{H}\tilde{L}}, 1))] > p_G + w_G Q_G \\ \Rightarrow\ & (1-b_{\hat{H}\tilde{H}})[p_E + w_E Q_E(\lambda_E(0,0,1,1))] > p_G + w_G Q_G, \end{aligned}$$

since we have $\lambda_E(f^e_{\hat{L}\tilde{L}}, f^e_{\hat{L}\tilde{H}}, f^e_{\hat{H}\tilde{L}}, 1) < \lambda_E(0,0,1,1)$: if $f^e_{\hat{H}\tilde{H}} < 1$, we have $f^e_{\hat{L}\tilde{L}} = f^e_{\hat{L}\tilde{H}} = f^e_{\hat{H}\tilde{L}} = 0$ by Lemma EC.1, which implies $\lambda_E(f^e_{\hat{L}\tilde{L}}, f^e_{\hat{L}\tilde{H}}, f^e_{\hat{H}\tilde{L}}, 1) < \lambda_E(0,0,1,1)$. Hence, we have $C_{\hat{H}\tilde{H},E}(0,0,1,1) > C_{\hat{H}\tilde{H},G}(0,0,1,1)$, i.e., $(1-b_{\hat{H}\tilde{H}})[p_E + w_E Q_E(b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}})] > p_G + w_G Q_G$. This implies $(1-b_{\hat{H}})[p_E +$

$w_E Q_E(b_{\hat{L}}\lambda_{\hat{L}}+\lambda_{\hat{H}})] > p_G + w_G Q_G$, i.e., $C_{\hat{H},E}(0,1) > C_{\hat{H},G}(0,1)$, since we have $b_{\hat{H}} \le b_{\hat{H}\tilde{H}}$. Hence, we have $C_{\hat{H},E}(0,1) > C_{\hat{H},G}(0,1)$, which contradicts (EC.18). This proves that $f^e_{\hat{H}\tilde{H}} = 1$.

Hence, out of the 12 different equilibrium regions of (EC.13) in the presence of virtual triage, we have $\tilde{\mathbf{f}}^{\mathbf{e}} \in \{(0,0,0,1),(0,0,m,1),(0,0,1,1),(0,m,1,1),(0,1,1,1),(0,1,m,1),(0,1,0,1),(0,m,0,1)\}$ when $\hat{\mathbf{f}}^{\mathbf{e}} = (0,1)$. The characterization of the eight equilibrium regions in the presence of virtual triage follows the same argument as part (i).

(iv) When $\hat{\mathbf{f}}^{\mathbf{e}} = (m,1)$, we can show that $f^e_{\hat{H}\tilde{H}} = 1$, $f^e_{\hat{L}\tilde{L}} < 1$ and $f^e_{\hat{L}\tilde{H}} > 0$ following the same argument as part (ii). As a result, out of the 12 different equilibrium regions of (EC.13) in the presence of virtual triage, we have $\tilde{\mathbf{f}}^{\mathbf{e}} \in \{(m,1,1,1),(0,1,1,1),(0,m,1,1),\ (0,1,m,1),(0,1,0,1),(0,m,0,1)\}$ when $\hat{\mathbf{f}}^{\mathbf{e}} = (m,1)$. The characterization of the six equilibrium regions in the presence of virtual triage follows the same argument as part (i).

(v) When $\hat{\mathbf{f}}^{\mathbf{e}} = (1,1)$, we can show that $f^e_{\hat{L}\tilde{H}} = f^e_{\hat{H}\tilde{H}} = 1$ following the same argument as part (i). As a result, out of the 12 different equilibrium regions of (EC.13) in the presence of virtual triage, we have $\tilde{\mathbf{f}}^{\mathbf{e}} \in \{(1,1,1,1),(m,1,1,1),(0,1,1,1),(0,1,m,1),(0,1,0,1)\}$ when $\hat{\mathbf{f}}^{\mathbf{e}} = (1,1)$. The characterization of the five equilibrium regions in the presence of virtual triage follows the same argument as part (i). □

Proof of Proposition 2. For a given virtual triage tool with IROC curve $\tilde{\alpha} = r(\tilde{\beta})$, we have $b_{\hat{T}\tilde{L}} = \frac{r(\tilde{\beta})b_{\hat{T}}}{r(\tilde{\beta})b_{\hat{T}}+(1-\tilde{\beta})(1-b_{\hat{T}})}$ and $b_{\hat{T}\tilde{H}} = \frac{(1-r(\tilde{\beta}))b_{\hat{T}}}{(1-r(\tilde{\beta}))b_{\hat{T}}+\tilde{\beta}(1-b_{\hat{T}})}$, $\hat{T} \in \{\hat{L},\hat{H}\}$.

(i) We have $\lim_{\bar{s}\to 0^+} b_{\hat{T}\tilde{L}} = \lim_{\tilde{\beta}\to 1^-} \frac{r'(\tilde{\beta})b_{\hat{T}}}{r'(\tilde{\beta})b_{\hat{T}}-(1-b_{\hat{T}})}$ by L'Hopital's rule. Recall $r(\tilde{\beta})$ is decreasing and convex in $\tilde{\beta}, \tilde{\beta} \in [0,1]$. If $\lim_{\tilde{\beta}\to 1^-} r'(\tilde{\beta})$ is sufficiently large, i.e., close to 0, we then have $\lim_{\tilde{\beta}\to 1^-} b_{\hat{T}\tilde{L}}$ be close to 0 and therefore $\tilde{f}^e_{\hat{T}\tilde{L}} = 0$. On the other hand, we have $\lim_{\bar{s}\to 0^+} b_{\hat{T}\tilde{H}} = \lim_{\tilde{\beta}\to 1^-} \frac{(1-r(\tilde{\beta}))b_{\hat{T}}}{(1-r(\tilde{\beta}))b_{\hat{T}}+\tilde{\beta}(1-b_{\hat{T}})}$ be close to $b_{\hat{T}}$.

When $\hat{\mathbf{f}}^{\mathbf{e}} = (0,1)$, with $\bar{s} \to 0^+$, this implies that $\tilde{f}^e_{\hat{T}\tilde{L}} = 0$ and $\tilde{f}^e_{\hat{T}\tilde{H}} = \hat{f}^e_{\hat{T}}$. Hence, we have $\tilde{\mathbf{f}}^{\mathbf{e}} = (0,0,0,1)$, and therefore $\lambda_G(\hat{\mathbf{f}}^{\mathbf{e}}) < \lambda_G(\tilde{\mathbf{f}}^{\mathbf{e}})$ and $\lambda_E(\hat{\mathbf{f}}^{\mathbf{e}}) > \lambda_E(\tilde{\mathbf{f}}^{\mathbf{e}})$.

When $\hat{\mathbf{f}}^{\mathbf{e}} = (1,1)$, with $\bar{s} \to 0^+$, this implies that $\tilde{f}^e_{\hat{T}\tilde{L}} = 0$ and $\tilde{f}^e_{\hat{T}\tilde{H}} = \hat{f}^e_{\hat{T}}$. Hence, we have $\tilde{\mathbf{f}}^{\mathbf{e}} = (0,1,0,1)$, and therefore $\lambda_G(\hat{\mathbf{f}}^{\mathbf{e}}) < \lambda_G(\tilde{\mathbf{f}}^{\mathbf{e}})$ and $\lambda_E(\hat{\mathbf{f}}^{\mathbf{e}}) > \lambda_E(\tilde{\mathbf{f}}^{\mathbf{e}})$.

Hence, when $\hat{\mathbf{f}}^{\mathbf{e}} \in \{(0,1),(1,1)\}$, there exists virtual triage such that $\exists\ \bar{s}_u \in (0,1)$ s.t. $\forall\ \bar{s} \in (0,\bar{s}_u)$, we have $\lambda_G(\hat{\mathbf{f}}^{\mathbf{e}}) < \lambda_G(\tilde{\mathbf{f}}^{\mathbf{e}})$ and $\lambda_E(\hat{\mathbf{f}}^{\mathbf{e}}) > \lambda_E(\tilde{\mathbf{f}}^{\mathbf{e}})$.

(ii) We have $\lim_{\bar{s}\to 1^-} b_{\hat{T}\tilde{L}} = \lim_{\tilde{\beta}\to 0^+} \frac{r(\tilde{\beta})b_{\hat{T}}}{r(\tilde{\beta})b_{\hat{T}}+(1-\tilde{\beta})(1-b_{\hat{T}})}$ be close to $b_{\hat{T}}$. On the other hand, we have $\lim_{\bar{s}\to 1^-} b_{\hat{T}\tilde{H}} = \lim_{\tilde{\beta}\to 0^+} \frac{-r'(\tilde{\beta})b_{\hat{T}}}{-r'(\tilde{\beta})b_{\hat{T}}+(1-b_{\hat{T}})}$ by L'Hopital's rule. If $\lim_{\tilde{\beta}\to 0^+} r'(\tilde{\beta})$ is sufficiently small, we then have $\lim_{\tilde{\beta}\to 0^+} b_{\hat{T}\tilde{H}}$ be close to 1 and therefore $\tilde{f}^e_{\hat{T}\tilde{H}} = 1$.

When $\hat{\mathbf{f}}^{\mathbf{e}} = (0,0)$, with $\bar{s} \to 1^-$, this implies that $\tilde{f}^e_{\hat{T}\tilde{L}} = \hat{f}^e_{\hat{T}}$ and $\tilde{f}^e_{\hat{T}\tilde{H}} = 1$. Hence, we have $\tilde{\mathbf{f}}^{\mathbf{e}} = (0,1,0,1)$, and therefore $\lambda_G(\hat{\mathbf{f}}^{\mathbf{e}}) > \lambda_G(\tilde{\mathbf{f}}^{\mathbf{e}})$ and $\lambda_E(\hat{\mathbf{f}}^{\mathbf{e}}) < \lambda_E(\tilde{\mathbf{f}}^{\mathbf{e}})$.

When $\hat{\mathbf{f}}^{\mathbf{e}} = (0,1)$, with $\bar{s} \to 1^-$, this implies that $\tilde{f}^e_{\hat{T}\tilde{L}} = \hat{f}^e_{\hat{T}}$ and $\tilde{f}^e_{\hat{T}\tilde{H}} = 1$. Hence, we have $\tilde{\mathbf{f}}^{\mathbf{e}} = (0,1,1,1)$, and therefore $\lambda_G(\hat{\mathbf{f}}^{\mathbf{e}}) > \lambda_G(\tilde{\mathbf{f}}^{\mathbf{e}})$ and $\lambda_E(\hat{\mathbf{f}}^{\mathbf{e}}) < \lambda_E(\tilde{\mathbf{f}}^{\mathbf{e}})$.

Hence, when $\hat{\mathbf{f}}^{\mathbf{e}} \in \{(0,0),(0,1)\}$, there exists virtual triage such that $\exists\ \bar{s}_l \in (0,1)$ s.t. $\forall\ \bar{s} \in (\bar{s}_l, 1)$, we have $\lambda_G(\hat{\mathbf{f}}^{\mathbf{e}}) > \lambda_G(\tilde{\mathbf{f}}^{\mathbf{e}})$ and $\lambda_E(\hat{\mathbf{f}}^{\mathbf{e}}) < \lambda_E(\tilde{\mathbf{f}}^{\mathbf{e}})$. $\square$

EC.4.3. Proofs for Section 5

In this subsection, we provide the proofs for the analysis of the impact of off-the-shelf virtual triage on system performance. In particular, we have an exogenous virtual triage accuracy $\tilde{\alpha}$ and $\tilde{\beta}$, and we analyze the decomposed effect of lower $\tilde{\alpha}$ or lower $\tilde{\beta}$ while keeping the other one constant.

Proof of Lemma 3. (i) When $\tilde{\mathbf{f}}^{\mathbf{e}} \in R_{p,\sim}$, the adoption of virtual triage does not change patient care-seeking behavior. Hence we have $\frac{dC_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\alpha}} = \frac{dC_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\beta}} = 0$.

(ii) In equilibrium region $\tilde{\mathbf{f}}^{\mathbf{e}} = (0,1,1,1)$, we have $\lambda_G(0,1,1,1) = \lambda_{\hat{L}\tilde{L}} = [\tilde{\alpha} b_{\hat{L}} + (1-\tilde{\beta})(1-b_{\hat{L}})]\lambda_{\hat{L}}, \lambda_E(0,1,1,1) = b_{\hat{L}\tilde{L}}\lambda_{\hat{L}\tilde{L}} + \lambda_{\hat{L}\tilde{H}} + \lambda_{\hat{H}} = [b_{\hat{L}} + \tilde{\beta}(1-b_{\hat{L}})]\lambda_{\hat{L}} + \lambda_{\hat{H}}$. We then have $\frac{\partial \lambda_G(0,1,1,1)}{\partial \tilde{\alpha}} = b_{\hat{L}}\lambda_{\hat{L}}, \frac{\partial \lambda_G(0,1,1,1)}{\partial \tilde{\beta}} = -(1-b_{\hat{L}})\lambda_{\hat{L}}, \frac{\partial \lambda_E(0,1,1,1)}{\partial \tilde{\alpha}} = 0, \frac{\partial \lambda_E(0,1,1,1)}{\partial \tilde{\beta}} = (1-b_{\hat{L}})\lambda_{\hat{L}}$.

In equilibrium region $\tilde{\mathbf{f}}^{\mathbf{e}} = (0,0,0,1)$, we have $\lambda_G(0,0,0,1) = \lambda_{\hat{L}} + \lambda_{\hat{H}\tilde{L}} = \lambda_{\hat{L}} + [\tilde{\alpha} b_{\hat{H}} + (1-\tilde{\beta})(1-b_{\hat{H}})]\lambda_{\hat{H}}, \lambda_E(0,0,0,1) = b_{\hat{L}}\lambda_{\hat{L}} + b_{\hat{H}\tilde{L}}\lambda_{\hat{H}\tilde{L}} + \lambda_{\hat{H}\tilde{H}} = b_{\hat{L}}\lambda_{\hat{L}} + b_{\hat{H}}\lambda_{\hat{H}} + \tilde{\beta}(1-b_{\hat{H}})\lambda_{\hat{H}}$. We then have $\frac{\partial \lambda_G(0,0,0,1)}{\partial \tilde{\alpha}} = b_{\hat{H}}\lambda_{\hat{H}}, \frac{\partial \lambda_G(0,0,0,1)}{\partial \tilde{\beta}} = -(1-b_{\hat{H}})\lambda_{\hat{H}}, \frac{\partial \lambda_E(0,0,0,1)}{\partial \tilde{\alpha}} = 0, \frac{\partial \lambda_E(0,0,0,1)}{\partial \tilde{\beta}} = (1-b_{\hat{H}})\lambda_{\hat{H}}$.

In equilibrium region $\tilde{\mathbf{f}}^{\mathbf{e}} = (0,1,0,1)$, we have $\lambda_G(0,1,0,1) = \lambda_{\hat{L}\tilde{L}} + \lambda_{\hat{H}\tilde{L}} = \tilde{\alpha}(b_{\hat{L}}\lambda_{\hat{L}} + b_{\hat{H}}\lambda_{\hat{H}}) + (1-\tilde{\beta})[(1-b_{\hat{L}})\lambda_{\hat{L}} + (1-b_{\hat{H}})\lambda_{\hat{H}}], \lambda_E(0,1,0,1) = b_{\hat{L}\tilde{L}}\lambda_{\hat{L}\tilde{L}} + b_{\hat{H}\tilde{L}}\lambda_{\hat{H}\tilde{L}} + \lambda_{\hat{L}\tilde{H}} + \lambda_{\hat{H}\tilde{H}} = b_{\hat{L}}\lambda_{\hat{L}} + b_{\hat{H}}\lambda_{\hat{H}} + \tilde{\beta}[(1-b_{\hat{L}})\lambda_{\hat{L}} + (1-b_{\hat{H}})\lambda_{\hat{H}}]$. We then have $\frac{\partial \lambda_G(0,1,0,1)}{\partial \tilde{\alpha}} = b_{\hat{L}}\lambda_{\hat{L}} + b_{\hat{H}}\lambda_{\hat{H}}, \frac{\partial \lambda_G(0,1,0,1)}{\partial \tilde{\beta}} = -[(1-b_{\hat{L}})\lambda_{\hat{L}} + (1-b_{\hat{H}})\lambda_{\hat{H}}], \frac{\partial \lambda_E(0,1,0,1)}{\partial \tilde{\alpha}} = 0, \frac{\partial \lambda_E(0,1,0,1)}{\partial \tilde{\beta}} = (1-b_{\hat{L}})\lambda_{\hat{L}} + (1-b_{\hat{H}})\lambda_{\hat{H}}$.

Hence, when $\tilde{\mathbf{f}}^{\mathbf{e}} \in R_{p,\sim}$, we have $\frac{\partial \lambda_G(0,1,0,1)}{\partial \tilde{\alpha}} > 0, \frac{\partial \lambda_E(0,1,0,1)}{\partial \tilde{\alpha}} = 0$, and therefore $\frac{dC_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\alpha}} > 0$. On the other hand, we have $\frac{\partial \lambda_G(0,1,0,1)}{\partial \tilde{\beta}} < 0, \frac{\partial \lambda_E(0,1,0,1)}{\partial \tilde{\beta}} > 0$, and $\frac{\partial \lambda_G(0,1,0,1)}{\partial \tilde{\beta}} + \frac{\partial \lambda_E(0,1,0,1)}{\partial \tilde{\beta}} = 0$. Since an arrival to an ED is more costly than an arrival to a GP by assumption, we have $\frac{dC_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\beta}} > 0$. $\square$

Proof of Lemma 4. (i) In equilibrium region $\tilde{\mathbf{f}}^{\mathbf{e}} = (0,m,1,1)$, for a given $\tilde{\alpha}$ and $\tilde{\beta}$ s.t. we have $f^e_{\hat{L}\tilde{H}} \in (0,1)$, $f^e_{\hat{L}\tilde{H}}$ is determined by solving the following problem:

$$\min_{0<f_{\hat{L}\tilde{H}}<1} \Phi(f_{\hat{L}\tilde{H}}) = \int_0^{\lambda_G} w_G Q_G dx + \int_0^{\lambda_E} w_E Q_E(x) dx + \int_0^{\lambda_G} \lambda_G p_G dx + \int_0^{\lambda_E} \lambda_E p_E dx, \qquad \text{(EC.19)}$$

where $f^e_{\hat{L}\tilde{H}}$ is given by the following FOC of Problem (EC.19):

$$\frac{\partial \Phi(f_{\hat{L}\tilde{H}})}{\partial f_{\hat{L}\tilde{H}}} = -[(1-\tilde{\alpha})b_{\hat{L}} + \tilde{\beta}(1-b_{\hat{L}})]\lambda_{\hat{L}}(p_G + w_G Q_G) + \tilde{\beta}(1-b_{\hat{L}})\lambda_{\hat{L}}[p_E + w_E Q_E(\lambda_E)] = 0.$$

We omit the dependence of λ_G and λ_E on $\tilde{\mathbf{f}}^{\mathbf{e}}$ and $\tilde{\alpha}$ and $\tilde{\beta}$ here to simplify the exposition. When $\tilde{\mathbf{f}}^{\mathbf{e}} = (0,m,1,1)$, we have

$$\lambda_G = \lambda_{\hat{L}\tilde{L}} + (1-f^e_{\hat{L}\tilde{H}})\lambda_{\hat{L}\tilde{H}} = \lambda_{\hat{L}} - f^e_{\hat{L}\tilde{H}}[(1-\tilde{\alpha})b_{\hat{L}} + \tilde{\beta}(1-b_{\hat{L}})]\lambda_{\hat{L}},$$
$$\lambda_E = b_{\hat{L}\tilde{L}}\lambda_{\hat{L}\tilde{L}} + (1-f^e_{\hat{L}\tilde{H}})b_{\hat{L}\tilde{H}}\lambda_{\hat{L}\tilde{H}} + f^e_{\hat{L}\tilde{H}}\lambda_{\hat{L}\tilde{H}} + \lambda_{\hat{H}} = b_{\hat{L}}\lambda_{\hat{L}} + f^e_{\hat{L}\tilde{H}}\tilde{\beta}(1-b_{\hat{L}})\lambda_{\hat{L}} + \lambda_{\hat{H}}.$$

We also have

$$\left.\frac{\partial^2\Phi(f_{\hat{L}\tilde{H}})}{\partial f_{\hat{L}\tilde{H}}^2}\right|_{f_{\hat{L}\tilde{H}}=f_{\hat{L}\tilde{H}}^e} = [\tilde{\beta}(1-b_{\hat{L}})\lambda_{\hat{L}}]^2 w_E \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E} > 0, \tag{EC.20}$$

and

$$\left.\frac{\partial^2\Phi(f_{\hat{L}\tilde{H}})}{\partial f_{\hat{L}\tilde{H}}\partial\tilde{\alpha}}\right|_{f_{\hat{L}\tilde{H}}=f_{\hat{L}\tilde{H}}^e} = b_{\hat{L}}\lambda_{\hat{L}}(p_G + w_G Q_G) > 0, \tag{EC.21}$$

and we have

$$\begin{aligned}\left.\frac{\partial^2\Phi(f_{\hat{L}\tilde{H}})}{\partial f_{\hat{L}\tilde{H}}\partial\tilde{\beta}}\right|_{f_{\hat{L}\tilde{H}}=f_{\hat{L}\tilde{H}}^e} &= (1-b_{\hat{L}})\lambda_{\hat{L}}[-p_G - w_G Q_G + p_E + w_E Q_E(\lambda_E) + \tilde{\beta} f_{\hat{L}\tilde{H}}^e (1-b_{\hat{L}})\lambda_{\hat{L}} w_E \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}] \\ &> (1-b_{\hat{L}})\lambda_{\hat{L}}[-p_G - w_G Q_G + p_E + w_E Q_E(\lambda_E)] \\ &> (1-b_{\hat{L}})\lambda_{\hat{L}}[-p_G - w_G Q_G + p_E + w_E Q_E(\lambda_H)] \\ &> 0,\end{aligned} \tag{EC.22}$$

where the last inequality comes from the assumption on p_E in Section 3.3.

We have the equilibrium social cost:

$$C_s(\tilde{\mathbf{f}}^{\mathbf{e}}) = \lambda_G w_G Q_G + \lambda_E w_E Q_E(\lambda_E) + \lambda_G a_G + \lambda_E a_E,$$

where

$$\frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f_{\hat{L}\tilde{H}}^e} = -[(1-\tilde{\alpha})b_{\hat{L}} + \tilde{\beta}(1-b_{\hat{L}})]\lambda_{\hat{L}}(a_G + w_G Q_G) + \tilde{\beta}(1-b_{\hat{L}})\lambda_{\hat{L}}[a_E + w_E Q_E(\lambda_E) + \lambda_E w_E \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}].$$

(a) Moreover, we have

$$\begin{aligned}\frac{dC_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\alpha}} &= \frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial \tilde{\alpha}} + \frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f_{\hat{L}\tilde{H}}^e}\frac{\partial f_{\hat{L}\tilde{H}}^e}{\partial \tilde{\alpha}} \\ &= [\frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial \tilde{\alpha}}\frac{\partial^2\Phi(f_{\hat{L}\tilde{H}}^e)}{\partial f_{\hat{L}\tilde{H}}^{e2}} - \frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f_{\hat{L}\tilde{H}}^e}\frac{\partial^2\Phi(f_{\hat{L}\tilde{H}}^e)}{\partial f_{\hat{L}\tilde{H}}^e \partial\tilde{\alpha}}]/\frac{\partial^2\Phi(f_{\hat{L}\tilde{H}}^e)}{\partial f_{\hat{L}\tilde{H}}^{e2}},\end{aligned}$$

where

$$\frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial \tilde{\alpha}} = f_{\hat{L}\tilde{H}}^e b_{\hat{L}}\lambda_{\hat{L}}(a_G + w_G Q_G) > 0.$$

We then have

$$\begin{aligned}&\frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial \tilde{\alpha}}\frac{\partial^2\Phi(f_{\hat{L}\tilde{H}}^e)}{\partial f_{\hat{L}\tilde{H}}^{e2}} - \frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f_{\hat{L}\tilde{H}}^e}\frac{\partial^2\Phi(f_{\hat{L}\tilde{H}}^e)}{\partial f_{\hat{L}\tilde{H}}^e\partial\tilde{\alpha}} \\ &= f_{\hat{L}\tilde{H}}^e b_{\hat{L}}\lambda_{\hat{L}}(a_G + w_G Q_G)[\tilde{\beta}(1-b_{\hat{L}})\lambda_{\hat{L}}]^2 w_E \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E} - \{\tilde{\beta}(1-b_{\hat{L}})\lambda_{\hat{L}}[a_E + w_E Q_E(\lambda_E) + \lambda_E w_E \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}] \\ &- [(1-\tilde{\alpha})b_{\hat{L}} + \tilde{\beta}(1-b_{\hat{L}})]\lambda_{\hat{L}}(a_G + w_G Q_G)\} b_{\hat{L}}\lambda_{\hat{L}}(p_G + w_G Q_G).\end{aligned}$$

As $\frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial \tilde{\alpha}}\frac{\partial^2\Phi(f_{\hat{L}\tilde{H}}^e)}{\partial f_{\hat{L}\tilde{H}}^{e2}} - \frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f_{\hat{L}\tilde{H}}^e}\frac{\partial^2\Phi(f_{\hat{L}\tilde{H}}^e)}{\partial f_{\hat{L}\tilde{H}}^e\partial\tilde{\alpha}}$ is linear and decreasing in a_E, $\exists\ a_E^u$ s.t. we have $\frac{dC_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\alpha}} < 0$ if and only if $a_E > a_E^u$.

(b) Meanwhile, we have

$$\frac{dC_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\beta}} = \frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial \tilde{\beta}} + \frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{L}\tilde{H}}}\frac{\partial f^e_{\hat{L}\tilde{H}}}{\partial \tilde{\beta}}$$
$$= [\frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial \tilde{\beta}}\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^{e2}_{\hat{L}\tilde{H}}} - \frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{L}\tilde{H}}}\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^e_{\hat{L}\tilde{H}}\partial \tilde{\beta}}]/\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^{e2}_{\hat{L}\tilde{H}}},$$

where

$$\frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial \beta} = f^e_{\hat{L}\tilde{H}}(1-b_{\hat{L}})\lambda_{\hat{L}}[a_E + w_E Q_E(\lambda_E) + \lambda_E w \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E} - a_G - w_G Q_G] > 0,$$

given the assumption on a_E in Section 3.3. We then have

$$\frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial \tilde{\beta}}\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^{e2}_{\hat{L}\tilde{H}}} - \frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{L}\tilde{H}}}\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^e_{\hat{L}\tilde{H}}\partial \tilde{\beta}}$$
$$= f^e_{\hat{L}\tilde{H}}(1-b_{\hat{L}})\lambda_{\hat{L}}[a_E + w_E Q_E(\lambda_E) + \lambda_E w_E \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E} - a_G - w_G Q_G] \cdot [\tilde{\beta}(1-b_{\hat{L}})\lambda_{\hat{L}}]^2 w_E \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}$$
$$- \{\tilde{\beta}(1-b_{\hat{L}})\lambda_{\hat{L}}[a_E + w_E Q_E(\lambda_E) + \lambda_E w_E \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}] - [(1-\tilde{\alpha})b_{\hat{L}} + \tilde{\beta}(1-b_{\hat{L}})]\lambda_{\hat{L}}(a_G + w_G Q_G)\}$$
$$\cdot (1-b_{\hat{L}})\lambda_{\hat{L}}[-p_G - w_G Q_G + p_E + w_E Q_E(\lambda_E) + \tilde{\beta} f^e_{\hat{L}\tilde{H}}(1-b_{\hat{L}})\lambda_{\hat{L}} w_E \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}].$$

As $\frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial \tilde{\beta}}\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^{e2}_{\hat{L}\tilde{H}}} - \frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{L}\tilde{H}}}\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^e_{\hat{L}\tilde{H}}\partial \tilde{\beta}}$ is linear and decreasing in a_E, $\exists\ a^o_E$ s.t. we have $\frac{dC_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\beta}} < 0$ if and only if $a_E > a^o_E$.

The proof for the results with $\tilde{\mathbf{f}}^{\mathbf{e}} \in \{(0,m,0,1),(0,0,0,m)\}$ follows the same procedures.

(ii) In equilibrium region $\tilde{\mathbf{f}}^{\mathbf{e}} = (0,0,m,1)$, for a given $\tilde{\alpha}$ and $\tilde{\beta}$ s.t. we have $f^e_{\hat{H}\tilde{L}} \in (0,1)$, $f^e_{\hat{H}\tilde{L}}$ is determined by solving the following problem:

$$\min_{0<f_{\hat{H}\tilde{L}}<1} \Phi(f_{\hat{H}\tilde{L}}) = \int_0^{\lambda_G} w_G Q_G dx + \int_0^{\lambda_E} w_E Q_E(x)dx + \int_0^{\lambda_G} \lambda_G p_G dx + \int_0^{\lambda_E} \lambda_E p_E dx, \qquad \text{(EC.23)}$$

where $f^e_{\hat{H}\tilde{L}}$ is given by the following FOC of Problem (EC.23):

$$\frac{\partial \Phi(f_{\hat{H}\tilde{L}})}{\partial f_{\hat{H}\tilde{L}}} = -[\tilde{\alpha}b_{\hat{H}} + (1-\tilde{\beta})(1-b_{\hat{H}})]\lambda_{\hat{H}}(p_G + w_G Q_G) + (1-\tilde{\beta})(1-b_{\hat{H}})\lambda_{\hat{H}}[p_E + w_E Q_E(\lambda_E)] = 0.$$

When $\tilde{\mathbf{f}}^{\mathbf{e}} = (0,0,m,1)$, we have

$$\lambda_G = \lambda_{\hat{L}} + (1-f^e_{\hat{H}\tilde{L}})\lambda_{\hat{H}\tilde{L}} = \lambda_{\hat{L}} + (1-f^e_{\hat{H}\tilde{L}})[\tilde{\alpha}b_{\hat{H}} + (1-\tilde{\beta})(1-b_{\hat{H}})]\lambda_{\hat{H}},$$
$$\lambda_E = b_{\hat{L}}\lambda_{\hat{L}} + (1-f^e_{\hat{H}\tilde{L}})b_{\hat{H}\tilde{L}}\lambda_{\hat{H}\tilde{L}} + f^e_{\hat{H}\tilde{L}}\lambda_{\hat{H}\tilde{L}} + \lambda_{\hat{H}\tilde{H}} = \lambda_H + f^e_{\hat{H}\tilde{L}}(1-\tilde{\beta})(1-b_{\hat{H}})\lambda_{\hat{H}} + \tilde{\beta}(1-b_{\hat{H}})\lambda_{\hat{H}}.$$

We also have

$$\frac{\partial^2 \Phi(f_{\hat{H}\tilde{L}})}{\partial f^2_{\hat{H}\tilde{L}}}\bigg|_{f_{\hat{H}\tilde{L}}=f^e_{\hat{H}\tilde{L}}} = [(1-\tilde{\beta})(1-b_{\hat{H}})\lambda_{\hat{H}}]^2 w_E \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E} > 0,$$

and

$$\frac{\partial^2\Phi(f_{\hat{H}\tilde{L}})}{\partial f_{\hat{H}\tilde{L}}\partial\tilde{\alpha}}\bigg|_{f_{\hat{H}\tilde{L}}=f^e_{\hat{H}\tilde{L}}} = -b_{\hat{H}}\lambda_{\hat{H}}(p_G+w_GQ_G)<0,$$

and we have

$$\frac{\partial^2\Phi(f_{\hat{H}\tilde{L}})}{\partial f_{\hat{H}\tilde{L}}\partial\tilde{\beta}}\bigg|_{f_{\hat{H}\tilde{L}}=f^e_{\hat{H}\tilde{L}}} = (1-b_{\hat{H}})\lambda_{\hat{H}}[p_G+w_GQ_G-p_E-w_EQ_E(\lambda_E)+(1-\tilde{\beta})(1-f^e_{\hat{H}\tilde{L}})(1-b_{\hat{H}})\lambda_{\hat{H}}w_E\frac{\partial Q_E(\lambda_E)}{\partial\lambda_E}].$$

We have the equilibrium social cost:

$$C_s(\tilde{\mathbf{f}}^{\mathbf{e}}) = \lambda_Gw_GQ_G+\lambda_Ew_EQ_E(\lambda_E)+\lambda_Ga_G+\lambda_Ea_E,$$

where

$$\frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{H}\tilde{L}}} = -[\tilde{\alpha}b_{\hat{H}}+(1-\tilde{\beta})(1-b_{\hat{H}})]\lambda_{\hat{H}}(a_G+w_GQ_G)+(1-\tilde{\beta})(1-b_{\hat{H}})\lambda_{\hat{H}}[a_E+w_EQ_E(\lambda_E)+\lambda_Ew_E\frac{\partial Q_E(\lambda_E)}{\partial\lambda_E}].$$

(a) Moreover, we have

$$\begin{aligned}\frac{dC_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\alpha}} &= \frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial\tilde{\alpha}}+\frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{H}\tilde{L}}}\frac{\partial f^e_{\hat{H}\tilde{L}}}{\partial\tilde{\alpha}}\\ &= [\frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial\tilde{\alpha}}\frac{\partial^2\Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^{e2}_{\hat{H}\tilde{L}}}-\frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{H}\tilde{L}}}\frac{\partial^2\Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^e_{\hat{H}\tilde{L}}\partial\tilde{\alpha}}]/\frac{\partial^2\Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^{e2}_{\hat{H}\tilde{L}}},\end{aligned}$$

where

$$\frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial\tilde{\alpha}} = (1-f^e_{\hat{H}\tilde{L}})b_{\hat{H}}\lambda_{\hat{H}}(a_G+w_GQ_G)>0.$$

We then have

$$\begin{aligned}&\frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial\tilde{\alpha}}\frac{\partial^2\Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^{e2}_{\hat{H}\tilde{L}}}-\frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{H}\tilde{L}}}\frac{\partial^2\Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^e_{\hat{H}\tilde{L}}\partial\tilde{\alpha}}\\ &= (1-f^e_{\hat{H}\tilde{L}})b_{\hat{H}}\lambda_{\hat{H}}(a_G+w_GQ_G)[(1-\tilde{\beta})(1-b_{\hat{H}})\lambda_{\hat{H}}]^2w_E\frac{\partial Q_E(\lambda_E)}{\partial\lambda_E}\\ &+\{(1-\tilde{\beta})(1-b_{\hat{H}})\lambda_{\hat{H}}[a_E+w_EQ_E(\lambda_E)+\lambda_Ew_E\frac{\partial Q_E(\lambda_E)}{\partial\lambda_E}]-[\tilde{\alpha}b_{\hat{H}}+(1-\tilde{\beta})(1-b_{\hat{H}})]\lambda_{\hat{H}}(a_G+w_GQ_G)\}\\ &\cdot b_{\hat{H}}\lambda_{\hat{H}}(p_G+w_GQ_G).\end{aligned}$$

As $\frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial\tilde{\alpha}}\frac{\partial^2\Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^{e2}_{\hat{H}\tilde{L}}}-\frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{H}\tilde{L}}}\frac{\partial^2\Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^e_{\hat{H}\tilde{L}}\partial\tilde{\alpha}}$ is linear and increasing in a_E, $\exists\ a^u_E$ s.t. we have $\frac{dC_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\alpha}}<0$ if and only if $a_E<a^u_E$. Note that such a_E can exist given the assumption in Section 3.3.

(b) Meanwhile, we have

$$\begin{aligned}\frac{dC_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\beta}} &= \frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial\tilde{\beta}}+\frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{H}\tilde{L}}}\frac{\partial f^e_{\hat{H}\tilde{L}}}{\partial\tilde{\beta}}\\ &= [\frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial\tilde{\beta}}\frac{\partial^2\Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^{e2}_{\hat{H}\tilde{L}}}-\frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{H}\tilde{L}}}\frac{\partial^2\Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^e_{\hat{H}\tilde{L}}\partial\tilde{\beta}}]/\frac{\partial^2\Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^{e2}_{\hat{H}\tilde{L}}},\end{aligned}$$

where

$$\frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial \beta} = (1-f^e_{\hat{H}\tilde{L}})(1-b_{\hat{H}})\lambda_{\hat{H}}[a_E + w_E Q_E(\lambda_E) + \lambda_E w \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E} - a_G - w_G Q_G] > 0.$$

We then have

$$\begin{aligned}
&\frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial \tilde{\beta}} \frac{\partial^2 \Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^{e2}_{\hat{H}\tilde{L}}} - \frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{H}\tilde{L}}} \frac{\partial^2 \Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^e_{\hat{H}\tilde{L}} \partial \tilde{\beta}} \\
&= (1-f^e_{\hat{H}\tilde{L}})(1-b_{\hat{H}})\lambda_{\hat{H}}[a_E + w_E Q_E(\lambda_E) + \lambda_E w_E \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E} - a_G - w_G Q_G] \\
&\cdot [(1-\tilde{\beta})(1-b_{\hat{H}})\lambda_{\hat{H}}]^2 w_E \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E} \\
&- \{(1-\tilde{\beta})(1-b_{\hat{H}})\lambda_{\hat{H}}[a_E + w_E Q_E(\lambda_E) + \lambda_E w_E \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}] - [\tilde{\alpha} b_{\hat{H}} + (1-\tilde{\beta})(1-b_{\hat{H}})]\lambda_{\hat{H}}(a_G + w_G Q_G)\} \\
&\cdot (1-b_{\hat{H}})\lambda_{\hat{H}}[p_G + w_G Q_G - p_E - w_E Q_E(\lambda_E) + (1-\tilde{\beta})(1-f^e_{\hat{H}\tilde{L}})(1-b_{\hat{H}})\lambda_{\hat{H}} w_E \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}].
\end{aligned}$$

As $\frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial \tilde{\beta}} \frac{\partial^2 \Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^{e2}_{\hat{H}\tilde{L}}} - \frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{H}\tilde{L}}} \frac{\partial^2 \Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^e_{\hat{H}\tilde{L}} \partial \tilde{\beta}}$ is linear and increasing in a_E, $\exists\ a^o_E$ s.t. we have $\frac{dC_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\beta}} < 0$ if and only if $a_E < a^o_E$. Note that such a_E can exist given the assumption in Section 3.3.

The proof for the results with $\tilde{\mathbf{f}}^{\mathbf{e}} \in \{(0,1,m,1),(m,1,1,1)\}$ follows the same procedures. □

Proof of Proposition 3. When $\tilde{\alpha}+\tilde{\beta}=1$, we have $b_{\hat{L}\tilde{L}} = b_{\hat{L}\tilde{H}} = b_{\hat{L}}$ and $b_{\hat{H}\tilde{L}} = b_{\hat{H}\tilde{H}} = b_{\hat{H}}$. The adoption of virtual triage in this case is uninformative and therefore does not change patients' care-seeking behavior nor equilibrium social cost.

If $\hat{\mathbf{f}}^{\mathbf{e}} \in \{(0,m),(m,1)\}$, we have mixed strategy equilibria in the absence of virtual triage; and therefore, we also have mixed strategy equilibria in the presence of virtual triage with $\tilde{\alpha}+\tilde{\beta}=1$. Hence, we have $\tilde{\mathbf{f}}^{\mathbf{e}} \in R_{m,\tilde{L}} \cup R_{m,\tilde{H}}$ with $\tilde{\alpha}+\tilde{\beta}=1$. In this case, as $\tilde{\alpha}$ or $\tilde{\beta}$ decreases, we could have higher equilibrium social cost (conditional on ED service cost) by Lemma 4. Since we have $C_s(\tilde{\mathbf{f}}^{\mathbf{e}}) = C_s(\hat{\mathbf{f}}^{\mathbf{e}})$ with $\tilde{\alpha}+\tilde{\beta}=1$, this implies that $\exists\ \tilde{\alpha}+\tilde{\beta}<1$ and a_E s.t. $C_s(\hat{\mathbf{f}}^{\mathbf{e}}) < C_s(\tilde{\mathbf{f}}^{\mathbf{e}})$.

If $\hat{\mathbf{f}}^{\mathbf{e}} \in \{(0,0),(0,1),(1,1)\}$, we have pure strategy equilibria with $\tilde{\alpha}+\tilde{\beta}=1$ in the presence of virtual triage, and therefore we have $\tilde{\mathbf{f}}^{\mathbf{e}} \in R_{p,\infty}$ with $\tilde{\alpha}+\tilde{\beta}=1$. In this case, as $\tilde{\alpha}$ or $\tilde{\beta}$ decreases, $\tilde{\mathbf{f}}^{\mathbf{e}} \in R_{p,\infty}$ continues to hold along with $C_s(\tilde{\mathbf{f}}^{\mathbf{e}}) = C_s(\hat{\mathbf{f}}^{\mathbf{e}})$, until $\tilde{\alpha}$ or $\tilde{\beta}$ is sufficiently small such that virtual triage changes the care-seeking behavior of some patient type. Given the convexity of the potential function $\Phi(\mathbf{f})$, this means as $\tilde{\alpha}$ or $\tilde{\beta}$ decreases and becomes sufficiently small, we will move from a corner solution for minimizing the potential function where all patient types adopt pure strategy in equilibrium with $C_s(\tilde{\mathbf{f}}^{\mathbf{e}}) = C_s(\hat{\mathbf{f}}^{\mathbf{e}})$ to an interior solution where some patient type adopts mixed strategy in equilibrium. Hence, once entering a mixed strategy equilibrium region, as $\tilde{\alpha}$ or $\tilde{\beta}$ further decreases, we could have higher equilibrium social cost (conditional on ED service cost) by Lemma 4. This implies that $\exists\ \tilde{\alpha}+\tilde{\beta}<1$ and a_E s.t. $C_s(\hat{\mathbf{f}}^{\mathbf{e}}) < C_s(\tilde{\mathbf{f}}^{\mathbf{e}})$.

Hence, we have $\forall\ \hat{\mathbf{f}}^{\mathbf{e}} \in [0,1]^2$, $\exists\ \tilde{\alpha}+\tilde{\beta}<1$ and a_E s.t. $C_s(\hat{\mathbf{f}}^{\mathbf{e}}) < C_s(\tilde{\mathbf{f}}^{\mathbf{e}})$. □

Proof of Lemma 5. (i) When $\tilde{\mathbf{f}}^{\mathbf{e}} \in R_{p,\infty}$, the adoption of virtual triage does not change patient care-seeking behavior. Hence we have $\frac{dD_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\alpha}} = \frac{dD_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\beta}} = 0$.

(ii) In equilibrium region $\tilde{\mathbf{f}}^{\mathbf{e}} = (0,1,1,1)$, we have $D_s(\tilde{\mathbf{f}}^{\mathbf{e}}) = b_{\hat{L}\tilde{L}}\lambda_{\hat{L}\tilde{L}} = \tilde{\alpha} b_{\hat{L}}\lambda_{\hat{L}}$. In equilibrium region $\tilde{\mathbf{f}}^{\mathbf{e}} = (0,0,0,1)$, we have $D_s(\tilde{\mathbf{f}}^{\mathbf{e}}) = b_{\hat{L}}\lambda_{\hat{L}} + b_{\hat{H}\tilde{L}}\lambda_{\hat{H}\tilde{L}} = b_{\hat{L}}\lambda_{\hat{L}} + \tilde{\alpha} b_{\hat{H}}\lambda_{\hat{H}}$. In equilibrium region $\tilde{\mathbf{f}}^{\mathbf{e}} = (0,1,0,1)$, we have $D_s(\tilde{\mathbf{f}}^{\mathbf{e}}) = b_{\hat{L}\tilde{L}}\lambda_{\hat{L}\tilde{L}} + b_{\hat{L}\tilde{H}}\lambda_{\hat{H}\tilde{L}} = \tilde{\alpha} b_{\hat{L}}\lambda_{\hat{L}} + \tilde{\alpha} b_{\hat{H}}\lambda_{\hat{H}}$.

Hence, we have $\frac{dD_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\alpha}} > 0$ and $\frac{dD_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\beta}} = 0$. □

Proof of Lemma 6. (i) In equilibrium region $\tilde{\mathbf{f}}^{\mathbf{e}} = (0,m,1,1)$, we have $D_s(\tilde{\mathbf{f}}^{\mathbf{e}}) = b_{\hat{L}\tilde{L}}\lambda_{\hat{L}\tilde{L}} + (1 - f^e_{\hat{L}\tilde{H}})\lambda_{\hat{L}\tilde{H}} = [1 - f^e_{\hat{L}\tilde{H}}(1-\tilde{\alpha})] b_{\hat{L}}\lambda_{\hat{L}}$.

(a) We have

$$\begin{aligned}\frac{dD_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\alpha}} &= \frac{\partial D_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial \tilde{\alpha}} + \frac{\partial D_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{L}\tilde{H}}}\frac{\partial f^e_{\hat{L}\tilde{H}}}{\partial \tilde{\alpha}} \\ &= [\frac{\partial D_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial \tilde{\alpha}}\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^{e2}_{\hat{L}\tilde{H}}} - \frac{\partial D_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{L}\tilde{H}}}\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^e_{\hat{L}\tilde{H}}\partial \tilde{\alpha}}] / \frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^{e2}_{\hat{L}\tilde{H}}},\end{aligned}$$

and we have

$$\frac{\partial D_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial \tilde{\alpha}}\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^{e2}_{\hat{L}\tilde{H}}} - \frac{\partial D_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{L}\tilde{H}}}\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^e_{\hat{L}\tilde{H}}\partial \tilde{\alpha}} = f^e_{\hat{L}\tilde{H}} b_{\hat{L}}\lambda_{\hat{L}}\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^{e2}_{\hat{L}\tilde{H}}} + (1-\tilde{\alpha}) b_{\hat{L}}\lambda_{\hat{L}}\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^e_{\hat{L}\tilde{H}}\partial \tilde{\alpha}} > 0,$$

as we have $\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^{e2}_{\hat{L}\tilde{H}}} > 0$ and $\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^e_{\hat{L}\tilde{H}}\partial \tilde{\alpha}} > 0$ as shown in (EC.20) and (EC.21). Hence, we have $\frac{dD_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\alpha}} > 0$.

(b) Meanwhile, we have

$$\begin{aligned}\frac{dD_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\beta}} &= \frac{\partial D_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial \tilde{\beta}} + \frac{\partial D_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{L}\tilde{H}}}\frac{\partial f^e_{\hat{L}\tilde{H}}}{\partial \tilde{\beta}} \\ &= [\frac{\partial D_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial \tilde{\beta}}\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^{e2}_{\hat{L}\tilde{H}}} - \frac{\partial D_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{L}\tilde{H}}}\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^e_{\hat{L}\tilde{H}}\partial \tilde{\beta}}] / \frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^{e2}_{\hat{L}\tilde{H}}},\end{aligned}$$

and we have

$$\frac{\partial D_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial \tilde{\beta}}\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^{e2}_{\hat{L}\tilde{H}}} - \frac{\partial D_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{L}\tilde{H}}}\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^e_{\hat{L}\tilde{H}}\partial \tilde{\beta}} = (1-\tilde{\alpha}) b_{\hat{L}}\lambda_{\hat{L}}\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^e_{\hat{L}\tilde{H}}\partial \tilde{\beta}} > 0,$$

as we have $\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^e_{\hat{L}\tilde{H}}\partial \tilde{\beta}} > 0$ as shown in (EC.22). Hence, we have $\frac{dD_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\beta}} > 0$.

The proof for the results with $\tilde{\mathbf{f}} \in \{(0,m,0,1),(0,0,0,m)\}$ follows the same procedures.

(ii) In equilibrium region $\tilde{\mathbf{f}}^{\mathbf{e}} = (0,0,m,1)$, we have $D_s(\tilde{\mathbf{f}}^{\mathbf{e}}) = b_{\hat{L}}\lambda_{\hat{L}} + (1 - f^e_{\hat{H}\tilde{L}})\lambda_{\hat{H}\tilde{L}} = b_{\hat{L}}\lambda_{\hat{L}} + (1 - f^e_{\hat{H}\tilde{L}})\tilde{\alpha} b_{\hat{H}}\lambda_{\hat{H}}$.

(a) We have

$$\begin{aligned}\frac{dD_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\alpha}} &= \frac{\partial D_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial \tilde{\alpha}} + \frac{\partial D_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{H}\tilde{L}}}\frac{\partial f^e_{\hat{H}\tilde{L}}}{\partial \tilde{\alpha}} \\ &= [\frac{\partial D_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial \tilde{\alpha}}\frac{\partial^2 \Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^{e2}_{\hat{H}\tilde{L}}} - \frac{\partial D_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{H}\tilde{L}}}\frac{\partial^2 \Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^e_{\hat{H}\tilde{L}}\partial \tilde{\alpha}}] / \frac{\partial^2 \Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^{e2}_{\hat{H}\tilde{L}}},\end{aligned}$$

and we have

$$\frac{\partial D_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial \tilde{\alpha}}\frac{\partial^2\Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^{e2}_{\hat{H}\tilde{L}}} - \frac{\partial D_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{H}\tilde{L}}}\frac{\partial^2\Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^e_{\hat{H}\tilde{L}}\partial\tilde{\alpha}}$$
$$= (1-f^e_{\hat{H}\tilde{L}})b_{\hat{H}}\lambda_{\hat{H}}[(1-\tilde{\beta})(1-b_{\hat{H}})\lambda_{\hat{H}}]^2 w_E\frac{\partial Q_E(\lambda_E)}{\partial\lambda_E} - \tilde{\alpha}(b_{\hat{H}}\lambda_{\hat{H}})^2(p_G + w_G Q_G).$$

As $\frac{\partial D_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial \tilde{\alpha}}\frac{\partial^2\Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^{e2}_{\hat{H}\tilde{L}}} - \frac{\partial D_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{H}\tilde{L}}}\frac{\partial^2\Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^e_{\hat{H}\tilde{L}}\partial\tilde{\alpha}}$ is linear and decreasing in p_G, $\exists\ p^u_G$ s.t. we have $\frac{dD_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\alpha}} < 0$ if and only if $p_G > p^u_G$.

(b) Meanwhile, we have

$$\frac{dD_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\beta}} = \frac{\partial D_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial\tilde{\beta}} + \frac{\partial D_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{H}\tilde{L}}}\frac{\partial f^e_{\hat{H}\tilde{L}}}{\partial\tilde{\beta}}$$
$$= [\frac{\partial D_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial\tilde{\beta}}\frac{\partial^2\Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^{e2}_{\hat{H}\tilde{L}}} - \frac{\partial D_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{H}\tilde{L}}}\frac{\partial^2\Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^e_{\hat{H}\tilde{L}}\partial\tilde{\beta}}]/\frac{\partial^2\Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^{e2}_{\hat{H}\tilde{L}}},$$

and we have

$$\frac{\partial D_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial\tilde{\beta}}\frac{\partial^2\Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^{e2}_{\hat{H}\tilde{L}}} - \frac{\partial D_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{H}\tilde{L}}}\frac{\partial^2\Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^e_{\hat{H}\tilde{L}}\partial\tilde{\beta}}$$
$$= \tilde{\alpha}b_{\hat{H}}\lambda_{\hat{H}}(1-b_{\hat{H}})\lambda_{\hat{H}}[p_G + w_G Q_G - p_E - w_E Q_E(\lambda_E) + (1-\tilde{\beta})(1-f^e_{\hat{H}\tilde{L}})(1-b_{\hat{H}})\lambda_{\hat{H}} w_E\frac{\partial Q_E(\lambda_E)}{\partial\lambda_E}].$$

As $\frac{\partial D_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial\tilde{\beta}}\frac{\partial^2\Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^{e2}_{\hat{H}\tilde{L}}} - \frac{\partial D_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{H}\tilde{L}}}\frac{\partial^2\Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^e_{\hat{H}\tilde{L}}\partial\tilde{\beta}}$ is linear and increasing in p_G, $\exists\ p^o_G$ s.t. we have $\frac{dD_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\beta}} < 0$ if and only if $p_G < p^o_G$.

The proof for the results with $\tilde{\mathbf{f}}^{\mathbf{e}} \in \{(0,1,m,1),(m,1,1,1)\}$ follows the same procedures. □

EC.4.4. Proofs for Section 6

Proof of Proposition 4. For a given virtual triage tool, suppose $\tilde{\mathbf{f}}^{\mathbf{e}}(r(\tilde{\beta}^*),\tilde{\beta}^*)$ is in the interior of $R_{p,\sim}$. Recall that by "interior", we mean all patient types strictly prefer their choices of care locations for a given pure strategy equilibrium; in other words, it will lead to strictly higher patient costs for any patient that deviates from the equilibrium.

(i) If $\tilde{\mathbf{f}}^{\mathbf{e}}(r(\tilde{\beta}^*),\tilde{\beta}^*) = (0,0,0,1)$, we have

$$\lambda_G = \lambda_{\hat{L}} + [r(\tilde{\beta}^*)b_{\hat{H}} + (1-\tilde{\beta}^*)(1-b_{\hat{H}})]\lambda_{\hat{H}},$$
$$\lambda_E = b_{\hat{L}}\lambda_{\hat{L}} + [b_{\hat{H}} + \tilde{\beta}^*(1-b_{\hat{H}})]\lambda_{\hat{H}},$$

and

$$\frac{\partial^2 C_s(\tilde{\mathbf{f}}^{\mathbf{e}}(r(\tilde{\beta}),\tilde{\beta}))}{\partial\tilde{\beta}^2} > 0.$$

Hence, the unique $\tilde{\beta}^*$ is given by the solution to the following FOC

$$(a_G + w_G Q_G)[(1-b_{\hat{H}})\lambda_{\hat{H}} - r'(\tilde{\beta}^*)b_{\hat{H}}\lambda_{\hat{H}}] = [a_E + w_E Q_E(\lambda_E) + \lambda_E w_E\frac{\partial Q_E(\lambda_E)}{\partial\lambda_E}](1-b_{\hat{H}})\lambda_{\hat{H}}. \quad \text{(EC.24)}$$

The proof for (ii) and (iii) follows the same procedures. □

Proof of Proposition 5. For a given virtual triage tool, suppose $\tilde{\mathbf{f}}^{\mathbf{e}}(r(\tilde{\beta}^*),\tilde{\beta}^*)$ is in the interior of $R_{p,\sim}$.

If $\tilde{\mathbf{f}}^{\mathbf{e}}(r(\tilde{\beta}^*),\tilde{\beta}^*)=(0,0,0,1)$, we have

$$\begin{aligned}\frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}}(r(\tilde{\beta}),\tilde{\beta}))}{\partial\tilde{\beta}}\Big|_{\tilde{\beta}=\tilde{\beta}^*} &= (a_G+w_GQ_G)[r'(\tilde{\beta}^*)b_{\hat{H}}\lambda_{\hat{H}}-(1-b_{\hat{H}})\lambda_{\hat{H}}]\\ &+(a_E+w_EQ_E(\lambda_E)+\lambda_Ew_E\frac{\partial Q_E(\lambda_E)}{\partial\lambda_E})(1-b_{\hat{H}})\lambda_{\hat{H}}=0.\end{aligned} \tag{EC.25}$$

In this case, we have

$$\frac{\partial C_s^2(\tilde{\mathbf{f}}^{\mathbf{e}}(r(\tilde{\beta}),\tilde{\beta}))}{\partial\tilde{\beta}^2}=(a_G+w_GQ_G)r''(\tilde{\beta}^*)b_{\hat{H}}\lambda_{\hat{H}}+[2w_E\frac{\partial Q_E(\lambda_E)}{\partial\lambda_E}+\lambda_Ew_E\frac{\partial^2Q_E(\lambda_E)}{\partial\lambda_E^2}][(1-b_{\hat{H}})\lambda_{\hat{H}}]^2>0,$$

and

$$\begin{aligned}\frac{\partial C_s^2(\tilde{\mathbf{f}}^{\mathbf{e}}(r(\tilde{\beta}),\tilde{\beta}))}{\partial\tilde{\beta}\partial h} &= [2w_E\frac{\partial Q_E(\lambda_E)}{\partial\lambda_E}(1-\tilde{\beta}\hat{\beta})\lambda+\lambda_Ew_E\frac{\partial^2Q_E(\lambda_E)}{\partial\lambda_E^2}(1-\tilde{\beta}\hat{\beta})\lambda]\hat{\beta}(1-h)\lambda\\ &+(a_G+w_GQ_G)[r'(\tilde{\beta})(1-\hat{\alpha})\lambda+\hat{\beta}\lambda]-[a_E+w_EQ_E(\lambda_E)+\lambda_Ew_E\frac{\partial Q_E(\lambda_E)}{\partial\lambda_E}]\hat{\beta}\lambda,\end{aligned}$$

which is linear and decreasing in a_E. Hence, $\exists\ a_E^h$ s.t. we have

$$\frac{\partial\tilde{\beta}^*}{\partial h}=-\frac{\partial C_s^2(\tilde{\mathbf{f}}^{\mathbf{e}}(r(\tilde{\beta}),\tilde{\beta}))}{\partial\tilde{\beta}\partial h}\Big/\frac{\partial C_s^2(\tilde{\mathbf{f}}^{\mathbf{e}}(r(\tilde{\beta}),\tilde{\beta}))}{\partial\tilde{\beta}^2}\Big|_{\tilde{\beta}=\tilde{\beta}^*}>0$$

if and only if $a_E>a_E^h$. The proof for $\tilde{\mathbf{f}}^{\mathbf{e}}(r(\tilde{\beta}^*),\tilde{\beta}^*)\in\{(0,1,1,1),(0,1,0,1)\}$ follows the same procedures. □

Proof of Proposition 6. $C_s(\tilde{\mathbf{f}}^{\mathbf{e}}(r_1(\tilde{\beta}_1^*),\tilde{\beta}_1^*))>C_s(\tilde{\mathbf{f}}^{\mathbf{e}}(r_2(\tilde{\beta}_2^*),\tilde{\beta}_2^*))$ is straightforward as along $\tilde{\alpha}=r_2(\tilde{\beta})$, both virtual under-triage probability and virtual over-triage probability can be reduced compared with $\tilde{\alpha}=r_1(\tilde{\beta})$, and this leads to lower equilibrium social cost in pure strategy equilibrium regions by Lemma 3.

Suppose $\tilde{\mathbf{f}}^{\mathbf{e}}(r_1(\tilde{\beta}_1^*),\tilde{\beta}_1^*)=\tilde{\mathbf{f}}^{\mathbf{e}}(r_2(\tilde{\beta}_2^*),\tilde{\beta}_2^*)=(0,0,0,1)$. $\tilde{\beta}_1^*$ is given by the solution to

$$(a_G+w_GQ_G)[(1-b_{\hat{H}})\lambda_{\hat{H}}-r_1'(\tilde{\beta}_1^*)b_{\hat{H}}\lambda_{\hat{H}}]=[a_E+w_EQ_E(\lambda_E)+\lambda_Ew_E\frac{\partial Q_E(\lambda_E)}{\partial\lambda_E}](1-b_{\hat{H}})\lambda_{\hat{H}}.$$

(i) If

$$(a_G+w_GQ_G)[(1-b_{\hat{H}})\lambda_{\hat{H}}-r_2'(\tilde{\beta}_1^*)b_{\hat{H}}\lambda_{\hat{H}}]>[a_E+w_EQ_E(\lambda_E)+\lambda_Ew_E\frac{\partial Q_E(\lambda_E)}{\partial\lambda_E}](1-b_{\hat{H}})\lambda_{\hat{H}},$$

we have

$$0>r_1'(\tilde{\beta}_1^*)>r_2'(\tilde{\beta}_1^*).$$

In this case, we have $\tilde{\beta}_1^*<\tilde{\beta}_2^*$, and $r_1(\tilde{\beta}_1^*)>r_2(\tilde{\beta}_2^*)$.

(ii) If

$$(a_G + w_G Q_G)[(1-b_{\hat{H}})\lambda_{\hat{H}} - r_2'(\tilde{\beta}_1^*)b_{\hat{H}}\lambda_{\hat{H}}] < [a_E + w_E Q_E(\lambda_E) + \lambda_E w_E \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}](1-b_{\hat{H}})\lambda_{\hat{H}},$$

we have $r_1'(\tilde{\beta}_1^*) < r_2'(\tilde{\beta}_1^*)$. Define r_2^d s.t. it is given by

$$(a_G + w_G Q_G)[(1-b_{\hat{H}})\lambda_{\hat{H}} - r_2^d b_{\hat{H}}\lambda_{\hat{H}}] = [a_E + w_E Q_E(\lambda_E) + \lambda_E w_E \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}](1-b_{\hat{H}})\lambda_{\hat{H}}.$$

If we have

$$r_2'(\tilde{\beta}^c) > r_2^d,$$

we then have $\tilde{\beta}_1^* > \tilde{\beta}_2^*$, and $r_1(\tilde{\beta}_1^*) < r_2(\tilde{\beta}_2^*)$.

(iii) If

$$r_1'(\tilde{\beta}_1^*) < r_2'(\tilde{\beta}_1^*),$$

and

$$r_2'(\tilde{\beta}^c) < r_2^d,$$

we have $\tilde{\beta}_1^* > \tilde{\beta}_2^*$, and $r_1(\tilde{\beta}_1^*) > r_2(\tilde{\beta}_2^*)$.

The proof for $\tilde{\mathbf{f}}^{\mathbf{e}}(r(\tilde{\beta}^*), \tilde{\beta}^*) \in \{(0,1,1,1),(0,1,0,1)\}$ follows the same procedures. □

Proof of Proposition 7. The results follow from Lemma 5: When $\tilde{\mathbf{f}}^{\mathbf{e}} \in R_{p,\sim}$, we have $\frac{dD_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\alpha}} > 0$ and $\frac{dD_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\beta}} = 0$. Hence, in the interior of $R_{p,\sim}$, we can always improve triage safety subjective to a given IROC curve by increasing $\tilde{\beta}$ which leads to lower $\tilde{\alpha} = r(\tilde{\beta})$. □

EC.5. Extensions and Robustness

In this section, we extend our model and generalize several of our earlier assumptions.

EC.5.1. Heterogeneity of Patient Self-Triage Accuracy

In the model presented in Section 3, we assumed patient self-triage accuracy to be homogeneous, characterized by the pair of self under-triage and over-triage probabilities $(\hat{\alpha}, \hat{\beta})$. Yet in reality, patients may be heterogeneous in self-triage accuracy. We consider this scenario in Proposition EC.1.

PROPOSITION EC.1. *Suppose there are m groups of patients in the absence of self-triage and virtual triage, where group i patients have an arrival rate of GP-type patients $\lambda_{L,i} > 0$ and an arrival rate of ED-type patients $\lambda_{H,i} > 0, i \in \{1,2,...,m\}$, with $\sum_{i=1}^m \lambda_{L,i} = \lambda_L$ and $\sum_{i=1}^m \lambda_{H,i} = \lambda_H$. Group i patients have the self under-triage probability $\hat{\alpha}_i$ and self over-triage probability $\hat{\beta}_i$, with $\hat{\alpha}_i + \hat{\beta}_i \leq 1$. The structural insights in Sections 4, 5 and 6 continue to hold.*

The structural insights remain robust because the heterogeneity of self-triage accuracy for patients of different groups with the same care-seeking behavior in equilibrium has no impact in aggregate on the system performance metrics (i.e., social cost and triage safety) or on the decisions of the virtual triage provider. In this case, we can construct a hypothetical group of such patients with one belief of being ED-type and an aggregate arrival rate. Meanwhile, there will be at most one group containing patients who adopt a mixed strategy in equilibrium by Lemma EC.1. We can thus have a model equivalent to the one given in Section 3 for the analysis of system performance, and therefore the structural insights hold.

EC.5.2. Belief-dependent Disutility of Waiting

In the model presented in Section 3, we have assumed that patients' disutility of waiting per unit time is a function of their location of care choices, i.e., GP or ED, while independent of their belief of being ED-type. While we expect strategic patients of moderate acuity should have similar disutility of waiting per unit time under different beliefs of being ED-type, we relax this assumption nevertheless. Specifically, when making the decisions on care locations, patients' true type is unknown and characterized by their belief b of being ED-type. Naturally, we expect patients with higher belief of being ED-type to be more sensitive to waiting. To capture this, we assume that when a patient with belief b of being ED-type makes the decisions on care locations, their disutility of waiting per unit time at a GP is $w_G + bw$, while the disutility of waiting per unit time at the ED is $w_E + bw$. In this case, we can capture the heterogeneity of the patients' disutility of waiting per unit time across both care locations and belief of being ED-type. Moreover, we note that including belief-dependent disutility of waiting introduces additional interesting and natural dynamics in the model: for a patient with an initial belief b that visit the ED directly, their disutility of waiting per unit time at the ED is $w_E + bw$; on the other hand, if they visit a GP first and then referred to the ED, they are then revealed as ED-type with certainty (i.e., $b = 1$) and therefore their disutility of waiting per unit time at the ED is $w_E + w$. Hence, the belief-dependent disuitlity of waiting introduces an additional layer of heterogeneity for patients of the same initial type, due to the differences in their choices of care pathway. We are able to capture this in the extended model.

Meanwhile, this has implications for the calculation of social cost as well. Note that for the calculation of disutility of waiting as part of social cost, we should use the disutility of waiting based on observed patient's true type. This means that for the calculation of social cost, we have $w_E + 0 \cdot w = w_E$ as the disutility of waiting per unit time for a GP-type patient that visits the ED; $w_E + 1 \cdot w = w_E + w$ for an ED-type patient that visits the ED; $w_G + 0 \cdot w = w_G$ for a GP-type patient that visits a GP; and $w_G + 1 \cdot w = w_G + w$ for an ED-type patient that visits a GP. Hence, for patients with belief b that visit the ED, the disutility of waiting per unit time at the ED is

$w_E + w$ for b fraction of them and w_E for $1-b$ fraction of them; and therefore the average disutility of waiting per unit time at the ED for these patients is $b(w_E + w) + (1-b)w_E = w_E + bw$. Similarly, for patients with belief b that visit a GP, the average disutility of waiting per unit time at a GP for these patients is $b(w_G + w) + (1-b)w_G = w_G + bw$. We use these expressions for the calculation of disutility of waiting in social cost.

With the added layer of complexity, we can show by the following proposition that all the analytical results continue to hold. Notably, the results hold under the same assumptions on a_E and p_E as in Section 3.3.

PROPOSITION EC.2. *Suppose for a patient with belief b being ED-type, their disutility of waiting per unit time at a GP is $w_G + bw$, while the disutility of waiting per unit time at the ED is $w_E + bw$. The structural insights in Sections 4, 5 and 6 continue to hold.*

EC.5.3. Other Costs and Disutilities for Patients

In the model presented in Section 3, we have made the simplifying assumption that the only costs that patients incur arise from the disutility of waiting and from their GP and/or ED co-payment. Yet in reality, patients may incur other costs, such as transportation and parking fees, and other disutilities, such as those associated with pain and stress. Let c_G denote the expected cost the patient incurs for a GP visit in addition to the disutility of waiting and GP co-payment, and let c_E denote the expected cost the patient incurs for an ED visit in addition to the disutility of waiting and ED co-payment. We continue to assume that $a_E > a_G + w_G Q_G + c_G - w_E Q_E(\lambda_H) - c_E$ to capture the reality that an ED visit is more costly than a GP visit, and $p_E > p_G + w_G Q_G + c_G - w_E Q_E(\lambda_H) - c_E$ as otherwise all patients will visit the ED directly regardless of their self-triage decisions or virtual triage recommendations, which clearly does not reflect reality. In this case, we can easily show the robustness of our base model: As the functional forms of expected disutilities of waiting are general, any additional expected cost can be incorporated into the expected disutilities of waiting as a constant term. Hence, all the results will remain the same.

EC.5.4. Assumptions on Full Information and Rationality of Patients

In order to characterize patient care-seeking behavior, we have assumed that patients have full information and rationality in the model. While these assumptions are commonly invoked in the literature, we discuss possible relaxations and their implications on the managerial insights.

First, we note that not all our results are sensitive to the assumption of full information. In particular, in pure strategy equilibrium regions, patient care-seeking behavior is insensitive to full information on model parameters – we could have a range of parameter value combinations while

patients' decisions remain the same: the pure strategy equilibrium regions in Figure 2 illustrate one such scenario where patients' decisions are insensitive to virtual triage accuracy. For example, all patients follow virtual triage recommendations regardless self-triage decisions if patients are informed of small virtual under-triage and over-triage probabilities, not necessarily the precise accuracy. Hence, our analysis and insights associated with pure strategy equilibrium regions – either on patient behavior, system performance with off-the-shelf virtual triage, or system performance with custom virtual triage – remain robust and continue to hold in the absence of full information, as long as patients have a general sense of different system parameters.

On the other hand, the precise characterization of mixed strategy equilibrium regions does require full information – a marginal change in any system parameter will change the behavior of patients adopting mixed strategy in equilibrium. Moreover, the existence of mixed strategy equilibria also depends on the assumption that patients have full rationality, while in reality patients have bounded rationality at best and it is likely that they may actually not adopt mixed strategies. Realizing this, it triggers the following question: do we lose the insights related to mixed strategy equilibria in their absence? In particular, do the results on the potential non-monotonicity of system performance in off-the-shelf virtual triage accuracy (Lemmas 4 and 6 of the revised manuscript), which hinge on the existence of mixed strategy equilibria, still exist?

To understand this, we examine the alternative scenario where patients have both limited information and bounded rationality, and therefore patients have a simple decision to make, i.e., whether or not to follow virtual triage recommendations with certainty. As a result, we only have pure strategy equilibrium regions. In this case, we can in fact easily show that at the boundaries of pure strategy equilibrium regions, the system performance could have a jump discontinuity and potentially non-monotonicity in virtual triage accuracy. For example, when a patient type switches from visiting a GP first to visiting the ED directly as off-the-shelf virtual triage accuracy gets higher, they do so because the expected patient cost (i.e., sum of co-payment and disutility of waiting) of visiting the ED directly is lower than visiting a GP first. This decision is independent of ED service cost and could lead to higher equilibrium social cost if ED service cost is sufficiently high. Now we can see that with the assumptions on full information and rationality, the existence of mixed strategy equilibrium regions essentially eliminates jump discontinuity and make system performance being continuous in virtual triage accuracy. Hence, the structural insights associated with mixed strategy equilibrium outcomes in the presence of off-the-shelf virtual triage remain under limited information and bounded rationality.

EC.5.5. Proofs for EC.5

Proof of Proposition EC.1. For the given self under-triage probability $\hat{\alpha}_i$ and self over-triage probability $\hat{\beta}_i, i \in \{1,2,...,m\}$, let $b_{\hat{L},i}$ and $b_{\hat{H},i}$ denote the beliefs of being ED-type for self-triaged GP-type patients and ED-type patients of group i, where $\lambda_{\hat{L},i}$ and $\lambda_{\hat{H},i}$ denote their associated arrival rates. Upon patient self-triage, we then have $2m$ groups of patients, i.e., $\{(\hat{L},1),(\hat{H},1),...,(\hat{L},m),(\hat{H},m)\}$. Moreover, let $b_{\hat{L}\tilde{L},i}$, $b_{\hat{L}\tilde{H},i}$, $b_{\hat{H}\tilde{L},i}$ and $b_{\hat{H}\tilde{H},i}$ denote the beliefs of being ED-type after virtual triage recommendations of group i, where $\lambda_{\hat{L}\tilde{L},i}$, $\lambda_{\hat{L}\tilde{H},i}$, $\lambda_{\hat{H}\tilde{L},i}$ and $\lambda_{\hat{H}\tilde{H},i}$ denote their associated arrival rates. Hence, with virtual triage recommendations, we then have $4m$ groups of patients, i.e., $\{(\hat{L}\tilde{L},1),(\hat{L}\tilde{H},1),(\hat{H}\tilde{L},1),(\hat{H}\tilde{H},1),...,(\hat{L}\tilde{L},m),(\hat{L}\tilde{H},m),(\hat{H}\tilde{L},m),(\hat{H}\tilde{H},m)\}$.

By Lemma EC.1, we have at most one group of patients that adopt a mixed strategy in equilibrium, out of the $4m$ groups. Suppose we have a pure strategy equilibrium. Let $N_G \subseteq \{(\hat{L}\tilde{L},1),(\hat{L}\tilde{H},1),(\hat{H}\tilde{L},1),(\hat{H}\tilde{H},1),...,(\hat{L}\tilde{L},m),(\hat{L}\tilde{H},m),(\hat{H}\tilde{L},m),(\hat{H}\tilde{H},m)\}$ denotes the groups of patients that follow GP recommendations from virtual triage with certainty; let $N_E \subseteq \{(\hat{L}\tilde{L},1),(\hat{L}\tilde{H},1),(\hat{H}\tilde{L},1),(\hat{H}\tilde{H},1),...,(\hat{L}\tilde{L},m),(\hat{L}\tilde{H},m),(\hat{H}\tilde{L},m),(\hat{H}\tilde{H},m)\}$ denotes the groups of patients that follow ED recommendations from virtual triage with certainty. In this case, for N_G, we can have an equivalent hypothetical group of patients with their belief of being ED-type being $\frac{\sum_{i\in N_G} b_i\lambda_i}{\sum_{i\in N_G}\lambda_i}$ with an arrival rate $\sum_{i\in N_G}\lambda_i$; for N_E, we can have an equivalent hypothetical group of patients with their belief of being ED-type being $\frac{\sum_{i\in N_E} b_i\lambda_i}{\sum_{i\in N_E}\lambda_i}$ with an arrival rate $\sum_{i\in N_E}\lambda_i$. As a result, for the analysis of system performance metrics (i.e., social cost and triage safety), we have an equilibrium patient flow that is equivalent to $\tilde{\mathbf{f}}^{\mathbf{e}} \in R_{p,\asymp} \cup R_{p,\sim}$. Otherwise, suppose we have a mixed strategy equilibrium. In this case, for the analysis of system performance (social cost or triage safety), we have an equilibrium patient flow that is equivalent to $\tilde{\mathbf{f}}^{\mathbf{e}} \in R_{m,\tilde{L}} \cup R_{m,\tilde{H}}$. □

Proof of Proposition EC.2. Under belief-dependent disutility of waiting, we have the potential function

$$\Phi(\mathbf{f}) = \sum_{T\in\mathbf{T}} \int_0^{(1-f_T)\lambda_T} b_T w Q_G dx + \int_0^{\lambda_G(\mathbf{f})} (p_G + w_G Q_G) dx + \int_0^{\lambda_E(\mathbf{f})} (p_E + w_E Q_E(x)) dx, \quad \text{(EC.26)}$$

where $\mathbf{T}$ is the set of the types of patients. Note that we do not have terms involving $b_T w Q_E(\lambda_E(\mathbf{f}))$ in (EC.26): if a patient of type T visits a GP first, they will incur the disutility term $wQ_E(\lambda_E(\mathbf{f}))$ with probability b_T (i.e., when they are referred to the ED); if they visit the ED directly, they will incur the disutility term $b_T wQ_E(\lambda_E(\mathbf{f}))$ with certainty. We can easily see these two terms have the same value and therefore the disutility associated with term $b_T wQ_E(\lambda_E(\mathbf{f}))$ is patient flow invariant, and therefore it cancels out in (EC.26). Meanwhile, it is straightforward to show that $\Phi(\mathbf{f})$ continues to be convex in $\mathbf{f}$ in this case, as the disutility of waiting per unit time remains as a constant coefficient to the expected waiting time at a GP or the ED.

Correctness of Lemmas EC.1 and EC.2 under belief-dependent disutility of waiting:

Case 1: Suppose $\exists\, i \in \{1,2,...,n\}$ s.t. $f_i^e \in (0,1)$. In this case, we have an interior solution for f_i^e, and therefore we have

$$\left.\frac{\partial \Phi(\mathbf{f})}{\partial f_i}\right|_{\mathbf{f^e}} = -(p_G + b_i w Q_G + w_G Q_G)\lambda_i + (1-b_i)\lambda_i(p_E + w_E Q_E(\lambda_E(\mathbf{f^e}))) = 0.$$

Since $\forall\ j < i$ we have $b_j < b_i$ by assumption, this implies that

$$\begin{aligned}\left.\frac{\partial \Phi(\mathbf{f})}{\partial f_j}\right|_{\mathbf{f^e}} &= -(p_G + b_j w Q_G + w_G Q_G)\lambda_j + (1-b_j)\lambda_j(p_E + w_E Q_E(\lambda_E(\mathbf{f^e}))) > 0 \\ &\Leftrightarrow p_E + w_E Q_E(\lambda_E(\mathbf{f^e})) > p_G + b_j w Q_G + w_G Q_G + b_j(p_E + w_E Q_E(\lambda_E(\mathbf{f^e}))).\end{aligned} \tag{EC.27}$$

Hence, we have $f_j^e = 0, \forall\ j < i$: If $f_j^e > 0$, group j patients will find they can enjoy lower patient costs by visiting a GP first instead of visiting the ED directly according to (EC.27), which will lead to lower f_j^e. This contradicts to f_j^e being the equilibrium patient flow for group j patients. Similarly, $\forall\ k > i$ we have $b_k > b_i$ by assumption, and therefore

$$\begin{aligned}\left.\frac{\partial \Phi(\mathbf{f})}{\partial f_k}\right|_{\mathbf{f^e}} &= -(p_G + b_k w Q_G + w_G Q_G)\lambda_k + (1-b_k)\lambda_k(p_E + w_E Q_E(\lambda_E(\mathbf{f^e}))) < 0 \\ &\Leftrightarrow p_E + w_E Q_E(\lambda_E(\mathbf{f^e})) < p_G + b_k w Q_G + w_G Q_G + b_k(p_E + w_E Q_E(\lambda_E(\mathbf{f^e}))).\end{aligned} \tag{EC.28}$$

Hence, we have $f_k^e = 1, \forall\ k > i$: If $f_k^e < 1$, group k patients will find they can enjoy lower patient costs by visiting the ED directly instead of visiting a GP first according to (EC.28), which will lead to higher f_k^e. This contradicts to f_k^e being the equilibrium patient flow for group k patients.

Case 2: Suppose $\nexists\, i \in \{1,2,...,n\}$ s.t. $f_i^e \in (0,1)$. The proof follows the same procedures as the Case 2 of the proof for Lemma EC.1.

With Lemma EC.1 holds under belief-dependent disutility of waiting, it is straightforward to show that Lemma EC.2 also holds in this case.

Correctness of Proposition 1 under belief-dependent disutility of waiting:

First, the sets of all possible equilibrium regions both in the absence and in the presence of virtual triage remain the same given Lemmas EC.1 and EC.2 continue to hold. We then prove the subsets of equilibrium regions in the presence of virtual triage for any given equilibrium region in the absence of virtual triage remain the same.

Let $C_{\hat{T},l}(\hat{\mathbf{f}})$ denote the patient costs, i.e., sum of the total disutility of waiting and co-payment, for a patient of type $\hat{T}$ visiting l given a patient flow $\hat{\mathbf{f}}$ in the absence of virtual triage, where $\hat{T} \in \{\hat{L}, \hat{H}\}$ and $l \in \{G, E\}$; let $C_{\hat{T}\tilde{T},l}(\tilde{\mathbf{f}})$ denote the patient costs for a patient of type $\hat{T}\tilde{T}$ visiting l given a patient flow $\tilde{\mathbf{f}}$ in the presence of virtual triage, where $\tilde{T} \in \{\tilde{L}, \tilde{H}\}$. Specifically, we have

$$\begin{aligned}C_{\hat{T},G}(\hat{\mathbf{f}}) &= p_G + (b_{\hat{T}} w + w_G) Q_G + b_{\hat{T}}[p_E + (w + w_E) Q_E(\lambda_E(\hat{\mathbf{f}}))] \\ C_{\hat{T},E}(\hat{\mathbf{f}}) &= p_E + (b_{\hat{T}} w + w_E) Q_E(\lambda_E(\hat{\mathbf{f}})) \\ C_{\hat{T}\tilde{T},G}(\tilde{\mathbf{f}}) &= p_G + (b_{\hat{T}\tilde{T}} w + w_G) Q_G + b_{\hat{T}\tilde{T}}[p_E + (w + w_E) Q_E(\lambda_E(\tilde{\mathbf{f}}))] \\ C_{\hat{T}\tilde{T},E}(\tilde{\mathbf{f}}) &= p_E + (b_{\hat{T}\tilde{T}} w + w_E) Q_E(\lambda_E(\tilde{\mathbf{f}})).\end{aligned}$$

When $\hat{\mathbf{f}}^{\mathbf{e}} = (0,0)$, both $\hat{L}$ and $\hat{H}$ patients visit a GP first with certainty in equilibrium, as visiting the ED directly does not lead to lower patient costs for either of them. Hence, we have

$$\begin{aligned} & C_{\hat{H},E}(0,0) \geq C_{\hat{H},G}(0,0) \\ \Leftrightarrow\ & (1-b_{\hat{H}})[p_E + w_E Q_E(b_{\hat{L}}\lambda_{\hat{L}} + b_{\hat{H}}\lambda_{\hat{H}})] \geq p_G + (b_{\hat{H}}w + w_G)Q_G. \end{aligned} \tag{EC.29}$$

Note that we also have $C_{\hat{L},E}(0,0) \geq C_{\hat{L},G}(0,0)$, which is implied by (EC.29). In this case, in the presence of virtual triage, $\forall\ \tilde{\alpha} + \tilde{\beta} \leq 1$, we can show that $f^e_{\hat{H}\tilde{L}} = 0$ with proof by contradiction. Let $\tilde{\mathbf{f}}^{\mathbf{e}} = (f^e_{\hat{L}\tilde{L}}, f^e_{\hat{L}\tilde{H}}, f^e_{\hat{H}\tilde{L}}, f^e_{\hat{H}\tilde{H}})$ denote the equilibrium patient flow in the presence of virtual triage. If $f^e_{\hat{H}\tilde{L}} > 0$, we have $\hat{H}\tilde{L}$ patients find lower patient costs by visiting the ED directly under $\tilde{\mathbf{f}} = (f^e_{\hat{L}\tilde{L}}, f^e_{\hat{L}\tilde{H}}, 0, f^e_{\hat{H}\tilde{H}})$. Hence, we have

$$\begin{aligned} & C_{\hat{H}\tilde{L},E}(f^e_{\hat{L}\tilde{L}}, f^e_{\hat{L}\tilde{H}}, 0, f^e_{\hat{H}\tilde{H}}) < C_{\hat{H}\tilde{L},G}(f^e_{\hat{L}\tilde{L}}, f^e_{\hat{L}\tilde{H}}, 0, f^e_{\hat{H}\tilde{H}}) \\ \Leftrightarrow\ & (1-b_{\hat{H}\tilde{L}})[p_E + w_E Q_E(\lambda_E(f^e_{\hat{L}\tilde{L}}, f^e_{\hat{L}\tilde{H}}, 0, f^e_{\hat{H}\tilde{H}}))] < p_G + (b_{\hat{H}\tilde{L}}w + w_G)Q_G \\ \Rightarrow\ & (1-b_{\hat{H}\tilde{L}})[p_E + w_E Q_E(\lambda_E(0,0,0,0))] < p_G + (b_{\hat{H}\tilde{L}}w + w_G)Q_G, \end{aligned}$$

as we have $\lambda_E(f^e_{\hat{L}\tilde{L}}, f^e_{\hat{L}\tilde{H}}, 0, f^e_{\hat{H}\tilde{H}}) > \lambda_E(0,0,0,0)$. Hence, we have $C_{\hat{H}\tilde{L},E}(0,0,0,0) < C_{\hat{H}\tilde{L},G}(0,0,0,0)$, i.e., $(1-b_{\hat{H}\tilde{L}})[p_E + w_E Q_E(b_{\hat{L}}\lambda_{\hat{L}} + b_{\hat{H}}\lambda_{\hat{H}})] < p_G + (b_{\hat{H}\tilde{L}}w + w_G)Q_G$. This implies $(1-b_{\hat{H}})[p_E + w_E Q_E(b_{\hat{L}}\lambda_{\hat{L}} + b_{\hat{H}}\lambda_{\hat{H}})] < p_G + (b_{\hat{H}}w + w_G)Q_G$, i.e., $C_{\hat{H},E}(0,0) < C_{\hat{H},G}(0,0)$, as we have $b_{\hat{H}} \geq b_{\hat{H}\tilde{L}}$. Hence, we have $C_{\hat{H},E}(0,0) < C_{\hat{H},G}(0,0)$, which contradicts (EC.29). This proves that $f^e_{\hat{H}\tilde{L}} = 0$.

With $f^e_{\hat{H}\tilde{L}} = 0$, we also have $f^e_{\hat{L}\tilde{L}} = 0$ by Lemma EC.1 as $b_{\hat{L}\tilde{L}} \leq b_{\hat{H}\tilde{L}}$. Hence, out of the 12 different equilibrium regions of (EC.13) in the presence of virtual triage, we have $\tilde{\mathbf{f}}^e \in \{(0,0,0,0), (0,0,0,m), (0,0,0,1), (0,m,0,1), (0,1,0,1)\}$ when $\hat{\mathbf{f}}^{\mathbf{e}} = (0,0)$.

The rest of the proof follows the same procedure as the proof of Proposition 1.

Correctness of Proposition 2 under belief-dependent disutility of waiting:

Proposition 2 follows from the fundamental informativeness-volume trade-off of virtual triage recommendations subject to the ROC curve, along with the equilibrium regions as characterized by Proposition 1. Hence, the correctness of Proposition 1 implies the correctness of Proposition 2 in this case.

Correctness of Lemma 3 under belief-dependent disutility of waiting:

(i) When $\tilde{\mathbf{f}}^{\mathbf{e}} \in R_{p,\sim}$, the adoption of virtual triage does not change patient care-seeking behavior. Hence we have $\frac{dC_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\alpha}} = \frac{dC_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\beta}} = 0$.

(ii) In equilibrium region $\tilde{\mathbf{f}}^{\mathbf{e}} = (0,1,1,1)$, we have

$$\begin{aligned} C_s(\tilde{\mathbf{f}}^{\mathbf{e}}) = {} & \lambda_G a_G + \lambda_E a_E + (b_{\hat{L}\tilde{L}}w + w_G)Q_G\lambda_{\hat{L}\tilde{L}} + (w + w_E)Q_E(\lambda_E)b_{\hat{L}\tilde{L}}\lambda_{\hat{L}\tilde{L}} + (b_{\hat{L}\tilde{H}}w + w_E)Q_E(\lambda_E)\lambda_{\hat{L}\tilde{H}} \\ & + (b_{\hat{H}}w + w_E)Q_E(\lambda_E)\lambda_{\hat{H}}. \end{aligned}$$

Then we have

$$\frac{dC_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\alpha}} = b_{\hat{L}}\lambda_{\hat{L}}[a_G + (w+w_G)Q_G] > 0.$$

Moreover, we have

$$\begin{aligned}\frac{dC_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\beta}} =& -(1-b_{\hat{L}})\lambda_{\hat{L}}a_G - (1-b_{\hat{L}})\lambda_{\hat{L}}w_GQ_G + (1-b_{\hat{L}})\lambda_{\hat{L}}a_E + \tilde{\alpha}b_{\hat{L}}\lambda_{\hat{L}}(w+w_E)\frac{\partial Q_E(\lambda_E)}{\partial\lambda_E}(1-b_{\hat{L}})\lambda_{\hat{L}} \\ &+ (1-\tilde{\alpha})b_{\hat{L}}\lambda_{\hat{L}}w\frac{\partial Q_E(\lambda_E)}{\partial\lambda_E}(1-b_{\hat{L}})\lambda_{\hat{L}} + w_E\frac{\partial Q_E(\lambda_E)}{\partial\lambda_E}(1-b_{\hat{L}})\lambda_{\hat{L}}\lambda_{\hat{L}\tilde{H}} + w_EQ_E(\lambda_E)(1-b_{\hat{L}})\lambda_{\hat{L}} \\ &+ (b_{\hat{H}}w+w_E)\frac{\partial Q_E(\lambda_E)}{\partial\lambda_E}(1-b_{\hat{L}})\lambda_{\hat{L}}\lambda_{\hat{H}} > 0,\end{aligned}$$

under the same assumption on a_E as in Section 3.3.

The proof for equilibrium region $\tilde{\mathbf{f}}^{\mathbf{e}} = (0,0,0,1)$ and $\tilde{\mathbf{f}}^{\mathbf{e}} = (0,1,0,1)$ follows the same procedures.

Correctness of Lemma 4 under belief-dependent disutility of waiting:

(i) In equilibrium region $\tilde{\mathbf{f}}^{\mathbf{e}} = (0,m,1,1)$, for a given $\tilde{\alpha}$ and $\tilde{\beta}$ s.t. we have $f^e_{\hat{L}\tilde{H}} \in (0,1)$, $f^e_{\hat{L}\tilde{H}}$ is given by the solution of the FOC of (EC.26):

$$\frac{\partial\Phi(f_{\hat{L}\tilde{H}})}{\partial f_{\hat{L}\tilde{H}}} = -[(1-\tilde{\alpha})b_{\hat{L}} + \tilde{\beta}(1-b_{\hat{L}})]\lambda_{\hat{L}}(p_G + w_GQ_G) - (1-\tilde{\alpha})b_{\hat{L}}\lambda_{\hat{L}}wQ_G + \tilde{\beta}(1-b_{\hat{L}})\lambda_{\hat{L}}[p_E + w_EQ_E(\lambda_E)] = 0.$$

We then have

$$\left.\frac{\partial^2\Phi(f_{\hat{L}\tilde{H}})}{\partial f^2_{\hat{L}\tilde{H}}}\right|_{f_{\hat{L}\tilde{H}}=f^e_{\hat{L}\tilde{H}}} = [\tilde{\beta}(1-b_{\hat{L}})\lambda_{\hat{L}}]^2w_E\frac{\partial Q_E(\lambda_E)}{\partial\lambda_E} > 0,$$

and

$$\left.\frac{\partial^2\Phi(f_{\hat{L}\tilde{H}})}{\partial f_{\hat{L}\tilde{H}}\partial\tilde{\alpha}}\right|_{f_{\hat{L}\tilde{H}}=f^e_{\hat{L}\tilde{H}}} = b_{\hat{L}}\lambda_{\hat{L}}[p_G + (w+w_G)Q_G] > 0,$$

and we have

$$\begin{aligned}\left.\frac{\partial^2\Phi(f_{\hat{L}\tilde{H}})}{\partial f_{\hat{L}\tilde{H}}\partial\tilde{\beta}}\right|_{f_{\hat{L}\tilde{H}}=f^e_{\hat{L}\tilde{H}}} &= (1-b_{\hat{L}})\lambda_{\hat{L}}[-p_G - w_GQ_G + p_E + w_EQ_E(\lambda_E) + \tilde{\beta}f^e_{\hat{L}\tilde{H}}(1-b_{\hat{L}})\lambda_{\hat{L}}w_E\frac{\partial Q_E(\lambda_E)}{\partial\lambda_E}] \\ &> (1-b_{\hat{L}})\lambda_{\hat{L}}[-p_G - w_GQ_G + p_E + w_EQ_E(\lambda_E)] \\ &> (1-b_{\hat{L}})\lambda_{\hat{L}}[-p_G - w_GQ_G + p_E + w_EQ_E(\lambda_H)] \\ &> 0,\end{aligned}$$

under the same assumption on p_E as in Section 3.3.

Meanwhile, we have the equilibrium social cost:

$$\begin{aligned}C_s(\tilde{\mathbf{f}}^{\mathbf{e}}) =& (b_{\hat{L}\tilde{L}}w + w_G)Q_G\lambda_{\hat{L}\tilde{L}} + (w+w_E)Q_E(\lambda_E)b_{\hat{L}\tilde{L}}\lambda_{\hat{L}\tilde{L}} + (1-f^e_{\hat{L}\tilde{H}})(b_{\hat{L}\tilde{H}}w + w_G)Q_G\lambda_{\hat{L}\tilde{H}} \\ &+ (1-f^e_{\hat{L}\tilde{H}})(w+w_E)Q_E(\lambda_E)b_{\hat{L}\tilde{H}}\lambda_{\hat{L}\tilde{H}} + f^e_{\hat{L}\tilde{H}}(b_{\hat{L}\tilde{H}}w + w_E)Q_E(\lambda_E)\lambda_{\hat{L}\tilde{H}} + (b_{\hat{H}}w + w_E)Q_E(\lambda_E)\lambda_{\hat{H}} \\ &+ \lambda_Ga_G + \lambda_Ea_E.\end{aligned}$$

Same as the proof of Lemma 4 (i), it is straightforward to show that $\frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{L}\tilde{H}}}$ is linear and increasing in a_E, $\frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial \tilde{\alpha}}$ is independent of a_E, and $\frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial \tilde{\beta}}$ is linear and increasing in a_E.

(a) Hence, we have

$$\begin{aligned}\frac{dC_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\alpha}} &= \frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial \tilde{\alpha}} + \frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{L}\tilde{H}}}\frac{\partial f^e_{\hat{L}\tilde{H}}}{\partial \tilde{\alpha}} \\ &= [\frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial \tilde{\alpha}}\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^{e2}_{\hat{L}\tilde{H}}} - \frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{L}\tilde{H}}}\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^e_{\hat{L}\tilde{H}}\partial \tilde{\alpha}}]/\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^{e2}_{\hat{L}\tilde{H}}},\end{aligned}$$

where the numerator is linear and decreasing in a_E. This implies that $\exists\ a^u_E$ s.t. we have $\frac{dC_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\alpha}} < 0$ if and only if $a_E > a^u_E$.

(b) Meanwhile, we have

$$\begin{aligned}\frac{dC_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\beta}} &= \frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial \tilde{\beta}} + \frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{L}\tilde{H}}}\frac{\partial f^e_{\hat{L}\tilde{H}}}{\partial \tilde{\beta}} \\ &= [\frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial \tilde{\beta}}\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^{e2}_{\hat{L}\tilde{H}}} - \frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{L}\tilde{H}}}\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^e_{\hat{L}\tilde{H}}\partial \tilde{\beta}}]/\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^{e2}_{\hat{L}\tilde{H}}},\end{aligned}$$

where the numerator is linear and decreasing in a_E. This implies that $\exists\ a^o_E$ s.t. we have $\frac{dC_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\beta}} < 0$ if and only if $a_E > a^o_E$.

The proof for the results with $\tilde{\mathbf{f}} \in \{(0,m,0,1),(0,0,0,m)\}$ follows the same procedures.

(ii) In equilibrium region $\tilde{\mathbf{f}}^{\mathbf{e}} = (0,0,m,1)$, for a given $\tilde{\alpha}$ and $\tilde{\beta}$ s.t. we have $f^e_{\hat{H}\tilde{L}} \in (0,1)$, $f^e_{\hat{H}\tilde{L}}$ is given by the solution to the FOC of (EC.26):

$$\frac{\partial \Phi(f_{\hat{H}\tilde{L}})}{\partial f_{\hat{H}\tilde{L}}} = -[\tilde{\alpha}b_{\hat{H}} + (1-\tilde{\beta})(1-b_{\hat{H}})]\lambda_{\hat{H}}(p_G + w_G Q_G) - \tilde{\alpha}b_{\hat{H}}\lambda_{\hat{H}} w Q_G + (1-\tilde{\beta})(1-b_{\hat{H}})\lambda_{\hat{H}}[p_E + w_E Q_E(\lambda_E)] = 0.$$

We then have

$$\left.\frac{\partial^2 \Phi(f_{\hat{H}\tilde{L}})}{\partial f^2_{\hat{H}\tilde{L}}}\right|_{f_{\hat{H}\tilde{L}} = f^e_{\hat{H}\tilde{L}}} = [(1-\tilde{\beta})(1-b_{\hat{H}})\lambda_{\hat{H}}]^2 w_E \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E} > 0,$$

and

$$\left.\frac{\partial^2 \Phi(f_{\hat{H}\tilde{L}})}{\partial f_{\hat{H}\tilde{L}}\partial \tilde{\alpha}}\right|_{f_{\hat{H}\tilde{L}} = f^e_{\hat{H}\tilde{L}}} = -b_{\hat{H}}\lambda_{\hat{H}}[p_G + (w+w_G)Q_G] < 0,$$

and we have

$$\left.\frac{\partial^2 \Phi(f_{\hat{H}\tilde{L}})}{\partial f_{\hat{H}\tilde{L}}\partial \tilde{\beta}}\right|_{f_{\hat{H}\tilde{L}} = f^e_{\hat{H}\tilde{L}}} = (1-b_{\hat{H}})\lambda_{\hat{H}}[p_G + w_G Q_G - p_E - w_E Q_E(\lambda_E) + (1-\tilde{\beta})(1-f^e_{\hat{H}\tilde{L}})(1-b_{\hat{H}})\lambda_{\hat{H}} w_E \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}].$$

Meanwhile, we have the equilibrium social cost:

$$\begin{aligned}C_s(\tilde{\mathbf{f}}^{\mathbf{e}}) &= (b_{\hat{L}}w + w_G)Q_G\lambda_{\hat{L}} + b_{\hat{L}}(w+w_E)Q_E(\lambda_E)\lambda_{\hat{L}} + (1-f^e_{\hat{H}\tilde{L}})(b_{\hat{H}\tilde{L}}w + w_G)Q_G\lambda_{\hat{H}\tilde{L}} \\ &\quad + (1-f^e_{\hat{H}\tilde{L}})(w+w_E)Q_E(\lambda_E)b_{\hat{H}\tilde{L}}\lambda_{\hat{H}\tilde{L}} + f^e_{\hat{H}\tilde{L}}(b_{\hat{H}\tilde{L}}w + w_E)Q_E(\lambda_E)\lambda_{\hat{H}\tilde{L}} + (b_{\hat{H}\tilde{H}}w + w_E)Q_E(\lambda_E)\lambda_{\hat{H}\tilde{H}} \\ &\quad + \lambda_G a_G + \lambda_E a_E.\end{aligned}$$

Same as the proof of Lemma 4 (ii), it is straightforward to show that $\frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{H}\tilde{L}}}$ is linear and increasing in a_E, $\frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial \tilde{\alpha}}$ is independent of a_E, and $\frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial \tilde{\beta}}$ is linear and increasing in a_E.

(a) Hence, we have

$$\begin{aligned}\frac{dC_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\alpha}} &= \frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial \tilde{\alpha}} + \frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{H}\tilde{L}}}\frac{\partial f^e_{\hat{H}\tilde{L}}}{\partial \tilde{\alpha}} \\ &= [\frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial \tilde{\alpha}}\frac{\partial^2 \Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^{e2}_{\hat{H}\tilde{L}}} - \frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{H}\tilde{L}}}\frac{\partial^2 \Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^e_{\hat{H}\tilde{L}}\partial \tilde{\alpha}}]/\frac{\partial^2 \Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^{e2}_{\hat{H}\tilde{L}}},\end{aligned}$$

where the numerator is linear and increasing in a_E. This implies that $\exists\ a^u_E$ s.t. we have $\frac{dC_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\alpha}} < 0$ if and only if $a_E < a^u_E$.

(b) Meanwhile, we have

$$\begin{aligned}\frac{dC_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\beta}} &= \frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial \tilde{\beta}} + \frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{H}\tilde{L}}}\frac{\partial f^e_{\hat{H}\tilde{L}}}{\partial \tilde{\beta}} \\ &= [\frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial \tilde{\beta}}\frac{\partial^2 \Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^{e2}_{\hat{H}\tilde{L}}} - \frac{\partial C_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{\partial f^e_{\hat{H}\tilde{L}}}\frac{\partial^2 \Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^e_{\hat{H}\tilde{L}}\partial \tilde{\beta}}]/\frac{\partial^2 \Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^{e2}_{\hat{H}\tilde{L}}},\end{aligned}$$

where the numerator is linear and increasing in a_E. This implies that $\exists\ a^o_E$ s.t. we have $\frac{dC_s(\tilde{\mathbf{f}}^{\mathbf{e}})}{d\tilde{\beta}} < 0$ if and only if $a_E < a^o_E$.

The proof for the results with $\tilde{\mathbf{f}}^{\mathbf{e}} \in \{(0,1,m,1),(m,1,1,1)\}$ follows the same procedures.

Correctness of Proposition 3 under belief-dependent disutility of waiting:

It follows directly from the correctness of Proposition 1 and Lemma 4 under belief-dependent disutility of waiting.

Correctness of Lemma 5 under belief-dependent disutility of waiting:

It has the same proof as Lemma 5.

Correctness of Lemma 6 under belief-dependent disutility of waiting:

It follows directly from the correctness of Lemma 4 under belief-dependent disutility of waiting.

Correctness of Proposition 4 under belief-dependent disutility of waiting:

For a given virtual triage tool, suppose $\tilde{\mathbf{f}}^{\mathbf{e}}(r(\tilde{\beta}^*),\tilde{\beta}^*)$ is in the interior of $R_{p,\sim}$.

(i) If $\tilde{\mathbf{f}}^{\mathbf{e}}(r(\tilde{\beta}^*),\tilde{\beta}^*) = (0,0,0,1)$, we have

$$\begin{aligned}C_s(\tilde{\mathbf{f}}^{\mathbf{e}}(r(\tilde{\beta}),\tilde{\beta})) &= (b_{\hat{L}}w + w_G)Q_G\lambda_{\hat{L}} + (b_{\hat{H}\tilde{L}}w + w_G)Q_G\lambda_{\hat{H}\tilde{L}} + (w+w_E)Q_E(\lambda_E)(b_{\hat{L}}\lambda_{\hat{L}} + b_{\hat{H}\tilde{L}}\lambda_{\hat{H}\tilde{L}}) \\ &\quad + (b_{\hat{H}\tilde{H}}w + w_E)Q_E(\lambda_E)\lambda_{\hat{H}\tilde{H}} + \lambda_G a_G + \lambda_E a_E,\end{aligned}$$

with

$$\frac{\partial^2 C_s(\tilde{\mathbf{f}}^{\mathbf{e}}(r(\tilde{\beta}),\tilde{\beta}))}{\partial \tilde{\beta}^2} = r''(\tilde{\beta})b_{\hat{H}}\lambda_{\hat{H}}[a_G + (w+w_G)Q_G] + (1-b_{\hat{H}})\lambda_{\hat{H}}w_E\frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}(1-b_{\hat{H}})\lambda_{\hat{H}}$$

$$+[(b_{\hat{L}}\lambda_{\hat{L}}+b_{\hat{H}}\lambda_{\hat{H}})(w+w_E)+\tilde{\beta}(1-b_{\hat{H}})\lambda_{\hat{H}}w_E]\frac{\partial Q_E(\lambda_E)}{\partial\lambda_E}[(1-b_{\hat{H}})\lambda_{\hat{H}}]^2$$
$$+w_E\frac{\partial Q_E(\lambda_E)}{\partial\lambda_E}[(1-b_{\hat{H}})\lambda_{\hat{H}}]^2$$
$$>0.$$

Hence, the unique $\tilde{\beta}^*$ is given by the solution to the following FOC

$$(a_G+w_GQ_G)[(1-b_{\hat{H}})\lambda_{\hat{H}}-r'(\tilde{\beta}^*)b_{\hat{H}}\lambda_{\hat{H}}]-r'(\tilde{\beta})b_{\hat{H}}\lambda_{\hat{H}}wQ_G= \tag{EC.30}$$
$$[a_E+w_EQ_E(\lambda_E)+(\lambda_Hw+\lambda_Ew_E)\frac{\partial Q_E(\lambda_E)}{\partial\lambda_E}](1-b_{\hat{H}})\lambda_{\hat{H}}.$$

The proof for (ii) and (iii) follows the same procedures.

Correctness of Propositions 5 and 6 under belief-dependent disutility of waiting:

Compared with (EC.24), (EC.30) has an additional term $-r'(\tilde{\beta})b_{\hat{H}}\lambda_{\hat{H}}wQ_G$ on the LHS of the equation and an additional term $\lambda_Hw\frac{\partial Q_E(\lambda_E)}{\partial\lambda_E}(1-b_{\hat{H}})\lambda_{\hat{H}}$ on the RHS of the equation. It is straightforward to show that the directional results of these comparative statics remain the same in this case.

Correctness of Proposition 7 under belief-dependent disutility of waiting:

The results follow directly from Lemma 5 under belief-dependent disutility of waiting. □

References

Bostock N. 2019. Number of GP practices in England falls below 7,000. *GP* (August 26). https://www.gponline.com/number-gp-practices-england-falls-below-7000/article/1525443.

Care Quality Commission. 2022. Urgent and emergency care survey 2020 (July 26). https://www.cqc.org.uk/publications/surveys/urgent-emergency-care-survey-2020.

Clark D. 2021. Average full-time hourly wage in the UK 2021, by region. *Statista* (July 28). https://www.statista.com/statistics/416097/full-time-hourly-wage-uk-by-region/#statisticContainer.

NHS. 2021. Acute Patient Level Activity and Costing, 2019-20. (August 26). https://digital.nhs.uk/data-and-information/publications/statistical/acute-patient-level-activity-and-costing/2019-20.

NHS. 2019. Appointments in General Practice December 2018. (August 26). https://digital.nhs.uk/data-and-information/publications/statistical/appointments-in-general-practice/december-2018.

O'Keeffe C, Mason S, Jacques R, Nicholl J. 2018. Characterising non-urgent users of the emergency department (ED): A retrospective analysis of routine ED data. *PLoS ONE* 13(2):e0192855.

Roughgarden T. 2007. Routing games. Nisan N, Roughgarden T, Tardos É, Vazirani V. *Algorithmic Game Theory* (Cambridge University Press, New York) 461–486.

The King's Fund. 2022. Key facts and figures about the NHS (July 24). https://www.kingsfund.org.uk/audio-video/key-facts-figures-nhs.

The King's Fund. 2022. What's going on with A&E waiting times? (July 28). https://www.kingsfund.org.uk/projects/urgent-emergency-care/urgent-and-emergency-care-mythbusters.