1 Computation

1.1 Algorithms for Model Computation

Given a set of physicians $\{(V_j, R_j(\cdot), \tau_j)\}_{j=1}^J$, punishment function $P(\cdot)$, distribution functions for patient parameters $F(\kappa)$ and $G(\gamma)$, and one search model parameter z, which is λ in the Logit model, β in the sequential model, we develop an algorithm to compute market equilibrium.

Equilibrium aggregates are computed on the basis of Monte-Carlo matrix calculus. Set J as the number of physicians,[1] I as the size of the sample drawn randomly from $F(\kappa)$ and $G(\gamma)$. For both models we define a class which can output a matrix S where each column is a patient's strategy vector S_i following that model, when given as input the arrayed set of physicians' quality and visit cost $\{(V_j, \tau_j)\}_{i=1}^J$, an arrayed set of patients $\{(\kappa_i, \gamma_i)\}_{i=1}^I$, the model parameter z and a given vector of physician strategies $\{\bar{\kappa_j}\}_{j=1}^J$.

We input as "patients" our I samples from $F(\kappa)$ and $G(\gamma)$, then a $J \times I$ matrix U is computed, where each component u_{ji} corresponds to the utility the sampled patient i would get from a visit to physician j.[2] This step is the same for both classes.

What differs between both models is the computation of the matrix S of patient strategis out of the utility matrix U:

- **Implicit search model (Logit):** First, an 'α-matrix' is calculated over matrix U, where each α_{ji} is $e^{\lambda u_{ji}}$ if $u_{ji} > 0$ and 0 if not. Then, for each patient i, that is, for each column, each component s_{ji} of the S matrix takes on the values $s_{ji} = \alpha_{ji} / \sum_{k=1}^J \alpha_{ki}$.
- **Explicit search model (Sequential):** Recall Equation (??) characterizing patient thresholds. We first compute the I-dimensional vector of the sampled patients' respective $\bar{U}_i$. Define U_i as the set $\{u_{ji}\}_{j=1}^J$ of utility patient i recieves from a visit to each physician. In matrix terms U_i would be the ith column of the $J \times I$ matrix U. The operation to compute each $\bar{U}_i$ is the following:

$$\bar{U}_i \equiv \underset{x \in U_i}{\arg\min} \left\| x - \frac{\beta}{1-\beta} \sum_{j=1}^J \left\{ \frac{\mathbb{1}[u_{ji} \geq x] \cdot (u_{ji} - x)}{\mathbb{1}[u_{ji} \geq x]} \right\} \right\|$$

  where the norm $\| \cdot \|$ is defined in $\mathbb{R}$ as simply the absolute value $| \cdot |$. This is to say, for each sampled patient i we evaluate x for each u_{ji} in U_i. In plain words, if patient i where to say: "I will only visit physicians which grant me at least as much utility as physician j", the optimal choice of j and its respective

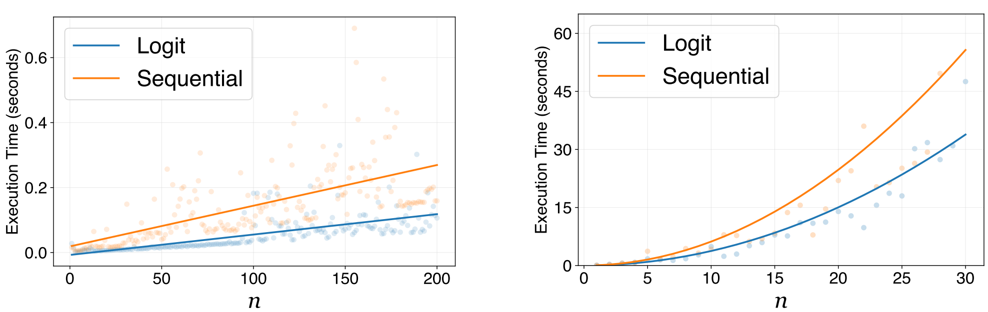
[1] Or, as we'll later interpret it, as the number of bins, where each bin j is a unique combination of $(V_j, \tau_j, R_j(\cdot), P_j(\cdot))$.

[2] As we have defined our matrices $J \times I$ for ease of visualization, we will refer to matrix components as x_{ji} in this section rather than x_{ij} as we do elsewhere in the paper.

u_{ji} would be the minimal in U_i for the evaluation of the absolute value in the left-hand side of (1.1). This is computationally less intensive than seeking to compute the exact root of Equation (??), which would be a redundant exercise, because there's a discrete number of physicians above that mark, and selecting instead to use "the lowest u_{ji} above the root of (??)" as threshold instead of the root proper would result in the same vector of strategies S_i for each patient.[3]

Once having computed our I-dimensional vector of patients' $\bar{U}_i$, we evaluate column-wise the binary operator $\mathbb{1}\{u_{ji} \geq \bar{U}_i\}$ over matrix U, and the $J \times I$ matrix S of patient strategies takes on the values $s_{ji} = \mathbb{1}\{u_{ji} \geq \bar{U}_i\} / \sum_{k=1}^{J} \mathbb{1}\{u_{ki} \geq \bar{U}_i\}$.

The S matrix in both cases is computed in linear time, $\mathcal{O}(n)$ in big-O notation, where input size n is the number of physicians J. Both consist of a series of ufunc or vectorized operations, where the inclusion of one more physician introduces a new row with I elements on which operations are performed at each stage. The additional step of calculating the $\bar{U}_i$ vector makes the execution of the 'explicit' search model relatively slower (see Panel Figure 1a).



(a) Linear polynomial fit of execution time for S matrix in both models

(b) Quadratic polynomial fit of execution time for $\bar{\kappa}^*$ vector in both models

Figure 1: Execution time comparison between the *implicit* and *explicit* search models

The steps following the computation of the S matrix are the same for both models. The J-dimensional vector Q of each physician's expected demand Q_j is achieved through the row-wise summation of all patient's demand for j, i.e. performing $\sum_{i=1}^{I} s_{ji}$ for each j. The vector X of each physician's expected sick leaves issued X_j performs that same summation, conditional on the patients' κ_i being above or at physician j's chosen threshold $\bar{\kappa}_j$, that is, $\sum_{i=1}^{I} \mathbb{1}[\kappa_i \geq \bar{\kappa}_j]\, s_{ji}$. Both sums are normalized to fit the actual number of patients in the physician-patient market.

For a given vector $\bar{\kappa}$, computing patient strategies, physician aggregates and then

[3]Granted, this is not strictly true, as in our formulation a 'u_{ji}' may be selected as threshold which is actually below the root proper in $\mathbb{R}$, but closer to it than the nearest one above it. The effect of this "estimation noise" on our overall results is negligible.

physicians' utility—defined as $\{R_j(Q_j) - P(X_j)\}$—entails a sequence of calculations done in linear time. However, the calculation of the *equilibrium* vector $\bar{\kappa}^*$ is performed in quadratic time $\mathcal{O}(n^2)$, as it requires each of the aforementioned steps to be done some x number of times *per* physician, such that the inclusion of an additional physician increases the number of operations required *per* physician as well as the amount of physicians whose $\bar{\kappa_j}$ needs to be computed. See Panel Figure 1b.

The algorithm to find the equilibrium J-vector $\bar{\kappa}^*$ of physician strategies is as follows:

i Input an initial guess $\bar{\kappa}^0$ for physician thresholds.

ii For each physician, we fix the the value of $\bar{\kappa}_{-j}$, the threshold values of the other $J-1$ physicians, and compute equilibrium aggregates and physician j's utility for different choices of $\bar{\kappa}_j$ across a grid. In particular, $U_j(\bar{\kappa}_j, \bar{\kappa}_{-j})$ is computed for each decimal step between 0 and $\bar{\kappa}_{\max}$, the maximum threshold physicians may choose.

iii The choice of threshold k_j which rendered j the most utility in the previous step is selected, and a grid is set-up spanning the 19 centesimal values in $[k_j - 0.05, k_j + 0.05]$,[4] each value therein input as physician j's threshold $\bar{\kappa}_j$ to compute U_j again, and the value within the grid which maximizes utility is chosen as $\bar{\kappa}^1_j$.

iv Having performed the previous step for all J physicians, we input as a new guess the vector $\bar{\kappa}^1 = \{\bar{\kappa}^1_1, ..., \bar{\kappa}^1_J\}$ to run steps ii and iii again. This defines an equilibrium-searching loop $\bar{\kappa}^n = \Phi(\bar{\kappa}^{n-1})$, and the loop is concluded when a fixed point is found, that is, the vector $\bar{\kappa}^*$ such that $\bar{\kappa}^* = \Phi(\bar{\kappa}^*)$.

v Optionally, having found a two decimals fixed point $\bar{\kappa}^*$, there's an algorithm in place to find an x-decimal fixed point $\bar{\kappa}^{*x}$. It starts by running a modified version of step iii on the two-decimal solution, setting up a grid of the 19 *millesimal* values in $[\bar{\kappa}^*_j - 0.005, \bar{\kappa}^*_j + 0.005]$ for each physician j,[5] selecting that which renders highest utility for each, and inputting this new vector as guess to run this step again, and so on until the fixed point $\bar{\kappa}^{*3} = \Phi(\bar{\kappa}^{*3})$ is found.

   This goes on like this, calculating the t decimals solution out of the $t-1$ decimals solution by setting up $10^{-(t-1)}$-sized grids for each physician in each iteration, until the specified amount of decimals wanted from the solution vector $\bar{\kappa}^{*x}$ is reached.

On the one hand, having to perform several operations *per* physician is quite computationally intensive and runs the cost of quadratic execution time. On the other hand, altough each iteration takes some time, for the right parameters the algorithm is quite efficient in the number of iterations needed for convergence, usually taking between 2 and 5. The option to find fixed points to $n+1$ decimal places incurs

[4] If k_j is 0 (or $\bar{\kappa}_{\max}$), only the 10 centesimal values above (below) and at k_j are computed.

[5] For $\bar{\kappa}^*_j = 0$ or $\bar{\kappa}_{\max}$ a similar logic to the footnote above follows.

progressively smaller execution time costs compared to n decimals, as the initial guess for $n+1$ decimals becomes increasingly accurate. See Figure 2.

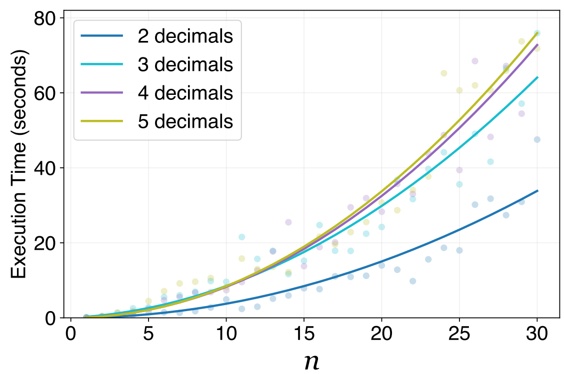


Figure 2: Quadratic polynomial fit of execution time of $\bar{\kappa_j}^*$
for different decimal approximations in Logit model

Once the equilibrium vector $\bar{\kappa_j}^*$ is calculated, it may be used to compute the equilibrium aggregates Q_j and X_j for all physicians, as well as physician utility.

1.2 Illustrative Examples

We use the following parametrization for illustration purposes:

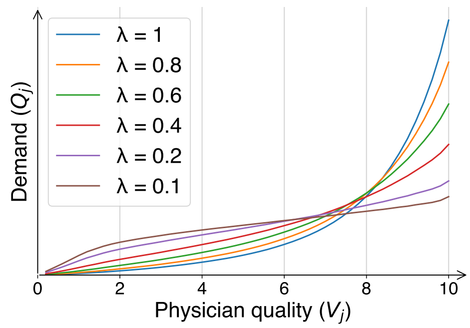
$F(\kappa)$	=	$U[0,1]$	$R(x)$	=	$1.5x$
$G(\gamma)$	=	$U[0,1]$	$P(x)$	=	$\frac{1}{100}x^2$
V_j	$\sim$	$U[0,10]$	τ	=	1

In particular, we consider $J = 50$ physicians whose V_j is evenly spaced in increments of 0.2 between 0.2 and 10. They share the same $R_j(\cdot)$ and τ_j.

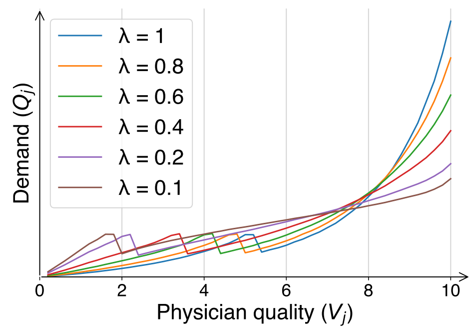
For the Logit model, we present some results for different parametrizations of λ in Figure 3. Observe first Panel 3a, graphing patient demand Q_j when all physicians choose $\bar{\kappa_j} = 0$: when $\lambda = 0.1$, patient demand is more evenly distributed across physicians of different quality, and as it increases towards 1 it becomes more concentrated at physicians of higher quality.

Panel 3b shows patient demand at the equilibrium $\bar{\kappa_j}*$ of each physician. For each level of λ, there is more patient demand at lower levels of V_j and then a drop-off, followed by a subsequent rise. The drop-off represents the transition from a $V_{mid-low}$ equilibrium to $V_{mid-high}$ (recall Figure ??), when physicians start opting for a higher $\bar{\kappa_j}$ in an inner solution. Before the drop-off, low V_j physicians opt for a threshold equal to 0, maximizing their possible patient demand, and then briefly slightly above 0. After a certain level of V_j, physicians find it better to be strict and decrease their patient demand slightly but scaling down their issued sick leaves by a larger measure, fulfilling the FOC condition (??).

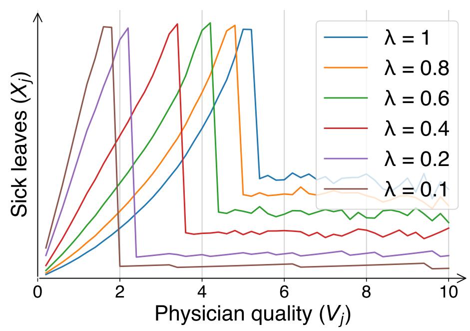
We can observe this behavior in Panel 3c. At each λ there's a significant drop-off in sick leaves at a certain V_j once physicians start preferring the stricter inner solution. Notice how after the drop, sick leaves remain flat, because physicians share $R_j(\cdot)$ and $P(\cdot)$ in our example, and in particular $R_j(\cdot) = r_j$, so as patient demand Q_j increases with V_j, all physicians must do in order to remain at the FOC physicians is increase their threshold $\bar{\kappa_j}$ to remain at the same level of X_j. We can see how $\bar{\kappa_j}^*$ increasies in V_j in Panel 3d.



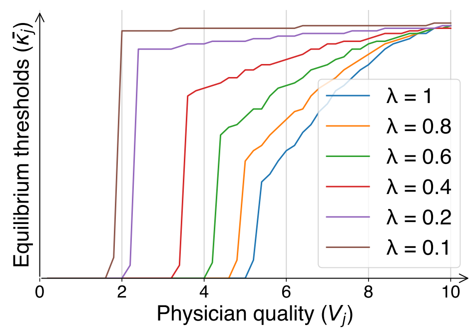
(a) Patient demand Q_j
by physician quality, $\bar{\kappa} = \vec{0}$



(b) Patient demand Q_j
by physician quality, equilibrium $\bar{\kappa}^*$



(c) Sick leaves issued X_j
by physician quality, equilibrium $\bar{\kappa}^*$



(d) Equilibrium $\bar{\kappa_j}^*$ by physician quality

Figure 3: Physician aggregates and strategies for different values of λ
in the *implicit* search model

Let's now look at the sequential model with the same parametrization, for different values of β (Figure 4). Something to not is that β is restricted to values within $(0, 1)$. As $\beta \to 1$, Q_j behaves similarly to when $\lambda \to \infty$ in the Logit model, which is why the concentration of patients at high V_j's with $\beta = 0.99$ is much more pronounced in Panel 4a than it is with $\lambda = 1$ in the previous Figure.

The shape of the demand curve points to two disimilarities from the Logit model:

- For physicians of low enough V_j demand is strictly 0 rather than just low, as they don't afford to any patient utility over their threshold $\bar{U}_i$.
- At high values of V_j the demand curve is concave rather than convex like before. This is because the formulation of the 'explicit' search model doesn't continuously (and exponentially) reward higher values of V_j, it only registers whether V_j is enough to have u_{ij} over patients' reserve utility $\bar{U}_i$. Past a certain point, most patients' $\bar{U}_i$ (which can be seen in Panel 4e) lay below the u_{ij} a certain physician j affords, and incremental gains to even higher V_j are low, hence the concavity.

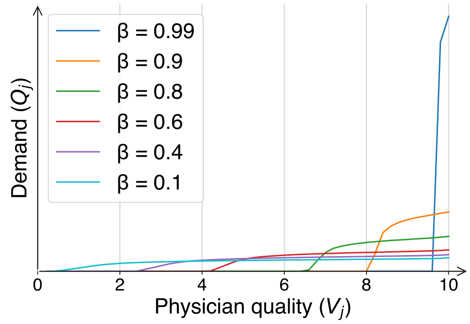
Panel 4b is not particularly illuminating, but Panel 4c is, and can show the difference between 4b and 4a. Panel 4c displays sick leaves granted X_j as a function of physician quality V_j for different values of the model parameter β, at an equilibrium physician strategies vector $\bar{\kappa}^*$. Graphically, it exhibits two peaks in X_j, which grow more pronounced and further apart from each other and the origin as β increases.

The first peak in each curve is composed of physicians who manage to get positive clientele from leaving their threshold $\bar{\kappa_j}$ at 0. The drop-off from the top signals the transition to an inner solution, where the patient base of physician j decreases for the sake of more starkly decreasing sick leaves issued (X_j), fulfilling the FOC condition (??). The second peak has a smoother lead-up and descent from it, in which physicians make the trade-off of gaining patients through lower $\bar{\kappa_j}$ but requiring sick leave lenience. The peak itself represents the tipping point, after which physicians gain a significant patient base just by virtue of their high V_j and can afford to cut back on sick leave issuance (i.e. increase strictness) with little downside to their overall Q_j. Then there's a final peak seen specifically in $\beta = 0.99$, which results out of the high concentration of patients around the maximum V_j, such that patients whose $\kappa_i \geq \bar{\kappa}_{\max}$, previously more spread out across physicians, visit the highest V_j's exclusively.

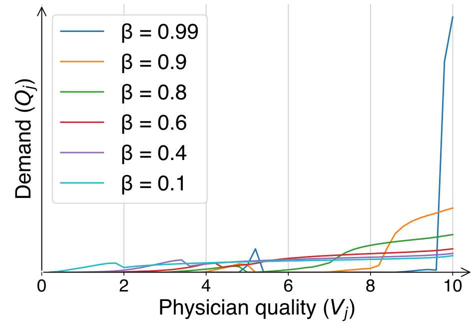
As β increases, so too the patients' reserve utility $\bar{U}_i$ as κ_i grows, as can be seen in Panel 4e. These higher values concentrate demand at higher values of V_j and also shorten the window in which patient demand may be gained through sick leave issuance (as opposed to just by virtue of a high V_j), which is why the peaks become "slimmer".

Despite these idiosyncracies, we can see in Panel 4d that physician thresholds $\bar{\kappa_j}$ grow with V_j in a very similar fashion to how they did in the previous model.

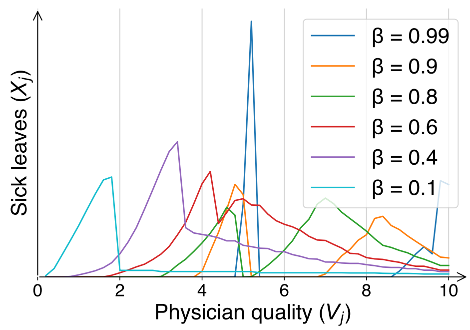
Some more figures from these simulations can be seen in the Appendix ??. They show how physician utility is strictly increasing in V_j (meaning the drop-offs in patient demand after transitioning to inner solutions of $\bar{\kappa_j}$ are strategic and not a detriment to utility), and also highlight the difference in Q_j and X_j between the equilibrium values of $\bar{\kappa}^*$ and other physician strategy vectors.



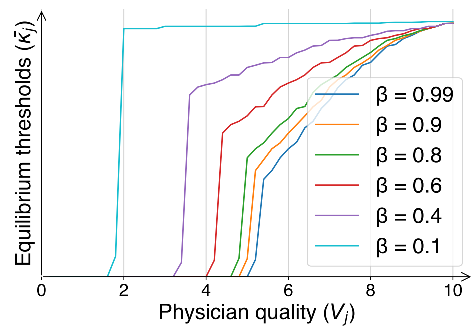
(a) Patient demand Q_j
by physician quality, $\bar{\kappa} = \vec{0}$



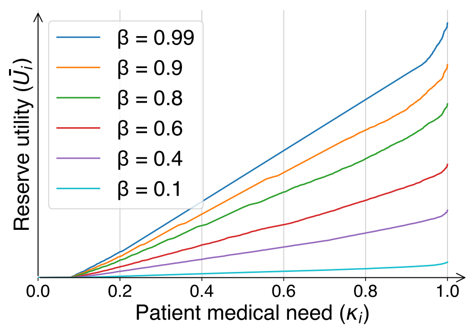
(b) Patient demand Q_j
by physician quality, equilibrium $\bar{\kappa}^*$



(c) Sick leaves issued X_j
by physician quality, equilibrium $\bar{\kappa}^*$



(d) Equilibrium $\bar{\kappa_j}^*$ by physician quality



(e) Patient reserve utility $\bar{U}_i$
by medical necessity (κ_i), equilibrium $\bar{\kappa}^*$

Figure 4: Physician aggregates and strategies for different values of β
in the *explicit* search model