PLOS ONE

RESEARCH ARTICLE

A flexible method for optimising sharing of healthcare resources and demand in the context of the COVID-19 pandemic

Lucas Lacasa[1,2]*, **Robert Challen**[3,4], **Ellen Brooks-Pollock**[5], **Leon Danon**[6,7]

1 School of Mathematical Sciences, Queen Mary University of London, London, United Kingdom, **2** Institute for Cross-Disciplinary Physics and Complex Systems IFISC (UIB-CSIC), Palma de Mallorca, Spain, **3** EPSRC Centre for Predictive Modelling in Healthcare, University of Exeter, Exeter, Devon, United Kingdom, **4** Taunton and Somerset NHS Foundation Trust, Taunton, Somerset, United Kingdom, **5** Bristol Medical School: Population Health Sciences, University of Bristol, Bristol, United Kingdom, **6** Data Science Institute, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, United Kingdom, **7** The Alan Turing Institute, British Library, London, United Kingdom

* l.lacasa@qmul.ac.uk

Check for updates

OPEN ACCESS

Citation: Lacasa L, Challen R, Brooks-Pollock E, Danon L (2020) A flexible method for optimising sharing of healthcare resources and demand in the context of the COVID-19 pandemic. PLoS ONE 15 (10): e0241027. https://doi.org/10.1371/journal.pone.0241027

Editor: M. Sohel Rahman, Bangladesh University of Engineering and Technology, BANGLADESH

Received: July 13, 2020

Accepted: September 22, 2020

Published: October 21, 2020

Data Availability Statement: All data and codes are available at https://github.com/lucaslacasa/loadsharing.

Funding: The authors report three grants from Engineering and Physical Sciences Research Council (EPSRC) which contributed to the study: LL EP/P01660X/1 EPSRC, RCh EP/N014391/1 EPSRC, and LD EP/N510129/1 EPSRC.

Competing interests: The authors have declared that no competing interests exist.

Abstract

As the number of cases of COVID-19 continues to grow, local health services are at risk of being overwhelmed with patients requiring intensive care. We develop and implement an algorithm to provide optimal re-routing strategies to either transfer patients requiring Intensive Care Units (ICU) or ventilators, constrained by feasibility of transfer. We validate our approach with realistic data from the United Kingdom and Spain. In the UK, we consider the National Health Service at the level of trusts and define a 4-regular geometric graph which indicates the four nearest neighbours of any given trust. In Spain we coarse-grain the healthcare system at the level of autonomous communities, and extract similar contact networks. Through random search optimisation we identify the best load sharing strategy, where the cost function to minimise is based on the total number of ICU units above capacity. Our framework is general and flexible allowing for additional criteria, alternative cost functions, and can be extended to other resources beyond ICU units or ventilators. Assuming a uniform ICU demand, we show that it is possible to enable access to ICU for up to 1000 additional cases in the UK in a single step of the algorithm. Under a more realistic and heterogeneous demand, our method is able to balance about 600 beds per step in the Spanish system only using local sharing, and over 1300 using countrywide sharing, potentially saving a large percentage of these lives that would otherwise not have access to ICU.

1 Background

The outbreak of COVID-19 [1], the disease caused by the novel coronavirus SARS-CoV-2, detected in China in December 2019 [2], has become pandemic and continues putting national health systems of different countries into significant levels of stress [3–6] (see [7] and references therein for a detailed overview). Either during the first or successive epidemic waves, the intensive care unit (ICU) demand of several hospitals might surpass their nominal capacity in particular regions in several countries, as has already happened in Italy or Spain [8]. The shortage of sanitary resources is unlikely to be limited to ICU units or ventilators, and other resources will face similar challenges, either during the first surge or in subsequent waves.

In the COVID-19 pandemic, demand for intensive care is not uniform across a country. Epidemic outbreaks can take place in different parts of a country and this can lead to substantial variations of demand both through space and time. Some hospitals may receive substantial numbers of patients early in an outbreak, whilst others may be only mildly affected. This demand heterogeneity opens the possibility of balancing the load of patient admissions such that excessive demand is re-routed to the places which have spare capacity. The clinical need for such a system was evidenced by a spontaneous initiative that took place in Madrid (Spain) in early April 2020 [9], when the Spanish capital was suffering a significant surge of COVID-19 cases. The intensive care lead of 76 hospitals in Madrid created an informal WhatsApp group to share daily information on the ICU demand and availability, with the goal of transferring patients across hospitals in the hope that the network could provide adequate treatment to all patients. Other tactical load balancing actions have been recently proposed in the US [10, 11]. Of course, this is an example of a quick, crisis emergency action, but as soon as multiple centres are overwhelmed, the demand pattern becomes very complex, and the number of possible transfer combinations increases exponentially in a graph with N nodes, the number of possible ways in which each of the nodes can transfer load to other nodes increases exponentially with N. Without a principled and organic approach to patient transfer it is possible to end up worsening the situation.

A natural question is thus, given the available resources of a national health system covering a specific region, whether there exist a principled, adaptive and *optimal* way of balancing the demand across hospitals by which a maximal number of patients can receive adequate treatment even during a pronounced epidemic peak, thereby relieving the stress of the whole system. Furthermore, the need to match intensive care supply to patient demand in different parts of the world is indeed currently urgent in areas of the world experiencing serious outbreaks. Here we address such questions by designing and implementing a simple and flexible load sharing procedure which can help to alleviate the level of stress that healthcare systems experience in a systematic way.

The methodology is in principle tailored to address the COVID-19 pandemic situation, but otherwise is general and thus applicable in different countries, at different resolution levels, and for any resource constrained clinical service. The method uses graph-embedded load balancing technology coupled with a simple optimisation kernel, and we showcase its usability by testing it on the UK National Health Service (NHS) and the Spanish health system as examples with different spatial granularity. Note that graph-embedded load balancing [12, 13] has been mainly explored in Computer Science (CS), usually taking a "vertex perspective" for graphical computation with the aim of achieving a centralised solution to load allocation, subject to locality and availability constraints [14]. Interestingly, this line usually relates to minimise large-scale computational efforts, rather than actually sharing physical resources. A similar approach overlaps with the so-called Social Choice Theory of allocating goods among a set of agents under some constraints that overlaps economics, social sciences and computer science [15–18]. More closely related to our approach is the concept of dynamic load balancing, theoretically explored in the CS literature recently [19, 20]. Similar approaches have also been investigated in the Operations Research (OR) literature, and in particular the topic of location theory is relevant here as well [21, 22]. All these provide a reasonably mature mathematical framework which we subsequently rely on. Indeed, here we build on conceptually similar approaches although we focus on a healthcare network where resources to be shared consist of ICU beds or ventilators, within the context of the COVID-19 pandemic. After presenting the algorithmic modelling, as a proof of concept we apply our framework to two realistic cases at different spatial resolutions: the United Kingdom's full NHS trust network, and the Spanish contact network between autonomous communities. We focus on the problem of ICU demand, propose and implement a routine strategy to transfer resources across the network, and demonstrate that it is capable of useful and relevant outcomes.

2 Materials and methods

2.1 Transfer networks

We first define the network over which load and resources can be shared. Demand and capacity data –and thus, load sharing– can be coarse-grained at different resolutions: hospitals, postcodes, trusts, and broader regions. In this paper, we consider two levels of resolution: NHS trusts (UK) and autonomous communities (Spain).

2.1.1 NHS trust network. We coarse-grain data for the UK at the level of trusts, as the main units of NHS organisation. We have $N = 141$ trusts across the UK, where each trust corresponds to a conglomerate of m hospitals. For each trust, we define a single central position by finding the centroid of the polygon whose vertices are the hospitals belonging to that trust. While spatial coordinates are given in terms of latitude and longitude, we make a small angle approximation and interpret latitude and longitude as cartesian coordinates Since $\sin\alpha \approx \alpha$ when $\alpha \ll 1$, it is easy to see that an increase of a small angle α leads to a linear increase $h \approx R\alpha$ where R is a constant. In particular, under this approximation the centroid coordinates of trust reduces to the arithmetic mean of the coordinates of each hospital in the trust

$$(x, y)_i = \left(\frac{1}{m}\sum_{j=1}^{m}\mathrm{lat}(j), \frac{1}{m}\sum_{j=1}^{m}\mathrm{lon}(j)\right)_i \quad (1)$$

In the event that the net capacity c_i of each hospital is also available, then instead of computing the centroid, one can compute the center of mass by appropriately weighting the contribution of each hospital:

$$(x, y)_i = \left(\frac{1}{m}\sum_{j=1}^{m}\mathrm{lat}(j)\bar{c}_j, \frac{1}{m}\sum_{j=1}^{m}\mathrm{lon}(j)\bar{c}_j\right)_i, \quad (2)$$

where $\bar{c}_j = c_j / \sum_{k=1}^{m} c_k$ is the normalised capacity of hospital j we normalise it such that x and y still have dimensions of length, and m is the number of hospitals in trust i. The distance between two trusts corresponds to the Euclidean distance between the centroids or the centers of mass if more precise coordinates need to be used, instead of these we can use Haversine formula.

$$d_{ij} = ||(x, y)_i - (x, y)_j||_2 = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \quad (3)$$

In our case we do not have information on the actual ICU capacity of each specific hospital within a given trust, so we choose to use centroids instead of centers of mass.

Once we have defined the location each of the 141 NHS trusts, we assign a vertex to this spatial location and proceed to tessellate this set. We build a regular geometric graph with degree $k = 4$, where each vertex i is connected to the four closest vertices according to the distance d_{ij} defined above. The resulting graph is depicted in panel a of Fig 1. Each trust will only be allowed to transfer patients or resources to the trusts in their topological neighborhood, modelling the fact that transfers only take place between close trusts.

2.1.2 Spain's autonomous community networks. Spain has a decentralised health system, so we consider that load sharing between hospitals can only take place within each

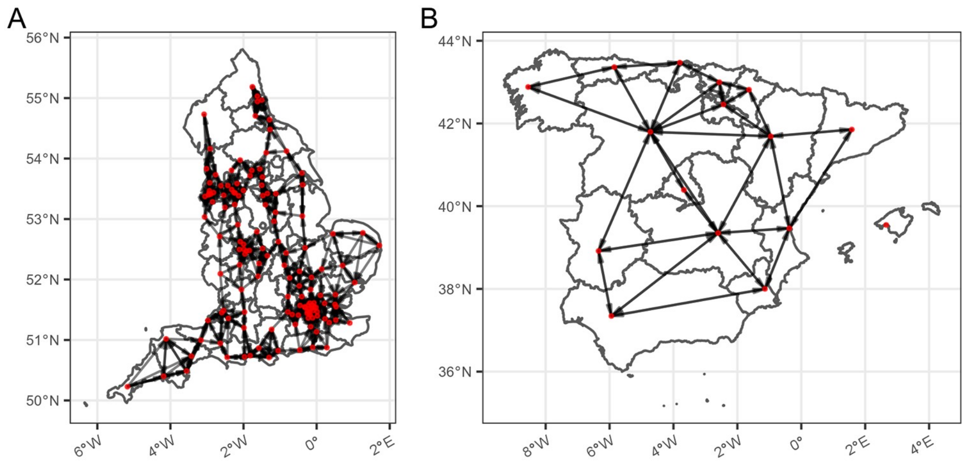

Fig 1. (a) NHS trust network, where nodes are NHS trusts and the network is a 4-regular geometric graph tesselating the set of nodes, i.e. any node is connected to its four closest nodes. (b) Spanish contact network, where nodes are autonomous communities and two nodes are linked if the communities share a common border. In this network we have discarded both archipelagos (Canary islands and Balearic islands) as transfer between these and mailand is not realistic. We have not plotted Canary islands as it is off-scale. Background images have been generated using Natural Earth [23] and GADM [24].

https://doi.org/10.1371/journal.pone.0241027.g001

autonomous community (intra-community). Because of that, as a second example here we will consider load sharing at the inter-community level. The network therefore has $N = 17$ nodes, each of them characterising a certain autonomous community. We will consider two different networks: a contact network and a fully connected network. In the contact network, two nodes are connected if the respective autonomous communities share a border. This makes this network more heterogeneous than the NHS trust network, where the maximal degree is $k = 9$ (for the community of *Castilla y Leon*). We assume that load sharing can only be performed by road, meaning that this network is disconnected; as two autonomous communities are not part of mainland Spain (Balearic islands and Canary islands). So, we only consider the large connected component, formed by $N = 15$ nodes with varying degree $2 \leq k \leq 9$. The resulting graph is depicted in panel (b) of Fig 1 (note that Canary islands have not been drawn because they are an off-scale disconnected node). Distance is not a constraint in this case.

Additionally, we will also consider a fully connected network formed of $N = 15$ nodes on the mainland, where all possible links are present. This models the ideal situation where the transfer of patients/ventilators between any two autonomous communities is possible, e.g. using the national train network, as already proposed [25]. The Balearic islands and the Canary islands are, again, not part of this network.

2.2 Local load sharing model

The basic architecture of the local load sharing model is depicted in Fig 2. For each node, the algorithm takes `projected-ICU-demand` data (aggregated at the NHS trust level or the autonomous community level, depending on the example), matches with its `baseline-ICU-capacity` (aggregated number of ICU beds or ventilators which are available to be used at that specific time by the trust itself or others), and generates a `local-stress` value for each node accordingly:

$$[\mathtt{local-stress}] = [\mathtt{projected-ICU-demand}] - [\mathtt{baseline-ICU-capacity}].$$

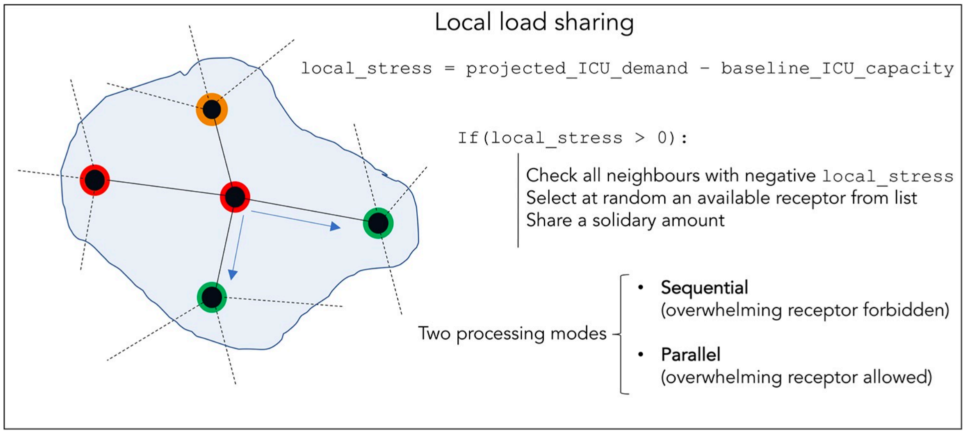

Fig 2. Scheme of the local load sharing model. Red and orange denotes an overwhelmed unit with varying levels of stress, green denotes a unit with capacity.

https://doi.org/10.1371/journal.pone.0241027.g002

For those nodes where the local stress is positive (meaning that demand surpasses the available capacity and thus there is a need to load-share), the algorithm explores which neighboring nodes (extracted from the topological neighborhood of the node under analysis) could accept a transfer. A transfer is possible if two conditions are met: (i) there is at least a node in the neighborhood of the origin node whose `local-stress` is negative (i.e. the receptor node has freely available space after its own demand is met), and (ii) the physical distance between the origin and the receptor node is smaller than a certain upper bound $d_{\max}$. This maximum distance models at the same time several possible constraints, e.g. the fact that ICU patients can only be outside an hospital for a limited amount of time or that effective transfers require the distance between origin and receptor to be small.

Once the receptor is chosen, a 'solidary' load is shared to the receptor. As a rule of thumb, we choose this load to be either 50% of the excess capacity of the receptor (that is, of |`local-stress`|, i.e. capacity after having met internal demand), or the total excess demand of the origin trust, whichever is smaller. The rationale for this definition is based on the fact that receptor nodes would probably be willing to accept to release only a percentage of their capacity while they keep another percentage in anticipation for future internal demand. The exact number (50% in this paper) is flexible and different practical implementations can assume different percentages.

2.3 Sequential vs parallel update

In this work we have systematically considered two alternative algorithmic updates, mirroring the fact that the local decision of a given node to transfer or not can be carried out either **sequentially** or in **parallel** for the rest of the nodes.

Let us first discuss the parallel mode. In this case, the `projected-ICU-demand` of all N nodes are updated in the new step independently from each other, and transfer between nodes take place independently. Note that this update can be problematic in practice, as e.g. two nodes might decide independently to transfer part of their excess load to a third node, what might mean overwhelming the third node. On the other hand, as we will show below, configurations which are globally more optimal are available using this mode.

In sequential update, the `projected-ICU-demand` of each node is sequentially updated after each local load share is performed. In other words, we consider the update of all

N nodes in order. That means, for instance, that in a given step of the algorithm all N nodes are updated in order, such that the node p is updated taking into account the old status of nodes $p + 1, p + 2, \ldots, n$ (that haven't been updated yet) and the *new* status of nodes $1, 2, \ldots, p - 1$. Sequential updates have the positive implication that no receptor will be overwhelmed from the *simultaneous* load sharing of different nodes.

Incidentally, the algorithmic difference between the sequential and the parallel update mode is similar to the difference between Jacobi and Gauss-Seidel numerical schemes when solving systems of linear equations. In practice, the code we have implemented asks the user to choose which processing mode is used (sequential or parallel).

2.4 Random search optimisation

The basic local load sharing model is run for all nodes (NHS trusts or autonomous communities), and as a result a possible load sharing configuration is extracted, consisting of the specified origin and destination of all the packets of ICU patients shared:

`Trust i shared x loads to trust j`

To assess the global impact of such load sharing configuration, we define the global stress of the whole system

$$\texttt{global} - \texttt{stress} = \sum_j \Theta[\texttt{local} - \texttt{stress}(j)], \tag{4}$$

where the sum runs over all trusts j, and $\Theta(x)$ is a rectified linear unit (ReLu), defined by $\Theta(x) = x$ if $x > 0$ and zero otherwise. So essentially `global-stress` counts the total demand of ICU units in excess of capacity, in all those trusts which are projected to be overwhelmed.

Now, in the event there is a node with positive local stress (i.e. with an excess of demand and a need to transfer load) and more than one candidate receptor, how to choose the adequate node where the load is shared to? A natural choice would be to follow a *majority heuristic*, i.e. transfer the share to the receptor with largest availability, i.e. with lowest local stress. However this choice does not always yield solutions which are globally optimal. This situation is illustrated in Fig 3, where the majority heuristic would suggest that node '2' should transfer its excess load to node '3' as its local stress is lower than the one at node '1'. However doing this precludes node '4', also in need of load-sharing, to transfer loads to node '3'. In this case the optimal solution (in the sense of minimising `global-stress`) would be that node '2' load-shares to node '1', thus enabling node '4' to load-share to node '3'.

The example above is just a cartoon in an extremely simple graph. In more practical applications where transfer networks are more complex, the number of possible configurations is much larger and thus the issue is even more acute. To address this issue, here we implement a so-called random search optimisation approach, which consists in two steps. First, if more than one receptor is available for transfer in the topological neighborhood of a given node, then the algorithm selects the receptor at random. Second, once the algorithm chooses the configuration for all N nodes, it is then re-run 10^5 times, such that in each realisation a

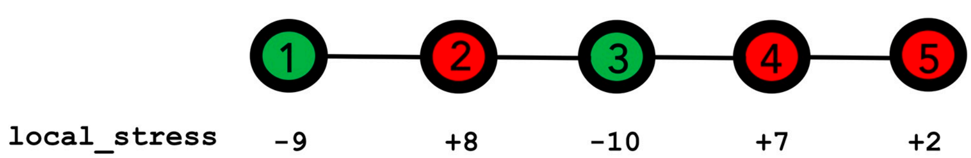

Fig 3. Cartoon of a simple chain graph where some nodes require to load share and a trivial majority heuristic rule provides a suboptimal solution.

https://doi.org/10.1371/journal.pone.0241027.g003

different configuration is stochastically chosen. In this way the algorithm stochastically samples the search space. The quasi-optimal run with the lowest `global-stress` is finally retained. While this approach is computationally scalable (and can be implemented to be run in real-time in practical cases), the solution is guaranteed to be at least a local optimum, but one cannot discard that other configurations not sampled might have lower `global-stress` since search is not exhaustive.

2.5 Input variables

Now we briefly discuss the main input data required to run the local load sharing model:

- `projected-ICU-demand`: This is an input data to the algorithm. In practice this demand is either informed by some surveillance protocol or otherwise modelled, as it could be estimated following a complex multi-step flow [12], which can be summarised as follows:

    1. The projected number of new infections next week: This quantity can be informed in the first place from an epidemiological model [8, 26] which provides predicted numbers of contagion at different spatial resolutions. Alternatively, or in the absence of such a model, it could be estimated from various sources of data [27] including prescription data [28] or through direct questionnaires Data can be retrieved and processed from apps and other surveillance systems such as centralised webpages where citizens submit their symptoms. These questionnaires, coupled with a classification algorithm, can estimate the number of latent infected people in a certain region or postcode. A post-processing of these numbers is then carried out, taking into account (i) age demographics and (ii) associated infection-to-ICU rates.

    2. The projected number of patients already in the hospital which progress to ICU by next week: this number is estimated from real data of hospital admissions and average admission-to ICU likelihood.

    3. The projected number of patients already in ICU this week which will still require ICU next week: this number takes into account both the fatality ratio and the estimated discharge time.
    As a proof of concept, in this work we assume different types of artificial ICU demands (uniform and heterogeneous distributions) in the UK case, whereas in the spanish case we consider realistic demand as of 30th March 2020, i.e. during the first epidemic wave. We test how the load sharing algorithm performs under different demands.

- `baseline-ICU-capacity`: This list is extracted from public available databases [29, 30]. In the case of autonomous communities these quantities already have into account some enhancement provided by surge capacity [30], whereas in the case of NHS trusts we only use baseline data, so we expect such capacity to be significantly increased in practice.

3 Results

3.1 Single-share in the UK NHS trust network

In this first section we assume that each trust can only submit a unique load to a unique receptor trust, to be selected randomly from the trust's topological neighborhood.

3.1.1 Stress test with fixed, uniform-load ICU demand. As an initial illustration, we first analyse a stress test case where `projected-ICU-demand` is artificially set to a uniform value of 20 ICU beds per trust (i.e. all trusts receive a demand of 20 beds) whereas we set all

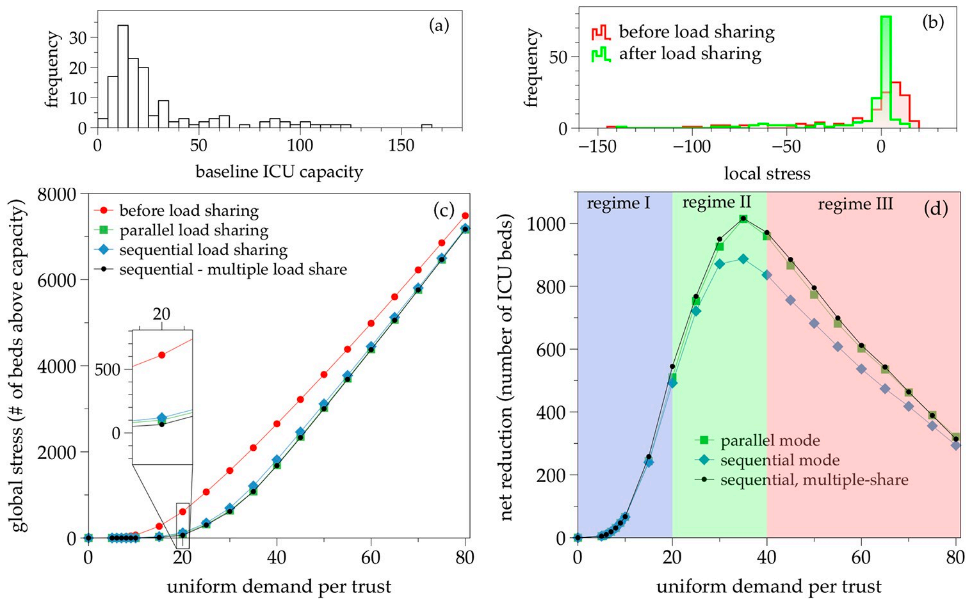

Fig 4. (a) Histogram of the `baseline-ICU-capacity` (number of beds) per trust for the NHS trust network. (b) Illustration of the histogram of `local-stress` (expected demand of number of beds above capacity) per trust, before and after applying the load sharing procedure. In the synthetic example, all trusts have a uniform `projected-ICU-demand` = 20, whereas the `baseline-ICU-capacity` is informed by data and shown in the left panel. Before the load sharing procedure, `global-stress` = 611, and after the procedure, the new `global-stress` = 101, i.e. a reduction of a total of 510 ICU patients (83%). (c and d) Response of the UK healthcare system in terms of `global-stress` (c) and net reduction in number of ICU beds (d) after the load sharing procedure is applied, as a function of the initial demand per trust (uniform demand across trusts). Different lines correspond to different modes: absence of load sharing (red); single-share, parallel mode (green); single-share, sequential mode (blue); multiple-share, sequential mode (black). Results are similar across modes. We can see three regimes: an initial regime where the load sharing procedure easily removes all signs of overwhelming, a second regime where although the procedure cannot remove all signs of overwhelming, the net reduction is maximised, and a third regime where the load sharing procedure is less and less efficient due to the fact that the whole system is overwhelmed.

https://doi.org/10.1371/journal.pone.0241027.g004

`baseline-ICU-capacity` to its real value, and $d_{\max} = \infty$. The histogram of `baseline-ICU-capacity` is reported in panel (a) of Fig 4, whereas the histogram of `local-stress`, before and after the load sharing procedure is performed, is depicted in panel (b) of the same figure (we are only showing the parallel mode here). The procedure is capable of reducing the global stress of the system from an initial value of `global-stress` = 611 ICU beds in excess in overwhelmed trusts, to a final value of `global-stress` = 101 after the optimal load sharing is performed, i.e. a transfer and subsequent treatment of 510 ICU patients.

3.1.2 Pipeline of uniform-load stress tests. In a second step, we explore how the system behaves when initial demand per trust varies. To do that, we consider a suite of stress tests and assume for each test that all trusts receive the same load –leading to a uniform demand per trust–, and we compute the `local-stress` before and after the load sharing procedure is applied. Accordingly, the `global-stress` of the whole system and the net reduction in the number of ICU beds in deficit (in collapsed trusts) is also computed.

Results are shown for both the sequential and parallel mode in panels (c) and (d) of Fig 4. Panel (c) plots the `global-stress` before and after the load sharing procedure is applied, as a function of the initial demand uniformly applied to all trusts. The net reduction (number of ICU patients or ventilators transferred) is then plotted in panel (d). As expected, the `global-stress` curves increase when the demand per trust is increased. At the beginning (for a uniform demand between 0 and 20 ICU beds per trust), the load sharing procedure works very well and completely removes any sign of overwhelming of the system (i.e. keeping the `global-stress` close to zero). When the demand per trust increases further we enter a second regime (between 20 and 40 ICU beds per trust) where the system shows signs of overwhelming but the load sharing procedure still removes a large portion of it (between 40 and 80%). If the demand per trust increases above 40 ICU beds, the whole system becomes overwhelmed, and the load sharing procedure becomes less and less capable of clearing demand, and the resulting net reduction decreases. Results are systematically better for the parallel mode than the sequential mode, but as previously mentioned, this comes at the expense of overwhelming some receptor trusts. Sequential mode still provides very good results and precludes receiving trusts from being overwhelmed.

3.2 Multiple-share in the UK NHS trust network

In this second section we relax the single-share assumption and allow each trust to share multiple loads to multiple receiving trusts, selected from the trust's topological neighborhood at random. For this analysis we drop the parallel mode and only consider the sequential processing mode, where real values of `local-stress` are updated in a sequential way as load sharing is performed.

In the uniform-load stress test, enabling a multiple-share option in the sequential mode provides an improvement in the net reduction of cases when compared to the single-share case. However, the improvement is not large (see panel (d) of Fig 4), and puts the multiple-share sequential mode on a similar footing to the single-share parallel mode, while guaranteeing that no receptor trust is overwhelmed. This result is easy to interpret: there is not much gain in being able to share loads to several receptor nodes at once (as opposed to only one), because on average this possibility will only be useful in a handful of cases. In other words, this result is a byproduct of imposing a uniform-load.

A different result is expected if the initial demand on each node is not uniform. Suppose, for instance, that we have a few trusts that are extremely overwhelmed, and could in principle share loads with several receptors (more than one available receptor in its topological neighborhood), but suppose that those receptors are small trusts with only a small number of available ICU beds. In that case, a single-share approach is clearly deficient, but a multiple-share approach could indeed provide a notable improvement. We illustrate this case in what follows.

3.2.1 Heterogeneous-load stress test. Instead of loading a uniform demand in each trust, we now test the scenario where demand is heterogeneous, and we only overwhelm 'large' trusts. To model such demand, we assume that if the trust originally has a `baseline-ICU-capacity` larger than a certain pre-defined threshold τ, then we set an initial value for `projected-ICU-demand` for this trust equivalent to 120% its `baseline-ICU-capacity` (i.e. we set put that node in a situation with positive stress, 20% above capacity). Similarly, for those trusts whose `baseline-ICU-capacity` is smaller than the threshold τ, we set an initial `projected-ICU-demand` equivalent to 80% of their corresponding `baseline-ICU-capacity` (i.e. 20% below capacity).

We then apply the load sharing procedure sequentially and compare the net reduction of the global level of stress (number of ICU patients that can be efficiently transferred) for the single-share and the multiple-share options. In Fig 5 we plot these results as a function of the threshold τ. First, note that for very small or very large values of τ both methods are similar. This is expected because in those limiting cases, either all nodes are overwhelmed (τ very small) or virtually no nodes are overwhelmed (τ very large). Thus in both cases there is no gain

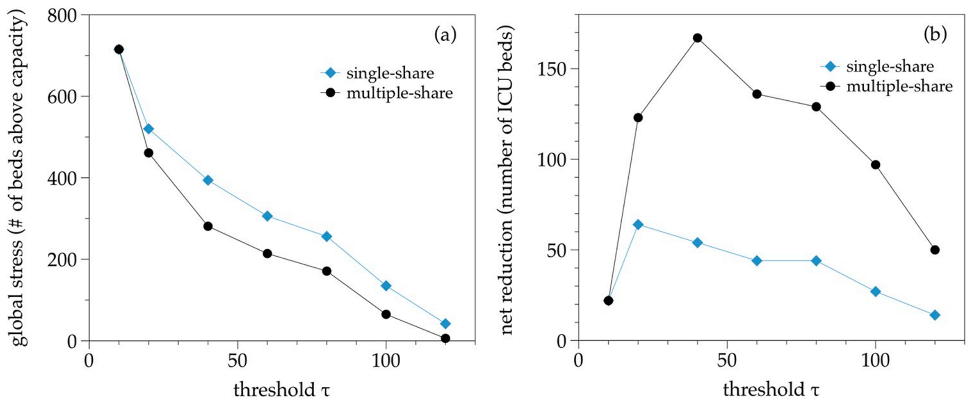

Fig 5. Global stress (panel a) and Net reduction (panel b) offered by the sequential load sharing procedure vs the threshold τ (see the text), for a single-share and a multiple-share option, in the UK system undergoing a synthetic heterogeneous-load stress test. The multiple-share option clearly outperforms the single-share one in this case.

https://doi.org/10.1371/journal.pone.0241027.g005

in performing a multiple-share over a single-share, as the net reduction of stress is small (either because we cannot transfer any load as all nodes are overwhelmed, or because no nodes are overwhelmed and there is no loads to share). Second, for the large range of intermediate values of τ where load is heterogeneous, we indeed find that the multiple-share option is much more efficient than the single-share one for a large range of values of τ, as expected.

3.3 Multiple-share in the Spanish autonomous communities contact networks

We now consider the second case: the Spanish healthcare system at the level of Spanish autonomous communities. Recall that there are 17 autonomous communities in Spain, and healthcare is decentralised so that each autonomous community runs its own system in a semi-independent way. To explore load sharing effects at the inter-community level, instead of adapting the 4-regular network to this context we have constructed two transfer networks: (i) a local contact network of 15 nodes (all autonomous communities in mainland Spain), where two nodes are linked if they share a border, and (ii) a fully connected network of all 15 autonomous communities in mainland Spain. The former allows for faster transfers, whereas the latter requires using national rail resources [23].

In both cases we use a sequential multiple-share mode. The `ICU-baseline-capacity` for each node is extracted from public data and considers both baseline and surge capacity on 30 March 2020 [28], and the `projected-ICU-demand` is initially set in terms of the ICU occupation number on 30 March 2020 [28]. The average is 63% of the national health system capacity, i.e. all autonomous communities are below capacity showing 63% load. We then increase the demand in each autonomous community, and explore how the load sharing procedure alleviates overwhelming. In the top panels of Fig 6 we illustrate a scenario, where the Spanish health system is globally overwhelmed (about 200% of the initial demand recorded on the 30th March 2020, or 130% above surge capacity). After load sharing using the contact network, some autonomous communities substantially alleviate such excess and for some others such excess is completely removed. In the ideal scenario where a fully connected network can be used, the load sharing is greatly enhanced. In the bottom panels of Fig 6 we plot the global stress and net reduction (total number of ICU beds or ventilators which are effectively

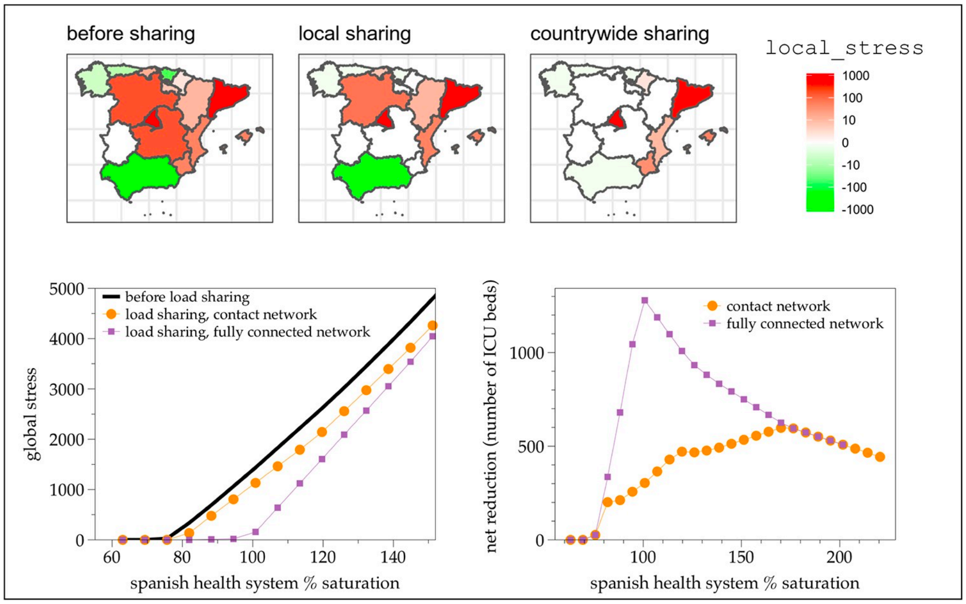

Fig 6. (Top panels) Color-coded local stress of each autonomous community in Spain, before (left panel) and after load sharing (middle and right panels). Background images have been generated using Natural Earth [23] and GADM [24]. Canary islands is absent because it is an isolated node and is off-scale. The initial demand is 200% above the real demand as of 30th March 2020, i.e. an approximate demand 130% above surge capacity. In the middle panel a local contact network is used, whereas in the right panel a fully connected network is used. The local contact network is able to transfer and alleviate about 467 ICU beds or ventilators, whereas the fully connected network can transfer 933 units. (Bottom panels) Global stress (left panel) and Net reduction (total number of ICU beds or ventilators efficiently transferred, right panel) offered by the sequential, multiple-share load sharing procedure performed on the Spanish health system, coarse-grained at the level of autonomous communities, as a function of the global saturation percentage of the system (note that capacity has already been enhanced thanks to surge capacity). Orange dots correspond to the contact network, whereas purple squares correspond to the fully connected network.

https://doi.org/10.1371/journal.pone.0241027.g006

transferred) as a function of the national health system saturation (in %), for either using the local contact network or the fully connected network. Both cases enable substantial transfers (about 600 for the local contact network and up to 1300 if transfer is done countrywide, for a single step of the algorithm). Both cases are indeed able to delay global overwhelming, and in the case of the fully connected network the algorithm can maintain the local stress of every autonomous community below capacity even when the true global saturation is around 100%.

For the local contact network, we can distinguish a first phase of steep increase, where only a few communities are overwhelmed and the algorithm is maximally efficient, until the saturation reaches about 120% of the capacity. Then in a second phase, the procedure is still able to transfer many beds or ventilators –even if some autonomous communities will still be overwhelmed–), peaking at a maximum of about 600 beds or ventilators when the system is globally at 170% capacity. As the system gets more and more overwhelmed globally, the load sharing algorithm loses efficiency and the amount of loads that can be shared starts to decrease.

4 Discussion

The COVID-19 pandemic is putting the national health systems of several countries under significant pressure. In this scenario, it is important to devise strategies that distribute capacity of hospitals, not only in terms of the number of ICU beds or ventilators, but also overall capacity (critical care, acute capacity, etc). Here, we have detailed such methodology and have implemented and validated it at two different resolutions: at the level of NHS trusts in the UK and at the level of autonomous communities in Spain. All data and code are available https://github.com/lucaslacasa/loadsharing, and will be continually updated. We presented a proof of concept and implementation and showed that this procedure works well and can de-collapse the national health systems in the UK and Spain for a range of scenarios. The random search optimisation layer permits exploration of non-intuitive load sharing configurations which go beyond the simple heuristic of sharing load with the neighbor with highest capacity (this latter being a strategy which might be locally optimal but also might be leading to a global response far away from the global optimum). We have studied several options, and compared the results of single-share (where a trust can only share load with a single receptor) or multiple-share (where the trust can share parts of the load with different receptors of its neighborhood). While the search space increases exponentially with the number of nodes, random search optimisation is scalable and can be run in real-time for system sizes comparable to realistic healthcare systems, thereby allowing for operational implementations of the method.

In the context of COVID-19, adopting a load sharing strategy is likely to be beneficial when the whole system is not completely overwhelmed, the projected ICU demand can be accurately estimated, and facilities exist to transfer either patients between ICU departments or ventilators. This is likely early on in the exponential growth phase (of each wave), or in situations where demand is declining either due to interventions or towards the end of the pandemic. When the system is already fully overwhelmed or soon-to-be, this strategy is likely to be inefficient. Furthermore, we also expect this approach to be useful as the epidemic reaches a declining phase, helping to reduce demand and allowing hospitals to return back to normal in a fast an optimised way. Note that we chose to validate the method in two countries (UK and Spain) as we could focus at two different spatial granularities. However, the method is directly applicable to other countries as well, as long as any sort of transfer system can be put in place. From a clinical point of view, an important point to consider is whether the load sharing can be activated at the ICU stage –potentially leading to transferring highly unstable patients who require ambulance with ICU equipment as well as trained personnel– or if, in anticipation to this, transfer needs to be planned at the point of hospitalisation (admission). In the latter scenario, planning needs to further take into account not only baseline ICU capacity, but overall capacity, also factoring in the estimated lag between admission to hospital and the need for ventilators, which for COVID-19 is currently estimated at about 2 to 3 days. The adequate strategy will also depend on the operational capacity of the system and the country where it is applied to. For illustration, this work explicitly considers the transfer of ICU patients, however exactly the same approach can be followed if the load to be shared is not patients but ventilators (the units to be moved are not ICU patients but ventilators, so transfer simply happens in the opposite direction, from receptor to origin). Assuming the receptor has both room and personnel to handle additional ventilators, this alternative would indeed (i) eliminate the burden on transferring highly unstable patients and the associated resources required to make such transfers, and (ii) the risk of transferring infection along with patients. Of course, risk (ii) is removed if one only transfers non-COVID ICU patients. In reality, a combination of these mechanisms (transferring ICU patients and ventilators) for sharing load is possible.

This work is subject to several limitations which we hope will be addressed in future work.

First of all, the baseline ICU demand only takes into account surge capacity in the Spanish case: more realistic analysis of the UK case shall include surge capacity, that is expected to significantly increase the real ICU capacity of each trust.

Second, in the sequential case (where receptors cannot be overwhelmed), overwhelmed nodes can at most share all the excess load, but not more (this latter case would be beneficial if e.g. two-step sharing is needed), therefore multiple-step load sharing strategies have not been explored.

Third, the optimisation process implemented here is based on a stochastic search. This method was chosen for simplicity and computational efficiency, but there is no mathematical guarantee that the suggested configuration is indeed the global optimum. More sophisticated methods such as hill climbing, genetic algorithms or simulated annealing could be used to refine this layer, if at all needed. Other extensions of interest include questions related to dynamic load balancing where the demand varies dynamically.

Fourth, and on relation to having distance between nodes as a limiting factor, note that while we have implemented such restriction ($d_{\max}$) in the code, for simplicity in this work we have set $d_{\max} = \infty$. The justification is that in two out of three realistic cases considered in this work, distance is already implicitly considered in the topology of the transfer network. For instance, in the UK case (NHS trust network), transfers are already restricted to happen only within the *closest* four trusts of a given origin. Similarly, in the first Spanish case we are considering a contact network of *adjacent* autonomous communities, i.e. transfers are only allowed to happen between communities that share a border. The last example (fully connected network of Spanish autonomous communities) is presented to assess how much more stress could be reduced if we make use of e.g. national train system to transfer between distant communities. A finer model would in this case benefit from adding a weight to every link in the network detailing the distance between any pair of autonomous communities and penalise the transfer accordingly. In such a case, we could then consider a finite $d_{\max}$, or even different values of $d_{\max}$ for different regions. All these are interesting extensions which would be relevant for a practical application of the model we present.

Finally, we have assumed that the cost of transfer is zero, i.e. the number of ambulances or the human resources are not a constraint, and that there are enough vehicles to transfer ICU patients or ventilators effectively and enough qualified personnel to handle them. All these limitations can be addressed by suitably extending the specifications of the algorithm, leading to multi-criteria optimisation problems.

Author Contributions

Conceptualization: Lucas Lacasa, Leon Danon.

Data curation: Lucas Lacasa, Robert Challen, Leon Danon.

Formal analysis: Lucas Lacasa.

Funding acquisition: Lucas Lacasa.

Investigation: Lucas Lacasa.

Methodology: Lucas Lacasa.

Project administration: Lucas Lacasa.

Resources: Leon Danon.

Software: Lucas Lacasa, Robert Challen.

Supervision: Ellen Brooks-Pollock, Leon Danon.

Visualization: Lucas Lacasa, Robert Challen.

Writing – original draft: Lucas Lacasa.

Writing – review & editing: Lucas Lacasa, Robert Challen, Ellen Brooks-Pollock, Leon Danon.

References

1. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it
2. Huang C. et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 395, 497–506 (2020). https://doi.org/10.1016/S0140-6736(20)30183-5 PMID: 31986264
3. Remuzzi Andrea, and Remuzzi Giuseppe. "COVID-19 and Italy: what next?." *The Lancet* (2020). https://doi.org/10.1016/S0140-6736(20)30627-9
4. Grasselli G, Pesenti A, Cecconi M. Critical Care Utilization for the COVID-19 Outbreak in Lombardy, Italy: Early Experience and Forecast During an Emergency Response. *JAMA*. Published online March 13, 2020. https://doi.org/10.1001/jama.2020.4031
5. Xie J., Tong Z., Guan X. et al. Critical care crisis and some recommendations during the COVID-19 epidemic in China. *Intensive Care Med* (2020). https://doi.org/10.1007/s00134-020-05979-7
6. Arabi Y.M., Murthy S. & Webb S. COVID-19: a novel coronavirus and a novel challenge for critical care. *Intensive Care Med* (2020). https://doi.org/10.1007/s00134-020-05955-1
7. Imperial College COVID-19 Response Team, COVID-19 Reports, https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/
8. Arenas Alex, Cota Wesley, Gómez-Gardeñes Jesús, Gómez Sergio, Granell Clara, Matamalas Joan T., et al, A mathematical model for the spatiotemporal epidemic spreading of COVID19, *medRxiv* 2020.03.21.20040022; https://doi.org/10.1101/2020.03.21.20040022
9. https://www.eldiario.es/madrid/hospitales-Madrid-contabilizar-coordinacion-Gobierno_0_1012799680.html
10. Jennifer Chu, Reporting tool aims to balance hospitals' Covid-19 load, MIT News Office, April 23, 2020. https://news.mit.edu/2020/reporting-tool-balance-hospitals-covid-19-0424
11. Bill Siwicki, Arizona develops surge line for load-balancing COVID-19 cases, HealthcareITNews, 15th May 2020, https://www.healthcareitnews.com/news/arizona-develops-surge-line-load-balancing-covid-19-cases
12. Wooldridge M., *An introduction to multiagent systems* (John Wiley & Sons, 2009)
13. Nisan N., Roughgarden T., Tardos E., Vazirani V.V., *Algorithmic Game Theory* (Cambridge University Press, 2007)
14. D. Yan, J. Cheng, Y. Lu and W. Ng, Effective techniques for message reduction and load balancing in distributed graph computation, *Proc. 24th International Conference on World Wide Web* (2015) pp.1307-1317
15. D. Chakrabarty, J. Chuzhoy and S. Khanna, On allocating goods to maximize fairness, *50th Annual IEEE Symposium on Foundations of Computer Science* (2009) pp 107-116.
16. Varian H., Equity, envy and efficiency. *Journal of Economic Theory* 9, 1 (1974). https://doi.org/10.1016/0022-0531(74)90075-1
17. Young H., *Equity* (Princeton University Press, 1995).
18. Moulin H., *Fair Division and Collective Welfare* (MIT Press, 2003)
19. P. Berenbrink, C. Cooper, T. Friedetzky, T. Friedrich, and T. Sauerwald. Randomized diffusion for indivisible loads. *Journal of Computer and System Sciences* 81(1), (2015): 159–185
20. P. Berenbrink, T. Friedetzky, D. Kaaser, and P. Kling. Tight & simple load balancing. In *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pp. 718-726. IEEE, 2019
21. E. Lofgren, A. Vullikanti, Hospital stockpiling problems with inventory sharing. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.
22. P-C Chen DeLaurentis, Hospital stockpiling for influenza pandemics. PhD diss., Purdue University, 2009
23. https://www.naturalearthdata.com/
24. https://gadm.org/
25. J. Lamet, Talgo S-730: así es el tren medicalizado por Renfe para trasladar pacientes entre comunidades autónomas, El Mundo 10th April 2020, https://www.elmundo.es/espana/2020/04/10/5e8f7a64fdddffe6318b462a.html
26. Danon L., Brooks-Pollock E., Bailey M., & Keeling M. J. (2020). A spatial model of CoVID-19 transmission in England and Wales: early spread and peak timing. *medRxiv*, https://doi.org/10.1101/2020.02.12.20022566
27. J. Deasy, E. Rocheteau, K. Kohler, D.J. Stubbs, P. Barbiero, P. Lio, et al, Forecasting ultra-early intensive care strain from COVID-19 in England. *medRxiv* 2020.03.19.20039057; https://doi.org/10.1101/2020.03.19.20039057

28. L. Lacasa, R. Challen, S. Hendricks, E. Brooks-Pollock, L. Danon, Potential for monitoring COVID-19 cases via primary care prescription data, in preparation.

29. NHS England. Critical Care Bed Capacity and Urgent Operations Cancelled. https://www.england.nhs.uk/statistics/statistical-work-areas/critical-care-capacity/

30. El mapa de las UCI en España: siete comunidades autónomas, al borde del colapso por el coronavirus, https://cadenaser.com/ser/2020/03/30/sociedad/1585591269_594948.html