I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 14860

A Demand-Oriented Approach to Health Care Capacity Planning

Danny Wende
Thomas Kopetsch
Wolfram F. Richter

NOVEMBER 2021

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 14860

A Demand-Oriented Approach to Health Care Capacity Planning

Danny Wende
TU Dresden and bifg

Thomas Kopetsch
Formerly: National Association of Statutory Health Insurance Physicians

Wolfram F. Richter
TU Dortmund, CESifo and IZA

NOVEMBER 2021

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

A Demand-Oriented Approach to Health Care Capacity Planning

The planning practice of health care capacities suffers from sectoral and regional constraints and it remains difficult to ensure an equal access for patients. Moreover, standard planning approaches lack the choice-theoretic grounding necessary for making reliable predictions of the demand and competition for supplied care. This paper presents a general equilibrium model designed to overcome such shortcomings. The derived metric of access to care is demand-oriented measuring the time patients waste seeking treatment. It contrasts with the usual metrics based on the floating catchment area (FCA) method, which suffer from supply bias and ad hoc specification. The approach is illustrated using Germany as an example. Much in line with official planning figures, overcapacities are shown to exist in all specialities. However, a closer look at the data provides a differentiated picture. Overcapacities are typical for urban regions and they go hand in hand with supply deficits in rural areas, albeit to a specialty-specific extent. In smaller towns, the supply is more in line with demand.

JEL Classification: H72, I18, D58, C3

Keywords: access to health care, capacity planning, endogenous demand for health services, general equilibrium model, gravity equation, floating catchment area metric

Corresponding author:
Danny Wende
Burgkstr. 31
01159 Dresden
Germany
E-mail: danny.wende@mailbox.org

1 Introduction

All well-developed health systems use capacity planning to ensure adequate access to care and a balanced allocation of resources (Ono et al., 2013). Despite their central importance for health care, common planning methods suffer from two severe shortcomings. First, they typically assume that medical demand is exogenously determined although it cannot be denied that the demand is susceptible to subjective choice and social influence. Second, planning methods are almost without exception area- and sector-specific (Ono et al., 2013). Yet in reality, health care is characterised by a transfer of services between regions and sectors, especially between urban and rural areas and between the ambulatory and hospital care sectors. Thus, (Czihal et al., 2012) have shown for Germany that there are regions, which, as net importers of health services, meet only 39% of their local demand. By contrast, net exporters deliver up to 473% of services used locally.

We contribute to the literature by developing a model that allows overcoming these two shortcomings. We do so by drawing on general equilibrium models developed over the last fifty years to explain interregional and intersectoral trade in goods and services. These models allow demand and supply relationships to be represented by equations that can be given a gravity-theoretical interpretation and that can be used for predicting trade very accurately (Anderson, 2011). They do so because they have a firm grounding in choice theory (Redding, 2010).

Our paper synthesises the full equilibrium trade model of Anderson and van Wincoop (2003) and the partial equilibrium model set up by Bikker and de Vos (1992) for analysing inpatient care as the result of a supply-demand relationship. Using Germany as an example, we demonstrate that the presented approach is well suited to measure access to health care and to improve practical capacity planning. As one might expect, rural and deprived regions are shown to suffer from undersupply with health care resources while oversupply is a characteristic of urban areas. A closer look at the individual sectors reveals a more differentiated picture.

Our paper is structured as follows: After Section 2 has briefly summarized the relevant literature, Section 3 introduces a simple yet full equilibrium model that allows us to derive a gravity-theoretical representation of the demand for medical treatment. In Sections 4 and 5 it is shown how the model can be used for planning the spatial allocation of capacities that ensures an equal access to care. Section 6 provides an application to Germany. Section 7 concludes.

2 Related literature

Gravity models are derived from the law of gravity in classical physics, which states that two bodies attract each other and that the strength of the attraction depends on the masses of the bodies and their physical distance (Anderson, 2011). This basic idea is taken up by economic gravity models. The first

theoretical modelling in the context of international trade was performed by Anderson (1979),[1] who derived the gravity equation by assuming regionally distributed consumers and producers of commodities. The consumers maximize utility at constant elasticity of substitution (CES), the producers maximize profit, and interregional trade is costly. It follows from the model that, after controlling for size differences, trade between two regions depends on the ratio of bilateral trade costs to the average of all trade costs. Variants of the model in more recent publications supplement the assumption of CES preferences with an assumption of monopolistic competition in order to endogenize the specialization of producers. An early example of this is Bergstrand (1985).

For the health economics literature, the description of catchment areas by Huff (1964) is the first reference for the concept of gravity. Lowe and Sen (1996) use a gravity equation to predict patient flows and catchment areas. The commonest approach to measuring access to health care captures the idea of gravity in supply-demand relationships by building on the floating catchment area (FCA) metric[2] (Delamater, 2013). However, this approach suffers from a strong supply bias, with the demand and competition for medical treatment represented only by functional *ad-hoc* specifications. As a consequence, estimates of model parameters are often determined anecdotally. For example, catchment areas are simply set at 30 to 60 minutes or at corresponding distances (Delamater, 2013), or are replaced by proxy variables (Matthews et al., 2019). The model presented in this paper endogenizes the demand for health care. We therefore call it *demand oriented*. The FCA metric is shown to refer to the special case characterized by exogenous demand.

Trade models cannot be readily applied to the exchange of health services, as prices play different roles. In goods markets, it is clearly the task of prices to balance demand and supply. In health care, prices exist at best in the form of co-payments. Overwhelmingly, queues take over the allocative control function (Lindsay and Feigenbaum, 1984). Waiting times are even considered a "key policy concern" in many OECD countries, though not in all (Sá et al., 2019; OECD, 2020, Fig. 1.2). As a result, there is also a growing literature analysing the optimal management of queues when supply is taken as given. In contrast, the model presented here is intended to serve demand-oriented planning of regionally and sectorally differentiated capacities. It therefore assumes that when service capacities are scarce, waiting times form endogenously, which, in addition to spatial distances, dampen demand or redirect it to competing capacities and thus bring it into line with supply. It should be mentioned that, unlike waiting times, the quality of supply is not an object of the modelling, even though quality is undoubtedly demand-determining in practice. However, we refrain from modelling demand for quality, as it is not a primary consideration in capacity planning.

[1] It should be noted, however, that Leamer and Stern (1970) had already provided the trade gravity equation by intuitive reasoning.

[2] Variants of the FCA method are the two-step (2S-), the extended two-step (E2S-), and the kernel density two-step (KD2S-) FCA methods. They mainly differ from the base version in spatial weighting (Delamater , 2013). The integrated (i-) and the three-step (3S-) FCA methods are discussed in Section 3.4.

3 Modelling supply and demand for health care

Health care capacities are assumed to differ by the place of service delivery, $j \in \{1,\dots,\#j\}$, and the type of medical speciality, $s \in \{1,\dots,\#s\}$. They are demanded from individuals and used by those in their patient role. An individual's place of residence is not necessarily the place where health services are delivered. On the contrary, it is assumed that patients travel from their homes, $i \in \{1,\dots,\#i\}$, to the locations, j, where they receive treatment. The spatial distance between the place of residence i and the place of service delivery j is denoted by d_{ij}.

3.1 The demand for health care

An individual with residence at i derives utility U_i from health-neutral consumption c_i and the bundle of treatments, $n_{i\cdot} = \left(n_{ijs}\right)_{js} \equiv \left(n_{i11},..,n_{ijs},..\right)$, which she or he expects to receive from providers of health services of specialities s at locations j. Those health services, n_{ijs}, have a time dimension. The utility derived from them is modelled as a nested function, $V_i \equiv V(W_{\cdot}(n_{i\cdot}))$ with $W_{\cdot} \equiv (W_1,..,W_s,..)$. The function $W_s = W_s\left(n_{i1s},..,n_{ijs},..\right)$ captures utility derived from services of the speciality s at the various locations.

$$U_i(c_i, n_{i\cdot}) \equiv c_i^{1-\mu_i} \cdot V_i^{\mu_i} \text{ with } V \equiv \left[\sum_s \vartheta_s \cdot W_s^{\frac{\sigma-1}{\sigma}}\right]^{\frac{\sigma}{\sigma-1}} \text{ and } W_s \equiv \left[\sum_j n_{ijs}^{\frac{\omega_s-1}{\omega_s}}\right]^{\frac{\omega_s}{\omega_s-1}}. \quad (1)$$

Substitutability is assumed at each of the three utility levels, albeit at constant elasticity and to varying degrees. At the top level, the elasticity of substitution is assumed to be one between health services and health-neutral consumption (Cobb-Douglas) and the partial elasticity of health-neutral consumption is $1-\mu_i$. The parameter μ_i can be interpreted as need for health care. At the lowest level, the elasticity of substitution carries an index s. The idea is that the substitutability between treatments from service providers of different locations varies with the kind of speciality, notably between elective and emergency treatments. In the case of emergency, spatial substitutability will be large, while ω_s will be smaller for elective treatment. Substitutability between different specialities will be particularly small with a tendency towards complementarity in demand, $\sigma < 1$. The relative importance of specialities is captured by the preference weights ϑ_s.

Individuals must keep to two budgets. One is for financing health-neutral consumption by income from labour. The price of consumption is normalized to one and the real wage rate is w. The other budget is for spending time. It is restricted by y which can be read as "year". Time is used for labour supply, l_i, health services, n_{ijs}, and time wasted for receiving medical treatment. The latter includes both the time the patient needs to visit a doctor's location, $f_{ijs} \equiv f_s(d_{ij})$, and the time, τ_{js}, he or she has to wait for

receiving treatment. We assume that distances, unlike waiting times, are unalterable. The total time it takes to be treated is $t_{ijs} \equiv n_{ijs} \cdot \tau_{js} \cdot f_{ijs}$. For model-related reasons, the time components are assumed to be multiplicatively connected.

$$c_i = w \cdot l_i \qquad \text{(consumption budget)} \qquad (2)$$

$$\sum_{js} n_{ijs} \cdot \tau_{js} \cdot f_{ijs} = y - l_i \qquad \text{(time budget).} \qquad (3)$$

By maximizing the utility function (1) subject to the constraints (2) and (3) we obtain the supply of labour, l_i, as a fraction of y, the demand for health-neutral consumption, $c_i = c_i(w)$, as a function of the real wage rate, w, and the demand for health services as a function of time wasted for receiving treatment, n_{ijs}. As distances are assumed to be unalterable the demand for health services, $n_{ijs}(\tau.)$, is stated as a function of the waiting time profile, $\tau. \equiv (\tau_{js})_{js}$, while the dependence on f_{ijs} is supressed:

$$l_i = (1-\mu_i)y\ ,\ c_i(w) = w(1-\mu_i)y,\ n_{ijs}(\tau.) = \mu_i y \cdot \vartheta_s^{\sigma} \cdot \frac{\tau_{js}^{-\omega_s} \cdot f_{ijs}^{-\omega_s}}{T_i(\tau.)} \qquad (4)$$

The denominator,

$$T_i(\tau.) \equiv \sum_{js} \vartheta_s^{\sigma} \cdot \tau_{js}^{1-\omega_s} \cdot f_{ijs}^{1-\omega_s}, \qquad (5)$$

can be interpreted as a (preference-weighted) *index of wasted time* (per unit of demand for health care). The waste of time reduces the demand. It is worth noting that the index is independent of $\tau.$ and constant in i if $\omega_s = 1$ for all s (Cobb-Douglas). An implication of equations (3) and (4) is that $\mu_i y = y - l_i$ can be interpreted as the time which individuals expect to spend on the consumption of health services.

3.2 The supply of medical treatment and general equilibrium conditions

In many countries, the remuneration of health care services is strictly regulated. In such circumstances, doctors' scope of action is limited to the amount of labour they choose to supply. In its simplest form, the amount L_{js} chosen by a doctor practicing in speciality s at location j is merely a function of the doctor's real fee, h_{js}. Total labour supply, S_{js}, amounts to individual labour supply, $L_{js}(h_{js})$, multiplied by the number of doctors, A_{js}.

Patients demand health services that are time consuming. Let I_i be the number of individuals with residence i and $x_{ijs} = x_{ijs}(\tau.) \equiv I_i \cdot n_{ijs}(\tau.)$ the time demanded by those individuals for treatment in speciality s at location j. In equilibrium, the demand is met by supply,

$$\sum_i x_{ijs} = \sum_i I_i \cdot n_{ijs}\,(\tau.) = A_{js} \cdot L_{js} \equiv S_{js} \qquad \text{for all } js. \qquad (6)$$

The equation shows that waiting times, $\tau.$, have the function of balancing the demand for medical treatments and their supply. When high demand meets low supply, waiting times become long.

By assumption, consumer goods are produced at constant returns to scale using labour as the only input. In equilibrium, the supply price is determined by the cost of labour. Health services are financed by a payroll tax at rate b. Thus $1 - b = w$ must apply. The balancing of revenues and expenditures in the health budget requires

$$b \cdot \sum_i I_i \cdot l_i = \sum_{js} A_{js} \cdot h_{js} \cdot L_{js} \ . \tag{7}$$

The right-hand side of eq. (7) represents doctors' demand for consumption goods. By Walras' Law this demand is equal to $\sum_i I_i \cdot (l_i - c_i)$, which can be interpreted as the individuals' excess supply of consumption goods. The payroll tax rate, b, balances the budget.

3.3 A gravity-theoretic representation of the demand for medical treatment

The spatial distribution of the demand for medical treatment is given by x_{ijs}. In order to investigate this distribution more closely, we introduce the concept of *access to (health) care*,

$$\beta_{is} = \beta_{is}(\tau.) \equiv \sum_j \tau_{js}^{-\omega_s} \cdot f_{ijs}^{-\omega_s} \ . \tag{8}$$

The interpretation is suggested by the finding that eq. (4) implies a proportionality of β_{is} and the term $T_i \cdot \sum_j n_{ijs} / \mu_i$, which measures the time a representative patient with residence i wastes in receiving required medical care in specialty s. The factor $\sum_j n_{ijs}$ is individuals' demand and $\sum_j n_{ijs} / \mu_i$ patients' need of medical care. The latter has to be multiplied with T_i as this index measures wasted time per unit of demand. The concept of access to care is central to the FCA literature and was introduced by Joseph and Bantock (1982) using *ad-hoc* reasoning. The demand-oriented approach, in contrast, allows interpreting β_{is} as a *metric of wasted time*.

Remark 1: The demand-oriented approach suggests measuring access to care by the time patients waste seeking treatment.

Next we set

$$D_{is} = D_{is}(\tau.) \equiv \sum_j x_{ijs} = \vartheta_s^{\sigma} \cdot \tilde{I}_i \cdot \beta_{is} \quad \text{with} \quad \tilde{I}_i = \tilde{I}_i(\tau.) \equiv \frac{\mu_i y}{T_i(\tau.)} I_i \tag{9}$$

and interpret D_{is} as the spatially *aggregated demand for treatment* in specialization s at location i. It obviously increases proportionally with the product of $\tilde{I}_i$ and the index of access to care, β_{is}. The structure of the equation is as in Bikker and de Vos (1992) except for the factor $\tilde{I}_i / I_i = \mu_i y / T_i$, which can be interpreted as a population adjustment with no direct analogue in their approach. The adjustment is made with regard to the need for health care, μ_i, and the index of wasted time, T_i. Let us call $\tilde{I}_i$ the

adjusted number of individuals with residence i. In the literature on FCA, $D_{is}/\beta_{is} = \vartheta_s^{\sigma} \cdot \tilde{I}_i$ is interpreted as (access-independent) *demand potential* (Delamater, 2013). Combining equations (4), (8), and (9), we obtain

$$\phi_{ijs} = \phi_{ijs}(\tau.) \equiv \frac{x_{ijs}}{D_{is}} = \frac{\tau_{js}^{-\omega_s} \cdot f_{ijs}^{-\omega_s}}{\beta_{is}} . \tag{10}$$

In the literature, ϕ_{ijs} is interpreted as a *cross-border demand ratio* (Czihal et al., 2012). It measures the proportion of i's aggregate demand for treatment in s to be supplied from j.

Finally, we set

$$\alpha_{js} = \alpha_{js}(\tau.) \equiv \vartheta_s^{\sigma} \cdot \sum_i \tilde{I}_i(\tau.) \cdot f_{ijs}^{-\omega_s} \tag{11}$$

so that the supply of treatment time in speciality s at location j can be written as

$$S_{js} = S_{js}(\tau.) = \sum_i x_{ijs} = \vartheta_s^{\sigma} \cdot \tau_{js}^{-\omega_s} \cdot \sum_i \tilde{I}_i \cdot f_{ijs}^{-\omega_s} = \tau_{js}^{-\omega_s} \cdot \alpha_{js} . \tag{12}$$

From this equation it can be seen that α_{js} measures the number of treatments carried out in speciality s at location j. In the literature on FCA, α_{js} is interpreted as the *population's care potential* and S_{js}/α_{js} as the *supply ratio* (Delamater, 2013).

Remark 2: In equilibrium, the aggregate supply of treatment time, $S_s \equiv \sum_j S_{js}$, is allocated to the places of residence in proportion to the product of the adjusted number of individuals, $\tilde{I}_i$, and the metric of access to care, β_{is},

$$D_{is} = \frac{\tilde{I}_i \cdot \beta_{is}}{\sum_{i\prime} \tilde{I}_{i\prime} \cdot \beta_{i\prime s}} \cdot S_s . \tag{13}$$

The proof is straightforward. From the equations (12) and (8) we obtain

$$S_s = \sum_j S_{js} = \vartheta_s^{\sigma} \cdot \sum_i \tilde{I}_i \cdot \sum_j \tau_{js}^{-\omega_s} \cdot f_{ijs}^{-\omega_s} = \vartheta_s^{\sigma} \cdot \sum_i \tilde{I}_i \cdot \beta_{is}. \tag{14}$$

Eq. (13) follows when dividing eq. (9) by eq. (14).□

Proposition 1: The demand of individuals with residence i for treatment in speciality s at location j, x_{ijs}, is proportional to the demand potential, the supply ratio, and an indicator of distance

$$x_{ijs} = \left(\frac{D_{is}}{\beta_{is}}\right) \cdot \left(\frac{S_{js}}{\alpha_{js}}\right) \cdot f_{ijs}^{-\omega_s}. \tag{15}$$

Eq. (15) follows from equations (10) and (12). By interpreting the terms in brackets as "masses", the equation gets a gravity-theoretical meaning.

When relying on the definition (10) of the cross-border demand ratios eq. (15) can be written as

$$\beta_{is} = \phi_{ijs}^{-1} \cdot f_{ijs}^{-\omega_s} \cdot \frac{S_{js}}{\alpha_{js}}. \tag{16}$$

By setting $k_s \equiv [\prod_j (S_{js}/\alpha_{js})]^{\frac{1}{\#j}}$ and taking geometric means of both sides with respect to j we obtain the noteworthy equation

$$\beta_{is} = k_s \cdot [(\prod_j \phi_{ijs})^{\frac{1}{\#j}} \cdot (\prod_j f_{ijs})^{\frac{\omega_s}{\#j}}]^{-1}. \tag{17}$$

Proposition 2: The demand-oriented access metric is inversely proportional to the product of the geometric means of cross-border demand ratios and distances.

Unsurprisingly, access to care is negatively affected by high cross-border demand ratios and long distances. Furthermore, each geometric mean can be substituted by the other at constant elasticity.

3.4 A comparison of the gravity model with competing approaches

As has been mentioned, the idea of predicting health care utilisation by deriving the gravity equation (15) from a supply-demand relationship goes back to Bikker and de Vos (1992). The difference to the present approach comes from the derivation of supply and demand for health care. Bikker and de Vos simply postulate the existence of such functions. They focus on a particular medical speciality and analyse supply and demand in partial equilibrium. The present approach is one of full equilibrium analysis. Individuals must divide their time budget between the different health services, and demands have to be balanced by supplies. The difference is reflected in the index of wasted time, T_i, which has no analogue in the approach taken by Bikker and de Vos. The index varies with waiting times and individuals' willingness to substitute between specialities and locations. Thus, it indicates the strength of competition of demand for supply. Such an indicator should play a key role in health capacity planning.

In the literature, the dominant approaches to measuring access to care build on the FCA metric. In its base version, this metric, $\beta_{i,FCA}$, is defined in a two–step procedure. In a first step, the population is aggregated to a demand potential using a geographical weighting scheme, $g(d_{ij})$. In a second, the ratio of supply (often approximated by the number of service providers) to the demand potential is calculated

using the geographical weighting scheme to create the index (Delamater, 2013). If we suppress s and equate $g(d_{ij})$ with $f_{ij}^{-\omega}$ the metric can be written as

$$\beta_{i,FCA} = \sum_j \frac{S_j \cdot g(d_{ij})}{\sum_{i'} I_{i'} \cdot g(d_{i'j})} = \sum_j \frac{S_j \cdot f_{ij}^{-\omega}}{\sum_{i'} I_{i'} \cdot f_{ij}^{-\omega}} . \tag{18}$$

A structurally similar formula can be derived from the gravity model by relying on equations (8), (11), and (12):

$$\beta_{i,D} = \sum_j \tau_j^{-\omega} \cdot f_{ij}^{-\omega} = \sum_j \frac{S_j}{\alpha_j} \cdot f_{ij}^{-\omega} = \vartheta^{-\sigma} \cdot \sum_j \frac{S_j \cdot f_{ij}^{-\omega}}{\sum_{i'} \tilde{I}_{i'} \cdot f_{ij}^{-\omega}} \tag{19}$$

For simplicity, let us call $\beta_{i,D}$ the *demand-oriented metric of access to care*. It is a function of the adjusted number of individuals, $\tilde{I}_i = \frac{\mu_i y}{T_i} I_i$, whereas $\beta_{i,FCA}$ is one of the unadjusted number, I_i. This difference does not matter if the adjustment factor, $\mu_i y / T_i$, is constant in i. This is clearly the case if both μ_i and T_i are constant in i. Constancy of T_i is ensured if individuals' preferences are of the Cobb-Douglas type. The metrics, $\beta_{i,FCA}$ and $\beta_{i,D}$, are then equal up to a constant factor, as is easily seen when comparing the equations (18) and (19). This is at odds with empirical visualizations of $\beta_{i,FCA}$ and $\beta_{i,D}$, which differ considerably (see Section 6.2).

Proposition 3: The FCA metric is equal to the demand-oriented metric if the need to adjust the population for health care need, distances, and waiting times is ignored. In that case, the demand potential is proportional to the population size, $D_i / \beta_i = k \cdot I_i$.

However, it should not go unmentioned that the literature no longer regards the FCA metric as state of the art. Wan et al. (2012), for example, find it likely that it overestimates the demand for health care. Criticism of this kind has led to efforts to develop more refined metrics. Examples that deserve special mention are the three step (3S-) and the integrated (i-) FCA metrics proposed by Wan et al. (2012) and Luo (2014), respectively:

$$\beta_{i,3SFCA} = \sum_j \frac{S_j \cdot g(d_{ij}) \cdot G_{3SFCA}(d_{ij})}{\sum_{i'} I_{i'} \cdot g(d_{i'j}) \cdot G_{3SFCA}(d_{i'j})} \quad \text{with} \quad G_{3SFCA}(d_{ij}) = \frac{g(d_{ij})}{\sum_{j':\, d_{ij'} < \bar{d}} g(d_{ij'})} \tag{20},$$

$$\beta_{i,iFCA} = \sum_j \frac{S_j \cdot g(d_{ij}) \cdot G_{iFCA}(S_j, d_{i'j})}{\sum_{i'} I_{i'} \cdot g(d_{i'j}) \cdot G_{iFCA}(S_j, d_{i'j})} \quad \text{with} \quad G_{iFCA}(S_j, d_{ij}) = \frac{S_j \cdot g(d_{ij})}{\sum_{j':\, d_{ij'} < \bar{d}} S_{j'} \cdot g(d_{ij'})} \tag{21}$$

where $\bar{d}$ denotes some threshold distance defining a catchment. The obvious difference to the FCA metric comes from the factor G_{*FCA}. It has been added to the 3SFCA metric to incorporate the potential for competition among service providers when more than just one provider is located within i's catchment area (Delamater, 2013, p. 31). By contrast, Luo (2014) proposes the iFCA metric to

"moderate the over- or under-estimating of population demand that occurred with previous methods". To this end, the demand for health care is "adjusted by a Huff Model-based selection probability that reflects the impacts of both distance impedance and service site capacity". After running simulations, Paez et al. (2019) conclude that access to health care is best measured by the iFCA metric. However, all FCA-based metrics are open to the same kind of criticism. First, they are supply-biased which is not appropriate for demand-oriented resource planning. Secondly, competition among service providers is modelled by functional *ad-hoc* specifications. All this is different in the demand-oriented model. Demand is derived from utility maximization and the effect of competition is captured by the length of waiting times, τ_{js}, which are endogenously determined by the balancing of demand and supply.

4 Planning an equal access to care

Health equity is of paramount importance in any conception of social justice (Sen, 2002). This is recognized in most countries and the reason why health authorities pay the greatest attention to health care capacity planning and resource allocation (Ono et al., 2013). In the following, we interpret the equity objective as a mandate to establish equal access to care,

$$\bar{\beta}_s \equiv \bar{\beta}_{is} = \sum_j \bar{\tau}_{js}^{-\omega_s} \cdot f_{ijs}^{-\omega_s} \qquad \text{for all } is. \tag{22}$$

In other words, the time patients waste seeking treatment should not depend on their residence (Remark 1). Just to avoid misunderstandings: This does not mean an equalisation of the indices of wasted times, T_i. The metric of access to care, β_{is}, takes into account not only waiting times and distances, but also the need for medical treatment which T_i does not. In this section, we show how to determine the spatial allocation of capacities, which is compatible with an equal access to care.

4.1 Planning the spatial allocation of capacities

Distances are difficult to shorten in practice. For this reason, distances are assumed fixed and unalterable. The only way to vary $\bar{\beta}_{is}$ then is to control waiting times, $\bar{\tau}_{js}$. Eq. (22) defines a system of equations which under conditions of regularity can be used to determine the matrix $\bar{\tau}_{\cdot}^{-\omega_s} \equiv (\bar{\tau}_{js}^{-\omega_s})_{js} = (\bar{\tau}_{js}^{-\omega_s}(\bar{\beta}_s))_{js}$ as a function of $\bar{\beta}_s$.[3] By the system's linear structure, $\bar{\tau}_{js}^{-\omega_s}(\bar{\beta}_s) = \bar{\beta}_s \cdot \bar{\tau}_{js}^{-\omega_s}(1)$ must hold. Associated with $\bar{\beta}_{\cdot} = (\bar{\beta}_s)_s$, there is the index of wasted time,

$$\bar{T}_i(\bar{\beta}_{\cdot}) \equiv T_i\left(\bar{\tau}_{\cdot}(\bar{\beta}_{\cdot})\right) = \sum_{js} \vartheta_s^{\sigma} \cdot \bar{\tau}_{js}^{1-\omega_s}(\bar{\beta}_s) \cdot f_{ijs}^{1-\omega_s}$$

[3] Solvability requires invertible matrices $(f_{ijs}^{-\omega_s})_{ij}$ for all s and *a fortiori* equality of $\#j$ and $\#i$ (Gentle, 2007, p. 211).

$$= \sum_{js} \vartheta_s^\sigma \cdot \bar{\beta}_s^{1-1/\omega_s} \cdot \bar{\tau}_{js}^{1-\omega_s}(1) \cdot f_{ijs}^{1-\omega_s}. \tag{23}$$

Let $\bar{\tilde{I}}_i \equiv \tilde{I}_i(\bar{\beta}_\cdot) = \frac{\mu_i y}{\bar{T}_i(\bar{\beta}_\cdot)} I_i$ denote the adjusted number of individuals associated with $\bar{\beta}_\cdot$. Furthermore, let $\bar{D}_{is} = D_{is}(\bar{\tau}_\cdot(\bar{\beta}_s))$ and $\bar{S}_s = \sum_j \bar{S}_{js} = \sum_j S_{js}(\bar{\tau}_\cdot(\bar{\beta}_s))$ denote the associated demand and aggregate supply, respectively. For an equal access to health care, eq. (13) takes the particularly simple form of

$$\bar{D}_{is} = \frac{\bar{\tilde{I}}_i}{\sum_{i'} \bar{\tilde{I}}_{i'}} \cdot \bar{S}_s. \tag{24}$$

Adjusting waiting times requires rebalancing supply and demand. Of the two sides, only supply can be the direct object of planning. By contrast, demand at location i, $\bar{D}_{is}$, will respond endogenously to any changes in supplies and waiting times. The variables that allow planners to control the supply of health care are the number and location of service providers, $\bar{A}_{js}$, and the hours worked by doctors, $L_{js}(\bar{h}_{js})$, as functions of real fees, $\bar{h}_{js}$. In the short term, it makes sense to assume fixed and location-independent hours of working. This can mean, for example, that doctors work five days at 8 hours a week, while hospitals are open around the clock (24/7). Thus assuming $\bar{L}_s \equiv L_{js}(\bar{h}_{js})$ to be fixed for all js, the locally required number of service providers is obtained by dividing planned capacity $\bar{S}_{js}$ by $\bar{L}_s$:

$$\bar{A}_{js} = \frac{\bar{S}_{js}}{\bar{L}_s} \tag{25}$$

Hence, the primary task of planning is to compute the allocation of local capacities, $\bar{S}_{js}$, sustaining an equal access to care. We solve this computation by first deriving values of $\bar{S}_{js}$ as functions of $\bar{\beta}_1, \bar{\beta}_s, \ldots$ We then show how to determine the values of $\bar{\beta}_1, \bar{\beta}_s, \ldots$ We start with the first step.

Equations (14) and (22) imply

$$\frac{\bar{S}_s}{\bar{S}_1} = \frac{\vartheta_s^\sigma}{\vartheta_1^\sigma} \frac{\bar{\beta}_s}{\bar{\beta}_1}, \tag{26}$$

i.e. planned relative supplies need to equate the product of planned relative accesses to care, $\bar{\beta}_s/\bar{\beta}_1$, and the ratio of preference weights, $\vartheta_s^\sigma/\vartheta_1^\sigma \equiv \rho_s$.

By combining equations (12), (23), and (26) we are able to determine planned local supplies, $\bar{S}_{js}$, as functions of planned aggregate supplies, $\bar{S}_1, \bar{S}_s, \ldots$, and target metrics of access to care, $\bar{\beta}_1, \bar{\beta}_s, \ldots$:

$$\bar{S}_{js} = \vartheta_s^\sigma \cdot \bar{\tau}_{js}^{-\omega_s} \cdot \sum_i \bar{\tilde{I}}_i \cdot f_{ijs}^{-\omega_s} = \vartheta_s^\sigma \cdot \bar{\beta}_s \cdot \bar{\tau}_{js}^{-\omega_s}(1) \cdot \sum_i \frac{\mu_i y}{\sum_{j's'} \vartheta_{s'}^\sigma \cdot \bar{\beta}_{s'}^{1-1/\omega_{s'}} \cdot \bar{\tau}_{j's'}^{1-\omega_{s'}}(1) \cdot f_{ij's'}^{1-\omega_{s'}}} I_i \cdot f_{ijs}^{-\omega_s}$$

$$= \bar{S}_s \cdot \sum_i \frac{\mu_i y \cdot I_i \cdot \bar{\tau}_{js}^{-\omega_s}(1) \cdot f_{ijs}^{-\omega_s}}{\sum_{j's'} \bar{S}_{s'} \cdot \bar{\beta}_{s'}^{-1/\omega_{s'}} \cdot \bar{\tau}_{j's'}^{1-\omega_{s'}}(1) \cdot f_{ij's'}^{1-\omega_{s'}}} \tag{27}$$

4.2 Determining target metrics of access to care

In order to determine the unknown target values $\bar{\beta}_1, \bar{\beta}_s$, ..., we add up the equations (27) over j, rely on eq. (22), and reduce:

$$1 = \sum_i \frac{\mu_i y \cdot I_i \cdot \sum_j \bar{\tau}_{js}^{-\omega_s}(1) \cdot f_{ijs}^{-\omega_s}}{\sum_{j's'} \bar{S}_{s'} \cdot \bar{\beta}_{s'}^{-1/\omega_{s'}} \cdot \bar{\tau}_{j's'}^{1-\omega_{s'}}(1) \cdot f_{ij's'}^{1-\omega_{s'}}} = \sum_i \frac{\mu_i y \cdot I_i}{\sum_{j's'} \bar{S}_{s'} \cdot \bar{\beta}_{s'}^{-1/\omega_{s'}} \cdot \bar{\tau}_{j's'}^{1-\omega_{s'}}(1) \cdot f_{ij's'}^{1-\omega_{s'}}} \quad (28)$$

By again relying on eq. (26) we end up with

$$1 = \sum_i \mu_i y \cdot I_i \cdot \left(\sum_{s'} \bar{\beta}_1^{-1/\omega_{s'}} \cdot \rho_{s'}^{1/\omega_{s'}} \cdot \bar{S}_{s'}^{1-1/\omega_{s'}} \cdot \sum_{j'} \bar{\tau}_{j's'}^{1-\omega_{s'}}(1) \cdot f_{ij's'}^{1-\omega_{s'}}\right)^{-1} \quad (29)$$

Hence, we obtain $\bar{\beta}_1$ as an implicit function of $\bar{S}_s$, μ_i, I_i, ω_s, f_{ijs}, and $\vartheta_s^\sigma / \vartheta_1^\sigma = \rho_s$ (all ijs). The remaining target values of access to care, $\bar{\beta}_s$ ($s > 1$), are finally obtained from eq. (26).

To summarize, the following data allow us computing the planned number of local service providers, $\bar{A}_{js}$: (i) the need for health care, μ_i, (ii) the number of individuals I_i with residence i, (iii) the time patients need to visit doctors' locations, f_{ijs}, (iv) the elasticity of spatial substitution, ω_s, (v) the ratios of preference weights, ρ_s, (vi) standard working times, $\bar{L}_s$, and (vii) planned aggregate supplies, $\bar{S}_s$.

5 Parameterization

In this section, we propose an estimation strategy for the demand-oriented planning model and a test strategy for making an informed choice between the demand-oriented and the FCA-based metrics. By applying our approach to Germany, its practicality is demonstrated in Section 6.

5.1 Estimating distance decay functions

The use of eq. (22) for planning an equal access requires estimating the distance decay function $f_{ijs}^{-\omega_s} \equiv exp\, f_s(d_{ij})$. Functional specifications for f_s found in the literature are among others: logarithmic (Huff, 1963), linear (Fülöp et al., (2011), and quadratic (Luo and Qi, 2009). The latter is related to the Gaussian variogram model (Chilès and Delfiner, 1999). Gravity models are found to fit best when using the squared root of the travel time (Lowe and Sen, 1996), we therefore follow Luo and Qi and set

$$f_s(d_{ij}) \equiv \theta_s \cdot d_{ij}^2 \text{ with } \theta_s \equiv -\frac{3\omega_s}{\varphi_s^2}. \quad (30)$$

The parameter θ_s is estimated by relying on eq. (15):

$$x_{ijs} = exp\left(\theta_{is} + \theta_{js} + \theta_s \cdot d_{ij}^2\right) \quad (31)$$

The identification of θ_s takes advantage of the fact that the remaining factors on the right-hand side of eq. (15) vary with either i or j, but never with both indices simultaneously. This allows assuming that fixed effects capture all unobserved locational characteristics. Assuming the dependent variable to be Poisson distributed, a conditional Poisson regression can be used to derive an estimate for θ_s. However, note that such an estimate does not allow quantifying the elasticity of spatial substitution, ω_s.

In the variogram model, the estimated value of $\sqrt{-\frac{3}{\hat{\theta}_s}} = \frac{\varphi_s}{\sqrt{\omega_s}_s}$ is called *range* (Chilès and Delfiner, 1999). It measures the distance, below which 95 percent of all observable physician-patient contacts occur. The parameter determines the size of doctors' *catchment area*. By contrast, φ_s measures *patients' willingness to cover distance* to the place of service provision. For $\omega_s = 1$, the range equals the catchment area. For $\omega_s > 1$ observed distances are smaller than patients' willingness to travel. This typically applies to emergency treatment mainly provided by hospitals and in less severe cases by general practitioners or paediatricians. The alternative case with $\omega_s < 1$ is more likely for elective services when patients value proximity of treatment less than other characteristics such as service quality. In that case, observed distances are larger than patients' willingness to travel, yet spatial conditions are at odds.

The willingness of patients to travel is a property that has to be determined by survey technique. An example is provided by Sundmacher et al. (2018).

5.2 Estimating access to care

One notable advantage of the demand-oriented planning model stems from the fact that it allows for speciality dependent elasticities of spatial substitution, ω_s. However, this flexibility in the modelling of demand comes at a price. The measurement of access to care is considerably complicated if ω_s is allowed to deviate from one (Cobb-Douglas). Therefore, it is appropriate to show that the data speak against the simplifying assumption of $\omega_s = 1$ for all s. For this purpose, we compute values of β_{is} in two competing ways. One uses eq. (18) and reflects the assumption of $\omega_s = 1$ (Proposition 3). The other uses the formula

$$\beta_{is} = \sqrt[\#j]{\prod_j \frac{\sqrt[\#i]{\prod_{i'} x_{i'js} \cdot f_{i'js}^{\omega_s}}}{x_{ijs} \cdot f_{ijs}^{\omega_s} \sqrt[\#i]{\prod_{i'} D_{i's}}}} \cdot D_{is} \tag{32}$$

and exploits the available data without relying on any *a priori* constraint on ω_s. Eq. (32) is derived from eq. (15) by relying on the observation that β_{is} and α_{js} can be replaced in eq. (15) with $c_s \cdot \beta_{is}$ and α_{js}/c_s, respectively, without changing anything else. Hence, one can set the constant c_s at a value satisfying the constraint of $\sqrt[\#i]{\prod_i \beta_{is}} = 1$. By taking a geometric mean of eq. (15) with respect to i we obtain the

equation $\frac{S_{js}}{\alpha_{js}} = \sqrt[\#i]{\prod_i \frac{x_{ijs} \cdot f_{ijs}^{\omega_s}}{D_{is}}}$. By inserting this equation into eq. (15) and taking a further geometric mean with respect to j eq. (32) is obtained.

5.3 Planning the numbers of service providers

The task of capacity planning is to determine the numbers of service providers, $\bar{A}_{js}$, needed for all js. The method proposed builds on the equations (26), (27), and (29). The necessary data are obtained as follows. Population and morbidity data, I_i and $\mu_i y$, can be assumed to be known. Planned aggregate supplies, $\bar{S}_s$, have to be set by the planning authority. The parameter θ_s is estimated by relying on eq. (15). Patients' willingness to cover distance, φ_s, has to be determined by surveys. The equation $\omega_s = -\theta_s \cdot \varphi_s^2/3$ then allows computing the elasticity of spatial substitution. As we see no way to determine the ratios of preference weights, ρ_s, empirically, we set these parameters equal to one. With this simplifying assumption, all the quantities one needs to determine $\bar{A}_{js}$ are given.

6 The case of Germany

Germany lends itself to a study of health care capacity planning because the concept of "equal living standards" across the whole country has constitutional status. There is a free choice of doctors (no strict gatekeeping) and in the almost complete assumption of costs for all services offered by the statutory health insurance (SHI) system. It has to be noted, however, that there is a strict separation between ambulatory and hospital care in Germany. Ambulatory capacity planning is carried out nationally, but at a small-scale level, on the basis of the so-called Capacity Planning Guideline (Bedarfsplanungs-Richtlinie). The reference values against which current doctor-density ratios are measured are the ratios of doctors to population in 23 different medical specialities determined on a certain date in the past. They constitute the policy target for the number of ambulatory physicians in a full-time position to be financed by the SHI system.

In contrast, hospital capacities are planned at the level of Germany's sixteen federal states. The provision of hospital beds is planned politically usually using the Hill-Burton formula. This means that the target number of beds is derived from updating current values. This method is known to have the effect that the greatest capacity need is assessed where the existing capacity is highest.

The strict separation between ambulatory and hospital care is a characteristic feature of the German health care system and means that specialist capacities are available in both sectors. In practice, there is much overlap that is only partially taken into account in planning and ultimately leads to efficiency losses (Büyükdurmus et al., 2017). The separation of ambulatory and hospital capacity planning causes

particular problems for rural areas, as hospital planning aims at exploiting economies of scale and thus gives priority to urban areas.

For illustrating our method, we use data on the ambulatory sector provided by the National Association of Statutory Health Insurance Physicians (Kassenärztliche Bundesvereinigung, KBV) for the years 2015 and 2016. For hospital care, we use the hospitals' quality reports and an analysis of hospital billing (DRGs) provided by the Federal Statistical Office (Destatis). From the above sources we receive information about the number of treatment cases whose costs have been reimbursed by the SHI. We interpret those numbers as patients' take up, x_{ijs}. We determine the distance between service providers and patients' home addresses, d_{ij}, by means of postcode-to-postcode routing, using OpenStreetMap data. Values of planned aggregate supplies, $\overline{S}_s$, are taken from the target values reported by the German Capacity Planning Guideline. As no target value is published for hospital care, the missing value is approximated by the current take-up, $\sum_i D_{is}$. Population and morbidity data, I_i and $\mu_i y$, are taken from Capacity Planning Guideline reference values.[4] See appendix for a detailed description of data.

6.1 Estimating the size of catchment areas

Table 1 shows regression results for catchment areas. At roughly 24 minutes, the measured catchment area is smallest for GPs, followed by gynaecologists with 30 minutes, paediatricians with 36, and ophthalmologists with 37 minutes. At 56 minutes the measured catchment area is largest for psychotherapists, followed by urologists (49 minutes). With a 95% bound applied to distance weighting, the measured size of catchment areas largely confirms the values in Fülöp et al. (2011). Significant differences only exist for urologists - 49 here vs. 31 minutes in Fülöp et al., (2011) - and psychotherapists (56 vs. 62 minutes)[5].

Patients' willingness to travel is taken from Sundmacher et al. (2018). Catchment area sizes turn out to be smaller than patients' willingness to travel, which means an elasticity of spatial substitution larger than one. Due to lack of estimates for hospitals, catchment areas are set at 60 minutes and the elasticity ω_s at 1.5. Robustness checks reveal that projected capacity imbalances are not insignificantly dependent on the choice of ω_s. We justify our choice of 1.5 as follows.

The spatial elasticities of substitution range from 1.05 to 1.99 for outpatient doctors, with an average value of 1.45. A value smaller than 1.05 seems implausible for hospitals, because it would mean that treatments can be planned better by hospitals than by outpatient doctors. Such an assumption is not compatible with the emergency function of hospitals. Rather, it speaks for a higher elasticity. However,

[4] Values taken from the German Capacity Planning Guideline (17.12.2020): https://www.g-ba.de/richtlinien/4/.
[5] The differences have an explanation. Fülöp et al. (2011) base their analysis on selected German states and on data from 2010 counting multiple visits as one. As to psychotherapists, the difference could also result from an increase in their number.

the elasticity should also not significantly exceed the maximum value of 1.99 for outpatient doctors. In our opinion, a value of 1.5 is an appropriate compromise. For the sake of comparison, the supplement shows projected capacity imbalances for $\omega_s = 1.0$ and 2.0. The results show that with increasing elasticity, more capacities are allocated from urban to rural areas, both in inpatient and outpatient health care. This is an implication of the objective of sustaining an equal access to health care and the fact that individuals respond more strongly to distances when the spatial elasticity of substitution increases. The changes in the projected capacity imbalances also tend to decrease with increasing elasticity values. The difference from 1.0 to 1.5 is more pronounced than from 1.5 to 2.0. The explanation is that an elasticity of 1.0 is a rather extreme assumption which, according to eq. (5), implies a location-independent index of wasted time. Only when the elasticity of spatial substitution with respect to hospitals deviates from 1.0 do hospitals influence the spatial allocation of total wasted time and *a fortiori* the spatial allocation of the required capacities in all specialities. Loosely speaking, the outpatient doctors have to follow the hospitals.

Table 1: Regression results for catchment areas

	Distance decay parameter	**Size of catchment area in minutes**		**Patients' willingness to travel**	**Elasticity of spatial substitution**
	$\widehat{\theta}_s$	$\widehat{\sqrt{\frac{1}{\omega_s}\varphi_s}}$	**Confidence interval**	φ_s	ω_s
General Practitioners	-0,005	23.64	23.62 – 23.66	32.16	1,85
Ophthalmologists	-0,002	36.89	36.55 – 37.24	40.21	1,19
Surgeons	-0,002	39.91	39.40 – 40.44	51.27*	1,65
Gynaecologists	-0,003	30.12	30.08 – 30.16	42.50	1,99
ENT-specialists	-0,002	37.28	36.98 – 37.59	41.29	1,23
Dermatologists	-0,001	47.45	47.03 – 47.89	51.27*	1,17
Paediatricians	-0,002	35.99	35.71 – 36.27	46.70	1,68
Neurologists	-0,002	43.77	42.96 – 44.63	44.89	1,05
Orthopaedic specialists	-0,002	38.11	37.75 – 38.49	51.27*	1,81
Psychotherapists	-0,001	56.46	55.03 – 58.00	62.31	1,22
Urologists	-0,001	49.09	48.39 – 49.82	51.27*	1,09
Hospitals		60.00+			1,50

Note: + The size of hospitals' catchment area is simply set at 60 minutes. An empirical estimate fails due to missing data. * The entry is the one reported for "other physicians".

6.2 Estimating access to care

With the following visualizations of the German situation, we want to contrast the demand-oriented with the FCA-based measurement of access to care. For the purpose of illustration, we focus on general practitioners (GPs). The supplement provides visualizations of further specializations.

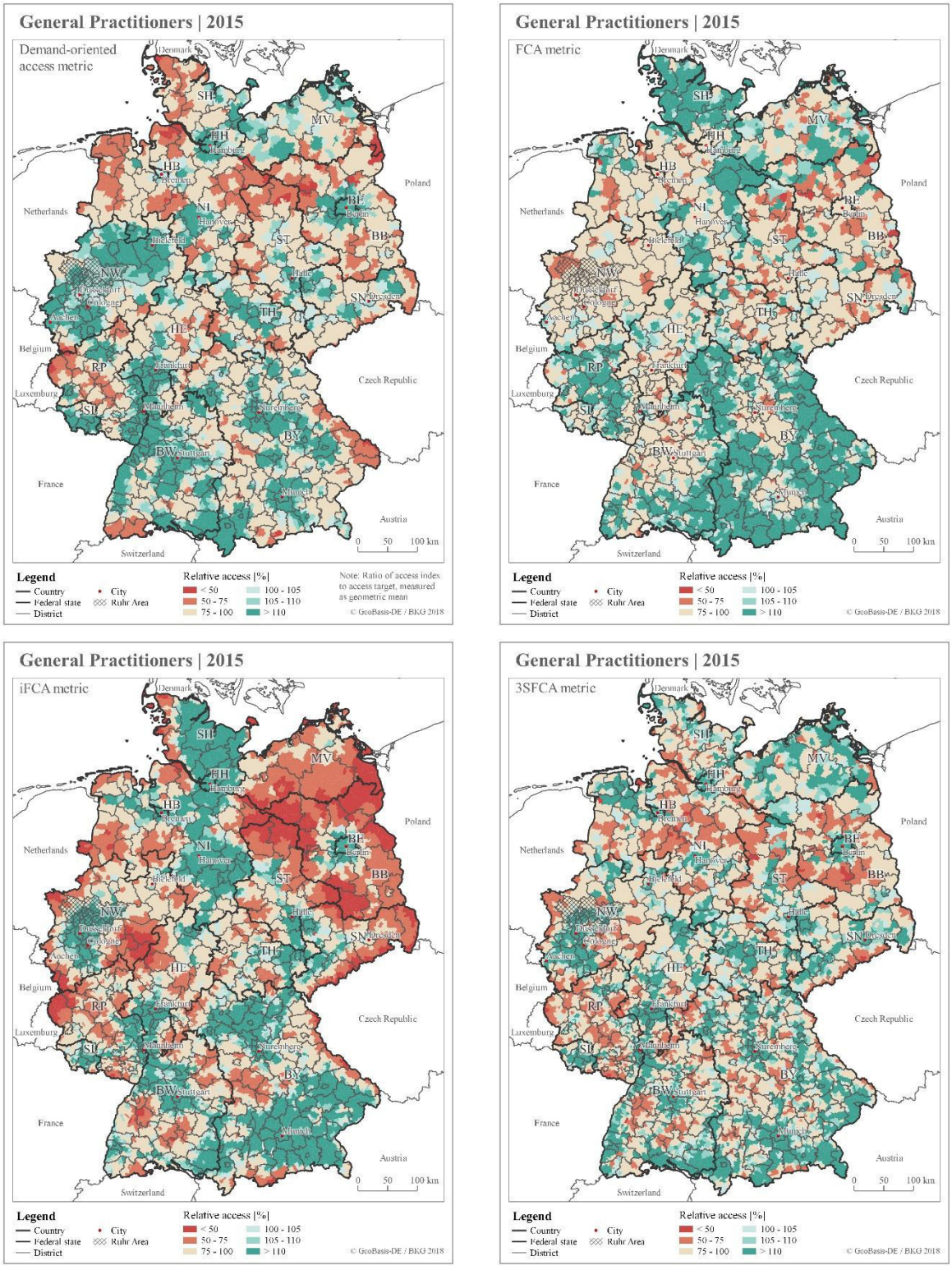

Figure 1: Access to General Practitioners
(top left: demand-oriented metric; top right: FCA metric; bottom left: iFCA metric; bottom right: 3SFCA metric)

The top left of Figure 1 relates to the demand-oriented access metric and the top right to the FCA metric. iFCA is visualized by bottom left and 3SFCA by bottom right. Green stands for oversupply and brown for undersupply. The first striking impression is the strong differences in the colouring. The regional

identification of under- and oversupply is almost opposing between top left and top right. The visualizations of iFCA and 3SFCA are closer to the demand-oriented access metric. However, iFCA has significantly more extreme values whereas 3SFCA draws a more disrupted and paler picture. It is worth mentioning that the results relating to iFCA confirm earlier ones of Bauer et al. (2018). Due to the same methodology and comparable data, this is hardly surprising.

According to the demand-oriented access metric, iFCA, and 3SFCA, larger cities tend to enjoy better access to GPs just as one would expect. Metropolises such as Hamburg, Munich, Stuttgart as well as the agglomerations of the Rhine-Main[6] and Rhine-Ruhr[7] areas are shown to enjoy particularly high levels of access. According to the FCA metric, the exact opposite is supposed to apply. The Rhine-Ruhr area and larger cities like Berlin, Munich and Hamburg are allegedly regions of more or less severe undersupply. This is not plausible and can only be explained by the FCA's missing sense of spatial directions in the exchange of services (Bauer et al., 2018). The mentioned cities and regions are known for their high capacities serving as medical care centres for their surroundings. The FCA's symmetrical distancing function does not do justice to the different roles of cities and environs. Unlike 3SFCA, iFCA increases not only the weight of distances, but also that of local supply. This apparently has the effect that the difference in access to care between sparsely populated regions and nearby agglomerations appears stronger. Examples of such spatial relationships are Brandenburg with Berlin, East Saxony with Dresden and Leipzig and Rothaar Mountains with Frankfurt and the Ruhr area.

In order to work out the differences in the competing access metrices quantitatively, Table 2 presents Spearman's rank correlations.

Table 2: Rank correlations between common measurements of access to care and demand-oriented measurement

	FCA	iFCA	3SFCA
General Practitioners	8,6%	56,3%	55,2%
Ophthalmologists	23,1%	49,6%	53,7%
Surgeons	30,3%	53,5%	52,3%
Gynaecologists	21,2%	57,0%	66,1%
ENT-specialists	25,2%	52,2%	55,7%
Dermatologists	8,2%	61,8%	66,1%
Paediatricians	27,4%	55,4%	61,7%
Neurologists	21,2%	58,7%	55,0%
Orthopaedic specialists	24,4%	54,6%	64,2%
Psychotherapists	23,2%	62,3%	64,7%
Urologists	35,0%	69,4%	69,4%
Hospitals	21,0%	41,2%	26,0%

[6] Metropolitan area around the city of Frankfurt including Mayence, Wiesbaden, Worms, Darmstadt, Aschaffenburg.

[7] Ruhr area plus the cities of Düsseldorf, Cologne, and Bonn. The Ruhr area is marked in Figures 2-4 by gridlines.

The correlation coefficients are relatively small indicating significant differences in detail. This speaks against the use of FCA-based metrics when measuring and planning access to care. This result is best explained by looking at the stylized model of two geographical entities. One of these, denoted by A, has all the features of a large city, i.e. a high supply of services and a population which can satisfy its demand locally. The second, denoted by B, has a small population, a low supply of services and many people demanding medical treatment in A. Because of the spatial symmetry, an unadjusted FCA-based measurement would indicate an equal level of access to care at both places. According to the competing measurements, the large city A would have to be considered the one with better access. This result is obtained for 3SFCA because distances are short in cities whereas it is obtained for iFCA because cities are also better served. The demand-oriented access metric additionally takes into account the important aspect that a larger share of demand is satisfied locally which helps to reduce wasted time.

6.3 Planning the number of health care providers

Table 3 compares available capacities of health care with those considered necessary when calculated using eq. (29). Across the country, there is overcapacity in all specialties ranging from 3.6% for general practitioners to almost 40% for psychotherapists.[8] This result, however, cannot surprise the experts and is consistent with the known fact that official planning figures are lower than the current numbers of active physicians (Kopetsch and Schmitz, 2014). A closer look at the data offers a more differentiated picture: First, capacity deficits are typical for rural areas, but overcompensated by overhangs in urban areas. Secondly and a bit surprising, even in rural areas there are too many surgeons rather than too few. About every 20th surgeon there seems dispensable. Thirdly, in urban areas, there is a marked oversupply in all specialities. The largest oversupply of about 60% concerns psychotherapists, orthopaedists, dermatologists and ENT specialists. The smallest oversupply seems to be for general practitioners and hospital beds, at around 25%. Fourthly, in smaller towns, the supply is more balanced. This is especially true for ophthalmologists, paediatricians, neurologists and general practitioners. Overcapacity of about 20% and more is found among surgeons, orthopaedists and dermatologists whereas undersupply holds for hospital beds and general practitioners.

- Table 3 about here -

The demand-oriented results of Table 3 suggest that a targeted reallocation of capacities could considerably improve equality in the access to health care. There are two possible counterarguments to this conclusion. First, health care facilities are not easily set up or closed, investments are largely irreversible, and due account must be taken of the proximity of other facilities and institutions such as universities. Secondly, the advantage of easy access has to be traded off against the economies to be achieved by centralisation. Indeed, many specialized treatments can only be safely performed if they are

[8] Hospitals are excluded here because of the assumption that all available beds are required.

conducted frequently enough for the staff involved to acquire and maintain sufficient experience.[9] Nevertheless, demand-oriented planning should help to identify imbalances in the spatial allocation of health care capacities.

7. Conclusions

The model presented in this paper is designed to improve health care capacity planning. It combines the trade model developed by Anderson and van Wincoop (2003) with the approach developed by Bikker and de Vos (1992) for modelling the demand for inpatient care. The advantage is practical as well as methodological. It is methodological because the demand for medical services and the competition for capacities of supply are endogenously determined in a general equilibrium framework. This advantage has a practical dimension because endogenous modelling is an important prerequisite for consistently predicting the impact of changed supply capacities on the demand and competition for them. In detail, we obtain the following results.

First, our demand-oriented approach suggests measuring access to health care by the time patients waste seeking treatment. The derived index is shown to be inversely proportional to the product of two geometric means. One is taken of cross-border demand ratios and the other of distances (Proposition 2). The most common approaches to measuring access to care build on the FCA metric. The shortcomings of the FCA metric have been a frequent topic in the literature and a number of attempts have been made to correct them (e.g. Delamater, 2013). Our metric generalizes the FCA by allowing adjustments of the population for morbidity and time wasted on waiting for medical treatment (Proposition 3). The need to adjust for waiting time is less obvious than that for morbidity. However, waiting time reflects competition for supply capacities. The insight into the need to consider such competition drives the proposal of the 3SFCA metric. While this refinement of the FCA metric emerged from an *ad-hoc* consideration, adjusting the population for waiting time is a consistently modelled improvement.

Secondly, the presented approach allows characterizing the demand for health care in the form of a gravity equation (Proposition 1). This equation is the empirical basis for measuring access to health care as well as for planning health care capacities.

Thirdly, we propose a test strategy for making an informed choice between the demand-oriented and the FCA-based metrics.

Fourthly, we demonstrate that our approach is well suited to improve practical capacity planning. Thus, we show that the gravity equation derived from our model can be used for estimating the size of catchment areas. While the literature mostly postulates ad-hoc values (e.g. 30 minutes), we determine

[9] The reimbursement of certain treatments by the SHI system requires hospitals to perform a minimum number of such interventions per year.

plausible values that are consistently estimated from our model. We finally point out how to compute the number of service providers needed to ensure an equal access to care.

The merits of our approach are illustrated by applying it to the German health system. Confirming findings in the literature, we show that there is a clear urban-rural divide. Rural areas suffer from undersupply whereas urban areas enjoy considerable oversupply. Aggregating deficits and overhangs, overcapacities are shown to exist in all specialities. A closer look at the data reveals a more differentiated picture.

No study of this kind would be complete without mentioning caveats. We do not include quality of care aspects in the model although quality differences definitely affect patient choices. However, we believe that quality cannot be the primary concern of capacity planning. There exist better-targeted instruments, such as pay-for-performance. Our empirical analysis also excludes individuals who did not consult an SHI doctor in the years 2015/16. This may bias the results, especially in regions with a high proportion of patients with private, rather than statutory, health insurance. A final caveat concerns the size of catchment areas, which may well be underestimated. The reason is that journey times are computed by using the postal codes of billing addresses. As a result, the travel time within the same postal code area is zero, a technical simplification which obviously underestimates some actual travel times. Spatial substitutability will equally be underestimated as a result.

Despite all these limitations, however, we strongly believe that our choice-theoretic grounding of the gravity equation and the empirical analysis of the German situation based on this equation make a significant contribution towards bringing health care planning more in line with patients' needs.

8. References

Anderson, J., 1979. A theoretical foundation for the gravity equation. The American Economic Review 69, 106–116.

Anderson, J., 2011. The gravity model. Annual Review of Economics, 133–160.

Anderson, J., van Wincoop, E., 2003. Gravity with gravitas: A solution to the border puzzle. The American Economic Review 93, 170–192.

Bauer, J., Maier, W., Müller, R., Groneberg, D.A., 2018. Hausärztliche Versorgung in Deutschland – Gleicher Zugang für alle? Deutsche Medizinische Wochenschrift 143, e9-e17.

Bergstrand, J.H., 1985. The gravity equation in international trade: Some microeconomic foundations and empirical evidence. The Review of Economics and Statistics 67, 474–481.

Bikker, J.A., de Vos, A.F., 1992. A regional supply and demand model for inpatient hospital care. Environment and Planning A 24, 1097–1116.

Büyükdurmus, T., Kopetsch, T., Schmitz, H., Tauchmann, H., 2017. On the interdependence of ambulatory and hospital care in the German health system. Health Economics Review 7, 2.

Chilès, J.P., Delfiner, P., 1999. Geostatistics: modelling spatial uncertainty. John Wiley & Sons, Hoboken, New Jersey.

Czihal, T., von Stillfried, D., Schallock, M., 2012. Regionale Mitversorgungsbeziehungen in der ambulanten Versorgung. Versorgungsatlas. Zentralinstitut für die kassenärztliche Versorgung in Deutschland, Berlin.

Delamater, P., 2013. Spatial accessibility in suboptimally configured health care systems: a modified two-step floating catchment area (M2SFCA) metric. Health & Place 24, 30–43.

Fülöp, G., Kopetsch, T., Schöpe, P., 2011. Catchment areas of medical practices and the role played by geographical distance in the patient's choice of doctor. The Annals of Regional Science 46, 691–706.

Huff, D.L., 1964. Defining and estimating a trading area. Journal of Marketing 28.

Joseph, A.E., Bantock, P.R., 1982. Measuring potential physical accessibility to general practitioners in rural areas: A method and case study. Social Science & Medicine 16, 85–90.

Kopetsch, T., Schmitz, H., 2014. Regional variation in the utilization of ambulatory services in Germany. Health Economics 23, 1481–1492.

Leamer, E.E., Stern, R.M., 1970. Quantitative International Economics. Allyn and Bacon, New Brunswick, New York, 209 pp.

Lindsay, C.M., Feigenbaum, B., 1984. Rationing by Waiting Lists. The American Economic Review 74, 404–417.

Lowe, J.M., Sen, A., 1996. Gravity model application in health planning: analysis of an urban hospital market. Journal of Regional Science 36, 437–461.

Luo, J., 2014. Integrating the Huff model and floating catchment area methods to analyze spatial access to healthcare services. Transactions in GIS 18, 436–448.

Luo, W., Qi, Y., 2009. An enhanced two-step floating catchment area (E2SFCA) method for measuring spatial accessibility to primary care physicians. Health & Place 15, 1100–1107.

OECD, 2020. Waiting Times for Health Services: Next in Line. OECD Health Policy Studies, Paris.

Ono, T., Lafortune, G., Schoenstein, M., 2013. Health workforce planning in OECD countries: A review of 26 projection models from 18 countries. OECD Health Working Papers 62, 1–130.

Paez, A., Higgins, C.D., Vivona, S.F., 2019. Demand and level of service inflation in Floating Catchment Area (FCA) methods. PLOS ONE 14, e0218773.

Redding, S.J., 2010. The Empirics of New Economic Geography. Journal of Regional Science 50, 297–311.

Sá, L., Siciliani, L., Straume, O.R., 2019. Dynamic hospital competition under rationing by waiting times. Journal of Health Economics 66, 260–282.

Sen, A., 2002. Why health equity? Health Economics 11, 659–666.

Sundmacher, L., Schang, L., Schüttig, W., Flemming, R., Frank-Tewaag, J., Geiger, I., Franke, S., Weinhold, I., Wende, D., Kistemann, T., Höser, C., Kemen, J., Hoffmann, W., van den Berg, N.,

Kleinke, F., Becker, U., Brechtel, T., 2018. Gutachten zur Weiterentwicklung der Bedarfsplanung i.S.d. §§ 99 ff. SGB V zur Sicherung der vertragsärztlichen Versorgung. order by Gemeinsamen Bundesausschusses (GBA), München, Köln, 810 pp.

Wan, N., Zou, B., Sternberg, T., 2012. A three-step floating catchment area method for analyzing spatial access to health services. International Journal of Geographical Information Science 26, 1073–1089.

Table 3: Predicted demand and projected capacity imbalances

Speciality	Capacities in urban areas			Capacities in minor towns			Capacities in rural areas			Overall capacities		
	Required (1)	Available (2)	$\frac{(1)-(2)}{(2)}$	Required (3)	Available (4)	$\frac{(3)-(4)}{(4)}$	Required (5)	Available (6)	$\frac{(5)-(6)}{(6)}$	Required (7)	Available (8)	$\frac{(7)-(8)}{(8)}$
General Practitioners	18194	24520	-25,8%	16976	15817	7,3%	14601	11291	29,3%	49771	51628	-3,6%
Ophthalmologists	1514	2696	-43,8%	1499	1535	-2,3%	1313	1048	25,3%	4326	5279	-18,1%
Surgeons	883	1872	-52,8%	883	1195	-26,1%	768	813	-5,5%	2534	3880	-34,7%
Gynaecologists	2433	5143	-52,7%	2347	2803	-16,3%	2058	1802	14,2%	6838	9748	-29,9%
ENT-specialists	931	2115	-56,0%	924	1069	-13,6%	809	726	11,4%	2664	3910	-31,9%
Dermatologists	741	1728	-57,1%	747	932	-19,8%	637	549	16,0%	2125	3209	-33,8%
Paediatricians	1675	2845	-41,1%	1653	1693	-2,4%	1450	1075	34,9%	4778	5613	-14,9%
Neurologists	1314	2575	-49,0%	1309	1374	-4,7%	1105	878	25,9%	3728	4827	-22,8%
Orthopaedic specialists	1190	2812	-57,7%	1184	1552	-23,7%	1035	1006	2,9%	3409	5370	-36,5%
Psychotherapists	5038	13720	-63,3%	5080	6202	-18,1%	4346	3342	30,0%	14464	23264	-37,8%
Urologists	669	1336	-49,9%	673	791	-14,9%	572	516	10,9%	1914	2643	-27,6%
Hospitals	221399	307828	-28,1%	223280	192688	15,9%	191727	135890	41,1%	636406	636406	0,0%

Note: Capacities are measured by the number of doctors in full-time equivalents with the exception of hospitals, which are measured by the number of beds.