ipea Instituto de Pesquisa Econômica Aplicada

Título do capítulo	CAPÍTULO 2 **ENTENDENDO OS CONCEITOS DE EFICIÊNCIA EM SAÚDE 60**
Autores (as)	Alexandre Marinho
DOI	

Título do livro	SUS: AVALIAÇÃO DA EFICIÊNCIA DO GASTO PÚBLICO EM SAÚDE
Organizadores (as)	Carlos Octávio Ocké-Reis
Volume	
Série	
Cidade	
Editora	Instituto de Pesquisa Econômica Aplicada (Ipea)
Ano	2023
Edição	1ª
ISBN	
DOI	

Entendendo os conceitos de eficiência em saúde

Alexandre Marinho

Introdução

No presente capítulo, apresentamos uma visão geral dos problemas da avaliação de eficiência no setor saúde. Sempre que possível, evitaremos entrar em tecnicalidades excessivas, embora algumas vezes isso não seja evitável, dada a complexidade do tema. Também tentaremos, na medida do possível, dar aos leitores uma perspectiva das possibilidades de avaliação de eficiência no Brasil, mas adiantamos que uma aplicação específica será detalhadamente explorada no capítulo 7 deste livro.

São várias as razões pelas quais as avaliações de eficiência devem ser realizadas. De modo não exaustivo, passamos a discutir brevemente algumas delas.

O setor saúde e, em uma perspectiva mais próxima aos objetivos deste livro, a prestação de cuidados médicos à saúde não configuram um mercado com competição perfeita. Se isso ocorresse, valeria o Primeiro Teorema do Bem-Estar Social[I], e a eficiência no sentido de Pareto estaria garantida. Eficiência no sentido de Pareto é um conceito importante em Economia, e a ele voltaremos neste capítulo. Por enquanto, dizemos que existe eficiência no sentido de Pareto se uma alocação de recursos é tal que não é possível melhorar qualquer agente econômico sem piorar algum outro[II]. A discussão das implicações das 'imperfeições' (presença de externalidades; bens públicos; bens de mérito; *moral hazard*; seleção adversa; distorção das preferências; assimetria de informações; dificuldades de aprendizado; oligopólios; instituições sem fins lucrativos etc.[III]) dos mer-

I Esse teorema pode ser visto com detalhes técnicos em Varian[1].

II Nesse caso, temos eficiência forte de Pareto. Se não for possível melhorar todos os agentes sem piorar algum agente, temos eficiência fraca de Pareto.

III Esses conceitos serão mais bem explorados ao longo deste capítulo.

cados de saúde foi estabelecida, inicialmente, de modo concatenado, no texto pioneiro de Arrow[2]. No que se refere, em particular, às imperfeições dos mercados de planos de saúde e as implicações sobre o Sistema Único de Saúde (SUS), no Brasil, encontramos discussões em Ocké-Reis[3], Bahia[4], Andrade[5] e Marinho[6].

Paralelamente, a utilização do conceito de eficiência e de suas medidas não está aqui lastreada apenas na teoria econômica. Ela é obrigatória no setor público brasileiro. A Constituição federal de 1988[7], no *caput* de seu § 37, determina:

> A administração pública direta e indireta de qualquer dos Poderes da União, dos Estados, do Distrito Federal e dos Municípios obedecerá aos princípios de legalidade, impessoalidade, moralidade, publicidade e **eficiência** [...]. (grifo nosso).

Neste capítulo, apresentaremos, brevemente, os conceitos de eficiência, e alguns outros conceitos correlatos indispensáveis para a sua compreensão. Estes se originam na teoria econômica, notadamente, na microeconomia. Essa peculiaridade nos levou a ter de trabalhar alguns detalhes imprescindíveis para que o leitor não economista possa acompanhar os métodos de cálculo de eficiência apresentados no capítulo 6, bem como a aplicação empírica desses conceitos e métodos no SUS, desenvolvida no capítulo 7. Como grande parte do interesse da avaliação de eficiência decorre da percepção – correta ou não –, bastante difundida na sociedade, de que saúde custa cada vez mais "caro" e que os resultados não acompanham essa elevação de custos, assim como que grande parte dessa suposta "culpa" desse descompasso recairia sobre sistemas públicos de saúde – culpabilidade essa com a qual não concordamos, por razões que este livro tenta demonstrar –, iniciamos o capítulo discutindo essa elevação de custos em saúde *vis-à-vis* os respectivos resultados.

Custos crescentes com resultados duvidosos. Qual a explicação?

Se os custos fossem crescentes e os resultados fossem excelentes e crescentes, talvez não houvesse tantas preocupações. Haveria a possibilidade de, pelo já citado Primeiro Teorema do Bem-Estar, a alocação geral de recursos ser a melhor possível (ótima) e atingida a eficiência no sentido de Pareto. Contudo, como esse não é o caso em saúde, cabe entender um pouco melhor as razões para os custos serem crescentes, e os resultados serem duvidosos. Nas linhas abaixo, arrolamos algumas dessas razões:

1. Envelhecimento da população: as pessoas, em média, vivem cada vez mais. Entretanto, elas sofrem de doenças crônicas na velhice (e até mesmo antes).

2. Informação: com a disseminação do uso da internet e das redes sociais, houve um aumento expressivo da informação disponível (boa ou ruim) sobre saúde. As expectativas (infundadas ou não) e as escolhas disponíveis (reais ou não) das pessoas aumentaram, levando à adoção de métodos de diagnósticos, prescrições e tratamentos muitas vezes caros e pouco resolutivos, que as pessoas exigem dos prestadores de serviços de saúde públicos e privados.

3. Os direitos da cidadania: no Brasil, a saúde de cada cidadão brasileiro é um direito constitucional, de acordo com o art. 196 da Constituição federal vigente[7]. Por causa disso, e do aumento das informações disponíveis, as pessoas exigem tudo que existe (ou que elas acham que existe) em saúde. Uma consequência é a judicialização da saúde e a expansão muitas vezes injustificadas dos custos. No entanto, as pessoas não se conformam mais em receber abaixo do que julgam ser os seus direitos, devido pelo Estado. Elas, geralmente, não aceitam passivamente morrer sem esgotar todos os recursos que, pela Constituição federal brasileira, são limitados apenas pelas tecnologias existentes (inclusive no exterior, segundo algumas interpretações da justiça brasileira).

4. Problemas informacionais e comportamentais dos pacientes e prestadores de serviços: a presença de informação assimétrica. Nesse caso, os profissionais de saúde e os pacientes detêm informações diferentes sobre o adoecimento. Os profissionais têm conhecimento técnico e científico sobre as doenças maior do que o detido pelos pacientes, mas esses, por sua vez, podem omitir, ou não saber, detalhes da história sanitária deles mesmos ou da sua família. Por causa disso, contratos (no caso privado) ou decisões erradas, excessivamente agressivas, ou defensivas, podem levar a custos maiores do que o necessário. O *moral hazard* (dano moral) ocorre de duas formas: porque o paciente, ao dispor do seguro social, ou direito constitucional, pode passar a se comportar de modo mais arriscado, aumentando as chances de adoecimento; ou porque ele passa a usar os serviços de saúde mesmo quando não seria necessário (*overuse*), aumentando os custos pela mera utilização, ou até pela submissão a tratamentos desnecessários.

5. Problema similar ao *moral hazard,* mas pelo lado da oferta de serviços, o agenciamento (*agency*) ocorre quando os serviços de saúde atuam em excesso, ou também de modo desnecessário, dado que os pacientes não pagam a conta diretamente (ou integralmente) no momento da utilização. Um exemplo: o Brasil, por razões financeiras, culturais, e de conveniência de pais e prestadores de serviços, tem elevada prevalência de partos cesáreos, com consequências nada positivas em termos sanitários[8-9]. Vale acentuar que o uso insuficiente dos serviços de saúde (*underuse*) também pode

acarretar custos totais maiores do que o nível eficiente no longo prazo. Nessa hipótese, a prevenção falha, e a cura sai muito mais cara no futuro. Um exemplo é a enorme quantidade de transplantes feitos no SUS devido à alta prevalência de tabagismo, diabetes, hipertensão e obesidade na população, que poderiam ser evitados ou bastante reduzidos, com educação e com uma atenção básica resolutiva e efetivamente integrada com os níveis superiores de atenção. O mau uso, ou uso errado (*misuse*), também pode ocasionar custos maiores, por diversas razões. A automedicação é um exemplo clássico. Falhas diversas do sistema de saúde também podem acarretar elevações de custos: erros de médicos e outros profissionais, exposições excessivas às radiações ionizantes etc. Outro problema muito retratado é a criação e utilização de novas tecnologias pouco eficientes, sem o controle efetivo dos gestores da saúde. Além da relação custo-benefício desfavorável, muitas vezes, as novas tecnologias são utilizadas de modo concomitante e superposto com as tecnologias antigas, sem substituição, com resultados incrementais desprezíveis, diante de aumentos de custos nada desprezíveis. Na ânsia de parecer moderno para atrair pacientes, e de competir por 'melhores' médicos (*medical arms race*), essas novas tecnologias são incorporadas de modo ineficiente – e muitas vezes definitivo – aos sistemas de saúde.

6. O setor saúde é intensivo em mão de obra e apresenta baixa produtividade do trabalho, pois grande parte dos tratamentos não pode ser feita por máquinas ou prescinde de mão de obra altamente qualificada, haja vista que o trabalho é o produto ofertado, com poucas possibilidades de substituição. Ademais, a introdução de novas tecnologias, a despeito de aumentarem a precisão dos diagnósticos, e de reduzirem os tempos de tratamento (incluindo internações) e de consultas, não substitui totalmente as tecnologias antigas, e não reduz as necessidades de emprego de mão de obra[10-11]. Os recentes progressos nos aplicativos de atendimento aos pacientes oferecem algumas possibilidades de substituição.

O remédio proposto pelo senso comum para mitigar o problema dos custos crescentes sem resultados compatíveis é o aumento da eficiência, traduzido coloquialmente como "fazer mais com menos" ou "produzir mais gastando menos". Essa prescrição costuma vir ancorada em uma suposição errônea – e muitas vezes com objetivos políticos e econômicos não explícitos – de que ser mais eficiente é cortar custos. Há uma lógica aparente nesse argumento: se os custos são elevados, bastaria cortar os custos. Assim, o conceito de eficiência, ou a busca da eficiência, teria uma simplicidade meridiana.

A despeito da aparente simplicidade do conceito de eficiência econômica, há bastante confusão sobre o real significado dele. Os termos eficiência, efetividade, eficácia e produtividade são, frequentemente, usados de modo intercambiado. Ainda

assim, eles não significam exatamente a mesma coisa. Entretanto, todos eles têm uma característica comum: referem-se aos recursos e aos resultados de uma unidade produtiva, ou seja, de um agente econômico, ou organização econômica, que transforma fatores de produção – os insumos – em produtos (ou, em *outcomes*, ou desfechos). A própria escolha dos insumos e dos resultados é uma tarefa complexa, que depende dos objetivos da avaliação e, obviamente, dos objetivos da organização analisada. Se todos os insumos e resultados desejados e indesejados de todas as organizações analisadas forem incluídos na análise, todas receberão o mesmo escore (nota) de eficiência. A grande questão é escolher quais insumos e resultados são mais relevantes para as organizações. Essa escolha pode ser mais importante até mesmo do que os métodos quantitativos utilizados. Por exemplo: uma lâmpada gera calor e luz. Se o objetivo for aquecer, o grau de eficiência terá um valor. Se o objetivo for iluminar, o grau de eficiência, provavelmente, será outro.

Vejamos agora alguns conceitos e noções, que serão tratados neste capítulo e ao longo do próprio livro. Conforme foi exposto na Introdução deste capítulo, não há como compreender precisamente o conceito de eficiência sem alguma discussão de conceitos correlatos ou antecedentes:

1. Insumos: são os bens materiais e imateriais (serviços) que são transformados em outros bens e serviços em um processo produtivo. São também chamados de fatores de produção. Sempre que possível, os insumos devem ser medidos como fluxos por unidade de tempo. Exemplo: horas-máquina; homens-hora. Todavia, é comum que sejam medidos em termos de estoque: por exemplo, ao usar a quantidade de médicos em vez das horas de trabalho médico, ou usar a quantidade de tomógrafos no lugar da quantidade de tomografias. O dinheiro não é um fator de produção *stricto sensu*, pois não é transformado no processo produtivo. Entretanto, é comum que valores monetários (representando custos ou despesas) sejam utilizados como se fossem um fator de produção, na ausência de registro de levantamento detalhado sobre os insumos usados, dado que um valor monetário é o produto de quantidades pelos preços de cada insumo.

2. Produtos (*outputs*): são os produtos finais que um agente econômico gera em seu processo produtivo (transformação de recursos em resultados). Infelizmente, nem sempre eles são indicadores finais de bem-estar. Nos produtos, também se aplica a necessidade, nem sempre respeitada, de mensuração de fluxos. Por outro lado, é bastante comum que tais indicadores de bem-estar nem mesmo existam, ou sejam de muito difícil mensuração. Por exemplo, é usual utilizar a quantidade de consultas, exames ou cirurgias como indicador de desempenho de hospitais, embora tais

indicadores não estejam, mesmo em termos per capita, necessariamente, conectados com o bem-estar dos pacientes.

3. Desfechos (*outcomes*): são os indicadores realmente finalísticos em saúde. No entanto, eles não são facilmente observáveis, mensuráveis e registráveis de modo sistemático. Por exemplo: a duração e a qualidade de vida após intervenções médicas; doenças evitadas; pressão arterial; e assim por diante.

4. Tecnologia: estabelece todas as combinações das quantidades dos diferentes insumos que são capazes de produzir determinadas combinações de produtos. Como as firmas não podem produzir qualquer coisa a partir de qualquer coisa, a tecnologia é uma restrição para as firmas. Em termos técnicos simplificados, uma tecnologia T é um conjunto de combinações de insumos e de produtos tal que x pode produzir y. $T=\{(x, y) \mid x$ pode produzir y$\}$. Frise-se que x e y são vetores (listas) com quantidades positivas de insumos e produtos respectivamente.

5. Função de produção: estabelece o máximo de produto que se pode obter a partir de diferentes quantidades de fatores de produção. Então, uma função de produção retrata as melhores escolhas da firma entre os diferentes processos produtivos possíveis, representando uma fronteira das possibilidades de produção. Tecnicamente, dizemos que $y=F(x)$. Exemplo: em um hospital, podemos estabelecer que a quantidade de cirurgias é função da quantidade de médicos; da quantidade de centro cirúrgicos; das quantidades de enfermeiros etc. Os economistas costumam descrever a função de produção usando fórmulas matemáticas que sejam capazes de representar bem o processo de transformação de insumos em resultados. No trecho economicamente relevante da função de produção[12], ela será crescente, ou seja, quanto maiores as quantidades usadas dos insumos, maior deverá ser a quantidade produzida (monotonicidade)[IV].

6. Custo: é o valor monetário empregado na produção de bens e serviços utilizados diretamente na produção de bens e serviços e devem depender de decisões relacionadas com a produção, mesmo que não visíveis imediatamente.

7. Função custo: representa o modo mais barato de produzir um determinado nível de produto. Aqui, já há uma certa noção de comportamento eficiente, pois sempre se pode produzir de modo mais caro do que o mínimo. Dizemos que $C=F(w, y)$, ou seja, a função custo depende dos preços dos insumos (w) e das quantidades produzidas (y)[V].

IV Para exemplos de funções de produção em saúde, consultar Butler[12].

V Para exemplos de funções custo e de produção em saúde, consultar Butler[12].

8. Eficiência (*Efficiency*): expressa a relação entre custos e benefícios ótimos e observados. Em termos operacionais, uma avaliação de eficiência deve resguardar a observação de critérios mínimos de efetividade. Isso porque, dependendo do modelo, não produzir e não gastar pode ser tecnicamente eficiente, mas socialmente indefensável. A eficiência pode ser medida com duas orientações: orientação para os produtos ou orientação para os insumos. Na orientação para os produtos, a eficiência é a razão matemática (e, portanto, um número puro ou adimensional) entre o máximo nível de produto possível e o produto observado. Em orientação para os insumos, será a razão matemática (número adimensional) entre o nível de insumos observado e o menor nível de insumos possível. Então, a eficiência será, nesses casos (há exceções), um número entre zero e a unidade (ou entre 0% e 100%). Contudo, resta questionar o que fazer em casos de produção de múltiplos produtos, a partir da utilização de múltiplos insumos. Para dirimir essa dúvida, recorremos ao conceito de eficiência de Pareto-Koopmans, doravante PK[13(p. 60)].

> Uma unidade produtiva é eficiente se um aumento em qualquer produto requer a redução da quantidade de pelo menos um outro produto, ou o aumento da quantidade de pelo menos um insumo; e se a redução de qualquer insumo requer um aumento da quantidade de pelo menos um outro insumo, ou a redução da quantidade de pelo menos um produto. Um produtor tecnicamente ineficiente poderia produzir as mesmas quantidades de todos os produtos utilizando menor quantidade de pelo menos um insumo, ou utilizar as mesmas quantidades de todos os insumos para produzir mais de pelo menos um produto.

Atendidos tais requisitos, entendemos o que significa dizer que uma unidade produtiva é eficiente no sentido de PK.

9. Dominância: seja o insumo x usado para produzir o produto y. Sejam duas firmas: a Firma 1 representada pelo par (insumo, produto) dado por (x_1, y_1), e a Firma 2 representada pelo par (insumo, produto) dado por (x_2, y_2). Dizemos que (x_2, y_2) domina (x_1, y_1) se, e somente se, $x_2 \leq x_1$, $y_2 \geq y_1$ e $(x_1, y_1) \neq (x_2, y_2)$. Então, a Firma 2 e a Firma 1 são diferentes, e a Firma 2 produz maior quantidade, em pelo menos um produto, sem produzir menor quantidade de qualquer produto. Exemplo: a Firma (x=1, y=2) domina a Firma (x=1, y=1); mas a Firma (x=1, y=2) não domina, nem é dominada, por (x=1, y=2) que é igual a ela. A Firma (x=1, y=2) também não domina nem é dominada por (x=2, y=3) que gasta maior quantidade de insumos, mas também produz maior quantidade de produto. A dominância implica, fracamente, que as preferências são crescentes nos produtos e decrescentes nos insumos, o que

significa que produzir mais é pelo menos tão bom quanto produzir menos, e que gastar menos insumos é pelo menos tão bom quanto gastar mais insumos. Firmas eficientes não podem ser dominadas por quaisquer outras firmas.

10. Efetividade: diz respeito à implementação e ao aprimoramento de objetivos. Em termos econômicos estritos, especifica uma função utilidade (medida de bem-estar) bem definida, que deve ser maximizada. Seja *A* uma unidade produtiva que produz o bem *Y*. A efetividade traduz a razão matemática entre a utilidade realmente obtida e a utilidade máxima possível. Então: Efetividade=$U_A(Y)/U_{ÓTIMO}$. Em termos operacionais, a efetividade está relacionada com a implementação e o aprimoramento de metas (ou o preenchimento das necessidades). As metas são a quantificação dos objetivos. A falta de informação sobre a função utilidade nos leva, com frequência, a avaliar a eficiência no lugar da efetividade. Em geral, em nossos termos, eficiência é uma condição necessária, mas não suficiente para a efetividade.

11. Eficácia: refere-se aos objetivos pretendidos, ou seja, o processo produtivo deve gerar os efeitos desejados, em outras palavras, os resultados esperados. A eficácia está relacionada com as condições controladas nas quais as atividades são concebidas e simuladas.

12. Produtividade: é o quociente obtido pela divisão da quantidade utilizada de um determinado produto pela quantidade utilizada de um determinado insumo. Exemplos: quantidade de cirurgias realizadas em um mês dividida pelas horas de trabalho médico gastas nas cirurgias; quantidade consultas ambulatoriais em um ano dividida pela quantidade de leitos ambulatoriais utilizadas no mesmo período. Cabem três observações importantes: a primeira é que, a rigor, sempre deveríamos utilizar os fluxos de serviços gerados pelo capital, quando o capital é um insumo. No caso de leitos hospitalares, por exemplo, deveríamos medir quantas horas, ou dias, os leitos foram utilizados, e não a quantidade de leitos em si, pois os leitos podem ter ficado ociosos em determinados momentos. Um leito ativo ocioso, que não gere serviços, afetaria a eficiência, mas, em uma contagem não muito rigorosa, não afetaria a produtividade, pois seria contado como insumo. Em segundo lugar, é importante assinalar que as quantidades se referem a insumos e produtos físicos, nunca financeiros, pois o dinheiro não sofre transformação no processo produtivo, nem gera fluxo de serviços (o dinheiro compra insumos), embora o tratamento de dinheiro como insumo seja frequente na literatura, inclusive no capítulo 6 deste livro, representando insumos que não foram devidamente registrados e não puderam ser listados explicitamen-

te[VI]. Por último, cabe assinalar que a produtividade é determinada pela eficiência; pela tecnologia; pelo ambiente (não apenas ambiente físico); e pelas aleatoriedades, ou choques aleatórios, que são eventos imprevistos positivos ou negativos, como os seguintes exemplos: greve de funcionários, falta de insumos, epidemias e desastres.

Medidas de eficiência

As medidas de eficiência são, todas, distâncias relativas[14-15], ou seja, uma métrica aplicada ao afastamento de uma unidade produtiva em relação a uma fronteira de eficiência estimada por determinados métodos quantitativos[VII].

Em saúde, frequentemente, não conhecemos exatamente a função de produção, que é aquela que determina o máximo de produto (vetor de quantidades) que pode ser obtido a partir de uma dada cesta (vetor de quantidades) de insumos. Então, recorremos à observação de referências virtuosas – os *benchmarks* (melhores práticas) – para estabelecer as quantidades ótimas de cada produto (no vetor de produtos) e as quantidades ótimas de cada insumo (no vetor de insumos) de cada firma. Por exemplo, não sabemos exatamente qual a quantidade ótima de cirurgias a ser produzida por um hospital, ou por um centro cirúrgico, a partir dos insumos que utilizamos (médicos, enfermeiros, materiais, medicamentos, equipamentos etc.). Então, podemos estabelecer uma coleção (amostra) de hospitais similares, e observar os menores valores de cada insumo, e os maiores valores de cada produto, para termos uma noção dos valores 'ótimos' possíveis para cada insumo e para cada produto. Assim, as melhores práticas (*benchmarks*) passam a servir como uma espécie de fronteira de eficiência. Essa fronteira determina os limites do conjunto de possibilidades de produção da amostra, embora com algumas restrições técnicas e operacionais.

Vejamos um exemplo hipotético muito simples em que dois hospitais produzem apenas cirurgias, a partir do uso de um único insumo: os médicos. Recorreremos frequentemente aos hospitais como nossas unidades produtivas, porque eles são uma unidade de saúde representativa em termos de produção de cuidados em saúde. Com o intuito de simplificar o exemplo, vamos supor que ambos os hospitais empregam dez médicos com as mesmas qualificações, habilidades, salários e cargas horárias. O eventual hospital 'mais eficiente' entre os dois hospitais da amostra fez, *ceteris paribus*, 80 cirurgias anuais utilizando 10 médicos. O outro hospital que também tem 10 médicos fez apenas

VI O papel do dinheiro no processo produtivo será tratado quando falarmos adiante em custos e em eficiência alocativa.

VII A discussão detalhada sobre essas funções de distância pode ser encontrada, por exemplo, em Fried, Lovell e Schmidt[14] e em Bogetoft e Otto[15].

40 cirurgias no período. O escore de eficiência do primeiro hospital, que produz maior quantidade de cirurgias, seria igual a 1,00 (ou 100%). Como a quantidade de médicos é a mesma em ambos os hospitais, somente precisamos comparar a quantidade de cirurgias que cada um deles realiza. Então, o escore de eficiência do segundo hospital é igual a 40/80=0,5 ou 50%. Esse seria um modelo orientado para produtos, porque somente olhamos para as quantidades produzidas do único produto, ou seja, as cirurgias. Um exemplo orientado para os produtos manteria fixa a quantidade de cirurgias e faria variar a quantidades dos insumos.

Essa abordagem, embora muito utilizada, e muito prática, tem pelo menos três problemas, que trataremos a seguir. O primeiro é que não sabemos se as unidades da amostra estão fazendo o seu melhor esforço. Vamos supor que, na verdade, o hospital mais eficiente, que fez 80 cirurgias, pudesse, na realidade, fazer 100 cirurgias. Todavia, por alguma razão (desleixo, greve, acidente etc.), ele fez apenas as 80 cirurgias que usamos no exemplo anterior. Nesse caso, temos uma fronteira de eficiência aquém do máximo realmente possível. De fato, o hospital que fez apenas 40 cirurgias recebeu um escore estimado de eficiência igual a 40/80=0,5. Contudo, como o máximo possível era de 100 cirurgias no hospital eficiente (ao invés das 80 que usamos no exemplo), ele deveria ter recebido um escore real igual a 40/100=0,4. Como o hospital eficiente não trabalhou direito, a nota (escore) do hospital ineficiente foi superestimada. Por isso, podemos dizer que fronteira de eficiência 'estimada' é viesada (benevolente) em relação à fronteira 'real'.

O segundo problema advém da própria complexidade dos modelos. No caso de um único produto e de um único insumo, o problema é matematicamente simples, como vimos no parágrafo anterior. No entanto, no caso de múltiplos produtos e de múltiplos insumos (produção múltipla), a solução não é mais tão simples, como no exemplo hipotético dos dois hospitais que somente produziam cirurgias, usando apenas médicos. Isso ocorre porque, com produção múltipla, a simples divisão das quantidades de cada produto pelas quantidades de cada insumo pode levar a resultados absolutamente divergentes. Nesse caso, como comentaremos, alguma forma de tratar a multiplicidade de variáveis será necessária. Por razões que veremos adiante, é comum que eficiência seja um número puro (escore) no intervalo fechado [0, 1] ou [0%, 100%], embora valores positivos fora desse intervalo também sejam admitidos. Em geral, quanto maior o valor da medida, mais eficiente será a unidade produtiva, embora exceções significativas possam ocorrer. Uma unidade plenamente eficiente terá, usualmente, um valor de eficiência igual a unidade ou 100%.

Existe um terceiro problema, que é de difícil solução. Cirurgias não são um produto único. Podemos ter desde correções de hernias até cirurgias de aneurismas de aorta rotos ou transplantes. Ou seja, vários produtos muito diferentes poderiam ser

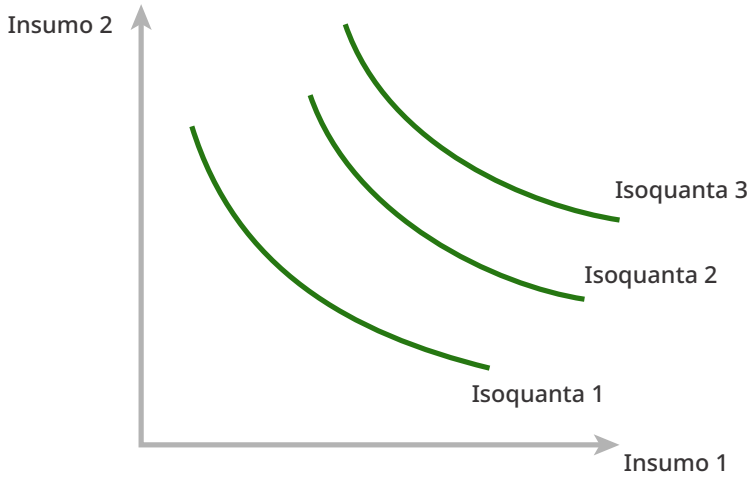
classificados como cirurgias. Daí a atenção que se precisa ter ao perfil de especialidades e complexidade do hospital. Ademais, mesmo um procedimento cirúrgico idêntico pode ter níveis de complexidade muito distintos dependendo do paciente. Assim, uma cirurgia de hernia em um jovem sem outros problemas de saúde é muito diferente de uma hernia que complicou em um idoso cheio de comorbidades. Nesse sentido, hospitais mais complexos podem aparecer como menos eficientes se forem comparados com outros menos complexos, que somente façam procedimentos simples em pacientes sem complicações. Quando se tem um sistema de acompanhamento de indicadores hospitalares, os hospitais podem resistir a aceitar paciente complicados para não terem queda em seus desempenhos (o famoso *cream-skimming*). Aliás, essa disparidade na gravidade dos pacientes está na origem da introdução do *Diagnosis-Related Groups* (DRG) para remuneração de procedimentos hospitalares nos Estados Unidos.

Antes de enveredarmos pelas soluções complexas dos casos multivariados, convém especificar ainda mais o conceito de eficiência. De fato, esse conceito não é único. Podemos falar de vários tipos de eficiência. A compreensão de cada um deles é crucial para que possamos compreender a utilidade, as limitações de cada um deles e os métodos disponíveis para calculá-los. A seguir, veremos apenas os conceitos mais básicos[VIII]. Antes, contudo, é necessário definir previamente dois outros conceitos fundamentais em microeconomia: isoquanta e isocusto:

1. Isoquanta: pode ser definida como o conjunto dos pontos que representam combinações de insumos que produzem o 'mesmo nível de produção'. A isoquanta é uma referência para a unidade produtiva (que pode ser um hospital ou posto de saúde), porque indica quais a combinações de cestas de insumos que produzem um determinado nível fixo de produção. Então, se a firma, para produzir um determinado nível de produto, usar mais insumos do que está indicado pela isoquanta da qual esse nível de produto faz parte, ele não é 'tecnicamente' eficiente, conforme a definição de eficiência técnica. O conjunto das isoquantas se chama mapa de isoquantas. Se supomos que a produção aumenta com o aumento dos insumos, é necessário, para manter fixo o nível de produção, que a quantidade de um insumo aumente, quando a quantidade de outro diminui e vice-versa. Se ambas as quantidades aumentam, a produção aumenta de modo recíproco. Então, a inclinação da isoquanta será negativa. Isoquantas mais altas representam maiores níveis de produção. Quanto mais para cima em um dito 'mapa' de isoquantas, maiores as quantidades empregadas dos insumos e maior o nível de produção.

VIII Referências para mais tipos de eficiência são Fried, Lovell e Schmidt[14] e Bogetoft e Otto[15].

Figura 1. Mapa de Isoquantas: o nível de produção permanece constante ao longo de uma isoquanta, e aumenta conforme as isoquantas se afastam da origem dos eixos cartesianos



As isoquantas são essenciais para compreensão dos conceitos de eficiência, pois elas permitem medir o afastamento, ou a 'distância', entre os valores observados de produção e de consumo de uma unidade produtiva, e os pontos de referência dados pela isoquanta. Então, é importante apresentar os conceitos de função distância mais utilizados nesse campo. Trataremos a Distância de Shephard e a Distância de Debreu-Farrell, em suas versões com orientação para insumos e com orientação para produto. As orientações se confundem com as perspectivas de análise: na orientação para produto, calculamos as possibilidades de expansão de produto com insumos fixos; e, inversamente, na orientação para insumos, calculamos quanto podemos poupar de insumos, dado um nível fixo de produção.

Modelo com orientação para insumos

Um conceito básico inicial é a tecnologia que, grosso modo, define o que é possível fazer com os insumos e o conhecimento disponíveis. A tecnologia pode ser descrita pela ótica dos insumos e pela ótica dos produtos. Pela ótica dos insumos, a tecnologia é um conjunto de insumos $L(y) = \{x: (y, x)$ é factível$\}$.

Em uma isoquanta $IsoqL(y) = \{x: x \in L(y), \lambda x \notin L(y), \lambda \in [0, 1)\}$, não é possível produzir o vetor original y com um vetor x radialmente menor do que o original. Radialmente significa na proporção $\lambda \in [0, 1)$, ou seja, reduzir todos os elementos do vetor de insumos na mesma proporção λ. Por exemplo, se $\lambda = 0{,}5$, todos os insumos terão suas quantidades reduzidas pela metade. Um subconjunto eficiente será $EFL(y) = \{x: x \in L(y), x' \notin L(y), x' \leq x\}$. Um subconjunto x de vetores eficientes não

contém nenhum vetor menor do que x. Então, x é o menor vetor de insumos (com menores quantidades de cada insumo) que viabiliza produzir y. Se existisse algum vetor menor viável, o vetor x não seria eficiente.

$x \geq x'$ significa que $x_i \geq x_i'$ para todo i = 1, ... l, mas $x \neq x'$ (é maior em algum i). Por exemplo, $x = \{1, 1, 1, 2\} > x' = \{1, 1, 1,1\}$.

Então, $EFL(y) \subseteq IsoqL(y)$. O conjunto eficiente está contido na isoquanta. Entretanto, nem todos os pontos da isoquanta serão eficientes. Isso depende da definição de eficiência, como veremos.

Distância de Shephard para insumos: $DI(y, x) = máx \{\lambda: x/\lambda \in L(y)\}$. $DI(y, x) \geq 1$.

Exatamente sobre a isoquanta, vemos que $IsoqL(y) = \{x: DI(y, x)=1\}$.

Exemplo: se $\lambda = 2$, poderíamos dividir todos os insumos por 2, e ainda produzir o vetor y original.

Distância de Debreu-Farrell para insumos: $DFI(y, x) = \min \{\lambda: \lambda x \in L(y)\}$. $DFI(y, x) \leq 1$.

Exemplo: se $\lambda = 0,5$, poderíamos multiplicar todos os insumos por 0,5 e ainda produzir o vetor y original.

$DI(y, x) = 1/DFI(y, x)$.

Vemos que $IsoqL(y) = \{x: DFI(y, x)=1\}$.

Modelo com orientação para produtos

Pela ótica dos produtos, a tecnologia é um conjunto de produtos $P(x) = \{y: (x, y)$ é factível$\}$.

Em uma isoquanta $IsoqP(x) = \{y: y \in P(x), \theta y \notin P(x), \theta \in (1, +\infty)\}$, não é possível produzir mais do que o vetor de produtos y original utilizando o vetor de insumos x original.

Um subconjunto eficiente será $EFP(x) = \{y: y \in P(x), y' \notin P(x), y' \geq y\}$. Se y é um subconjunto de vetores de produção eficientes, nenhum deles é maior do que o vetor y original, que é o maior factível. Se algum vetor maior fosse factível, os demais não seriam eficientes.

Então, $EFP(x) \subseteq IsoqP(x)$. O conjunto eficiente está contido na isoquanta. Contudo, nem todos os pontos da isoquanta serão eficientes. Isso depende da definição de eficiência, como será exposto mais adiante.

Distância de Shephard para produtos: $DO(x, y) = \min\{\theta: y/\theta \in P(x)\}$. $DO(x, y) \leq 1$.

A distância θ é o menor valor entre zero e a unidade pelo qual podemos dividir o vetor y, dado o vetor x original.

Exemplo: se $\theta = 0{,}5$ poderíamos dividir y por 0,5 ou multiplicar por 2.

Exatamente sobre a isoquanta, vemos que $IsoqP(x) = \{y: DO(x, y) = 1\}$.

Distância de Debreu-Farrell para produtos: $DFO(x, y) = \max\{\theta: \theta y \in P(x)\}$. $DFO(x, y) \geq 1$.

A distância θ é o maior valor pelo qual podemos multiplicar o vetor y, dado o vetor x original.

Exemplo: se $\theta = 2$, poderíamos multiplicar a produção por 2.

$DO(x, y) = 1/DFO(x, y)$.

Vemos que $IsoqP(x) = \{y: DFO(x, y)=1\}$.

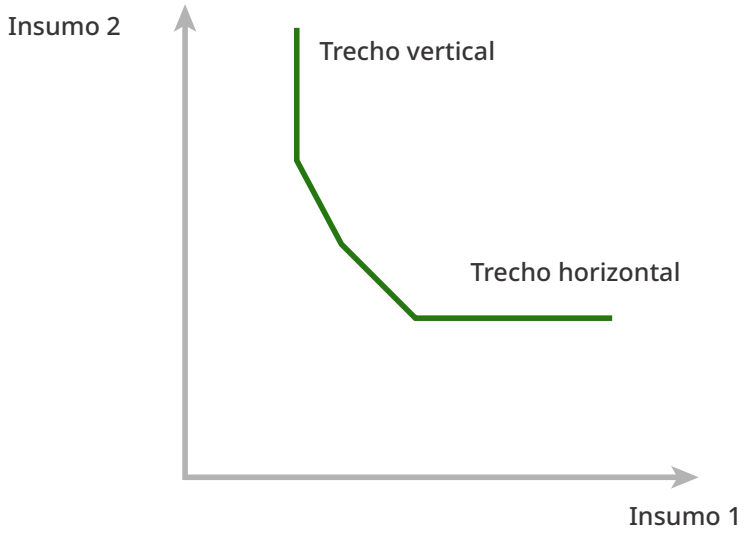
$DFI(y, x) = 1/DFO(x, y)$ se, e somente se, a tecnologia apresenta retornos constantes de escala.

Entendidas as isoquantas, podemos passar aos detalhes dos diferentes conceitos de eficiência propriamente ditos:

1. Eficiência Técnica (ET): suponha que um hospital produz uma determinada quantidade de produtos, que estaria sobre uma dada isoquanta. Vimos que a isoquanta representa um determinado nível de produção fixo. Contudo, vamos supor que o referido hospital não é tecnicamente eficiente, e usa uma quantidade maior de insumos do que seria necessário, para gerar aquele nível de produção. A ET mede o afastamento (distância) relativa entre as quantidades de insumos realmente utilizados, e as quantidades necessárias para estar exatamente na referida isoquanta. Como é uma medida (razão matemática) da relação entre uma posição observada e uma posição ótima (a isoquanta), a ET é um número puro (adimensional). Essa distância está relacionada apenas com quantidades dos insumos e não tem relação com os preços dos insumos ou dos produtos. Um ponto importante, e frequentemente ignorado, com sérias implicações, advém da geometria das isoquantas. Nos eventuais trechos horizontais ou verticais das isoquantas, podemos reduzir um dos insumos sem sair da isoquanta, ou seja, sem reduzir o nível de produção. Vimos que $EFP(x) \subseteq IsoqP(x)$, ou seja, conjunto eficiente está contido na isoquanta, mas nem todos os pontos da isoquanta serão eficientes. Diz-se que existem folgas, ou

disponibilidade forte de insumos (*strong disposability of inputs*). Em tais trechos, como existem folgas, existe Eficiência Fraca de Pareto (EPf), e o desempenho de algumas firmas (mas não todas) poderia melhorar. Diz-se que uma alocação de insumos e de produtos entre agentes econômicos é EPf se não é possível melhorar 'todos' os agentes sem piorar algum outro, mas ainda é possível melhorar pelo menos um agente. Por outro lado, diz-se que uma alocação de insumos e de produtos entre agentes econômicos apresenta Eficiência de Pareto Forte (EPF), se não é possível melhorar 'nenhum' dos agentes sem piorar algum outro. Nos trechos verticais e horizontais das isoquantas, poderíamos retirar certas quantidades de insumos (médicos, por exemplo) de um hospital e passar para algum outro hospital que, eventualmente, aumentaria a sua produção. Contudo, o hospital que perdeu médicos não reduziria a sua produção, daí a EPf. Nem sempre as isoquantas têm trechos horizontais ou verticais. Isso costuma acontecer em processos produtivos em que há proporções fixas entre os fatores de produção. Por exemplo, para cada ambulância, precisamos de somente um motorista; e caso haja duas ambulâncias, uma poderá ser dispensada, e alocada em outro hospital que eventualmente tenha um motorista desocupado. Assim, o hospital que perdeu uma ambulância não reduz a sua capacidade produtiva, e o hospital que ganhou uma ambulância aumenta a sua capacidade produtiva *ceteris paribus*. Resumidamente: uma unidade produtiva pode operar exatamente em cima de uma isoquanta e, ainda assim, não ser eficiente por um critério mais rígido (EPF).

Figura 2. Isoquanta com trechos paralelos aos eixos de coordenadas. Nesses trechos é possível reduzir a quantidade do insumo sem reduzir o nível de produção



2. ET de PK: retornamos a esse conceito, porque ele é muito importante. Vimos que uma unidade produtiva é eficiente no sentido de PK se um aumento em qualquer produto requer a redução da quantidade de pelo menos um outro produto, ou o aumento da quantidade de pelo menos um insumo[IX]; e se a redução de qualquer insumo requer um aumento da quantidade de pelo menos outro insumo, ou a redução da quantidade de pelo menos um produto. Um produtor 'tecnicamente ineficiente' poderia produzir as mesmas quantidades de todos os produtos, utilizando menor quantidade de pelo menos um insumo, ou utilizar as mesmas quantidades de todos os insumos, para produzir mais de pelo menos um produto. Seja o exemplo de um hospital que usa uma enfermeira e um médico para produzir uma consulta e um exame por hora; se ele for eficiente em PK, somente poderá aumentar a quantidade de consultas se reduzir a quantidade de exames, ou se aumentar a quantidade de médicos ou de enfermeiras. Isso vale para aumentar a quantidade de exames, que somente poderá ser feita à custa da redução da quantidade de consultas ou do aumento das quantidades de médicos ou de enfermeiras. Se o hospital pode aumentar as quantidades de consultas, ou de exames, sem reduzir o outro produto, nem aumentar o uso de nenhum insumo, e não o fez, é porque é ineficiente. Por outro lado, se for eficiente, somente poderá reduzir a quantidade de médicos se aumentar o uso de enfermeiras, ou se fizer menos exames, ou fizer menos consultas. Raciocínio análogo vale para a redução de quantidade de enfermeiras, que somente poderá ser obtida com a redução de algum produto ou aumento do uso de médicos. Então, resumidamente, fixada uma combinação de insumos e de produtos, uma unidade produtiva eficiente pelo critério PK não tem espaço para economizar qualquer insumo nem para expandir a produção de qualquer produto. Uma unidade eficiente faz o melhor possível.

3. Isocusto: é o conjunto das diferentes combinações de insumos que apresentam o mesmo custo monetário total, dados os preços fixos destes. No caso de dois insumos, a isocusto é uma reta com inclinação negativa, porque, se aumentarmos a quantidade de um insumo, temos de reduzir a quantidade do outro para manter o custo constante. Se aumentarmos as quantidades de ambos os insumos, o custo total aumenta, e se reduzirmos as quantidades dos dois insumos, o custo total diminui. A isocusto é uma referência para o produtor, porque indica quais a combinações de cestas de insumos que custam um determinado valor. Então, pagar mais por uma cesta de insumos do que está indicado pela isocusto

IX É essencial a suposição de que os insumos possam ser substituídos uns pelos outros em alguma medida. Isso vale para os produtos. No mundo real, nem sempre isso é possível.

da qual ela faz parte não é 'alocativamente' eficiente, conforme a definição de eficiência alocativa, dada a seguir.

$C=w_1x_1+w_2x_2+...+w_kx_k$. Em que:

C é o custo suportado pelo orçamento do produtor; os *ws* são os preços pagos (custos) por unidade de cada insumo *x*.

4. Eficiência de Custos (EC): É a razão matemática entre o custo mínimo possível para gerar um dado nível de produção e o custo efetivamente observado. Esse conceito não deve ser confundido com a isoscusto. Esta localiza cestas de insumos com o mesmo custo, independentemente do nível (vetor) de produção que elas possam gerar. O custo mínimo está relacionado com um dado nível de produção. Nenhuma cesta mais barata do que a cesta de custo mínimo (cesta ótima) poderá produzir aquele dado vetor de produtos.

   A função custo mostra o custo mínimo para produzir a combinação de produtos *y* aos preços de insumos *w* (supostos iguais para todas as Unidades Tomadoras de Decisões (*Decision Making Units* – DMU), dada uma tecnologia T:

   $C(w, y)=\min x\{wx \mid (x, y) \in T\}$. Então, o custo observado é maior ou igual ao custo mínimo: $wx \geq c\ (w, y)$ para todo $(x, y) \in T$.

   A EC será $EC=C(w, y)/wx \leq 1$ para todo $(x, y) \in T$.

5. Eficiência Alocativa (EA): suponha que um hospital produz uma determinada quantidade de produtos que estaria sobre uma dada isocusto, mas ele usa uma combinação de insumos que, dados os seus preços, custam mais caro do que o custo (constante) representado naquela isocusto. A EA mede a distância entre os custos observados (reais) e os custos (constantes) na isocusto capazes de gerar um dado nível de produção. Como a isocusto tem sua inclinação determinada pela razão entre os preços dos insumos, a EA avalia a relação entre essa razão de preços (preços relativos) e a proporção entre os insumos escolhida pelo hospital. Dito de outro modo, a EA mede o afastamento entre a isoquanta e a isocusto do hospital. Se a isoquanta e a isocusto são tangentes no ponto em que o hospital está produzindo – o que implica que elas têm a mesma inclinação –, o hospital será dito alocativamente eficiente. A EA é uma distância relativa entre uma dada isoquanta e a isocusto capaz de produzir o nível de produção dessa isoquanta e, portanto, é um número puro, normalmente no intervalo [0, 1] ou uma percentagem no intervalo [0%, 100%].

6. A Eficiência Total (ET) e a sua composição:

   Seja x^* a alocação de *inputs* que minimiza custos para uma produção fixa y; e x a alocação de *inputs* escolhida.

   A EC, ou eficiência econômica, será $EC=wx^*/wx$.

   Então o Problema Básico de Custos (PBC) será:

   Minimizar o custo wx sujeito a $(x, y)\epsilon T$.

   Seja x' a cesta tecnicamente eficiente e x a cesta escolhida. Então, $ET=x'/x$, em que ET é a eficiência técnica[X].

   Como w é positivo, podemos escrever a ET como: $ET=wx'/wx$.

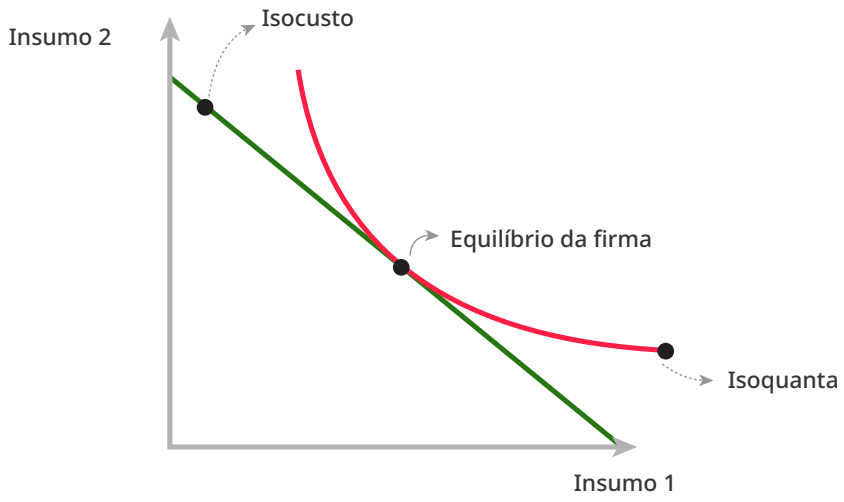
   Seja x^* a cesta mais barata para produzir y. A EA será $EA=wx^*/wx'$. Ela é a razão entre os custos das cestas minimizadoras de custos (estão na isocusto) e os custos das cestas tecnicamente eficientes (estão na isoquanta).

   A EC será: $EC=wx^*/wx=(wx^*/wx').(wx'/wx)$=EA.ET.

   A ET é o produto da ET pela EA. Assim, $EC=ET$xEA. Então, a ET será igual a unidade ou 100% se, e somente, se $EC=EA=ET=1$. Nesse caso, o hospital, produzindo um nível de produto localizado em uma isoquanta, tem os seus *insumos* em níveis exatamente necessários para produzir esse nível de produto. Além disso, a sua isocusto tangencia a isoquanta exatamente nesse ponto, de modo que a razão de entre as quantidades dos insumos é igual a razão entre os preços desses *insumos* (preços relativos).

X Observação: se, no modelo de Farrell, usarmos custos como insumos, a Eficiência de Custos CE=ET (Eficiência Técnica).

Figura 3. Isocusto, insumo e isoquanta são tangentes. É o ponto ótimo da firma (equilíbrio)



7. A Eficiência de Escala (EE): no curto prazo, admite-se que alguns insumos sejam fixos, e não podem ser ajustados, aceitando-se alguns desajustes em relação ao longo prazo, quando todos os insumos são variáveis. No longo prazo, a escala de operação é ótima por um período considerável. Por exemplo, um hospital não pode modificar a sua área construída rapidamente, mas pode modificá-la no longo prazo se isso for julgado conveniente. Então, dado um nível de produção do hospital, podemos construir uma isoquanta, supondo que todos os insumos são variáveis, e uma outra supondo que pelo menos um insumo é fixo. Se não consideramos o insumo fixo na análise, estamos construindo uma isoquanta de curto prazo, e a eventual distância entre o ponto em que o hospital opera e essa isoquanta será menor, do que se consideramos que tudo já poderia ter sido ajustado e que a referência válida é uma isoquanta de longo prazo – que estará, por sua vez, mais afastada do ponto determinado pela cesta de insumos usada pelo hospital. A distância entre relativa entre essas isoquantas é a EE. Se a isoquanta de curto prazo coincide com a de curto prazo, no ponto em que o hospital produz, ele será dito eficiente em escala por estar funcionando com tamanho ótimo (quantidades ótimas) de todos os insumos e da produção. Essa medida não considera preços de insumos nem de produtos (embora o hospital possa ter se equivocado ao escolher o seu tamanho por causa de algum preço). É também uma medida no intervalo [0, 1] ou [0%, 100%].

   Um ponto interessante é que essa medida, *per se*, não informa se um hospital tem ineficiência de escala por ser muito grande ou por ser muito pequeno. Será preciso

recorrer a modelos quantitativos sofisticados para aferir empiricamente o que está ocorrendo. A natureza dos retornos de escala é fundamental em saúde. Por exemplo, uma porcentagem grande dos hospitais do SUS tem menos de 50 leitos, de modo que estão abaixo da escala ótima observada em hospitais, que estaria em torno de 100-200 leitos, dependendo muito do tipo de hospital (hospitais universitários são necessariamente grandes, pois têm de oferecer aprendizado em múltiplas especialidades)[XI]. Não há como tornar eficientes hospitais que sejam muito pequenos, nem excessivamente grandes.

O problema da escala também surge na grande quantidade de municípios brasileiros muito pequenos, levando às necessidades de regionalização, ou de formação de consórcios, para que os serviços tenham uma escala de operação aceitável. Suponha que a fronteira de eficiência seja uma reta, que seria um caso muito simples, para efeitos de ilustração. Nessa situação, podemos, no caso de um insumo (x) e um produto (y), representar a fronteira por uma equação da seguinte reta $y=ax+b$. A constante (parâmetro) a é uma constante positiva, também chamada de declividade ou coeficiente angular da reta. A declividade será positiva, pois supomos que y aumenta quando x aumenta e vice-versa (monotonicidade). O parâmetro b é um número real positivo, negativo, ou nulo, e é chamado de intercepto ou coeficiente linear. Por exemplo: Se a=2, e b=1, para x=3 teremos y=2*3+1=7. O valor do parâmetro b determina a natureza dos retornos de escala da fronteira (e da tecnologia subjacente). Vejamos, brevemente, as principais possibilidades que existem para os retornos de escala:

7.1 Retornos constantes de escala (*Constant Returns to Scale* – CRS): ocorrem se, e somente se, b=0. Então, teremos $y=ax$. Nesse caso, o gráfico será uma reta passado pela origem dos eixos coordenados (0, 0). Exemplo: $y=2x$. Se x=0, temos y=0 e o gráfico passa pela origem dos eixos (0, 0). Nesse caso, se x=1, temos y=2. Se x=2, temos y=4. Ou seja, se dobramos a quantidade do insumo, a quantidade do produto também dobra por consequência. Então, dizemos que os retornos de escala são constantes (exatamente proporcionais).

7.2 Retornos variáveis de escala: pode ocorrer que uma fronteira apresente apenas um tipo de retorno de escala, mas também podem ocorrer dois ou três tipos de retornos de escala em trechos diferentes da fronteira. Podem ocorrer retornos constantes ou crescentes para níveis baixos de produção, e retornos decrescentes para níveis altos de produção.

XI Aqui, trata-se em particular de 'economias de escopo'.

7.3 Retornos crescentes de escala (*Increasing Returns to Scale* – IRS): ocorrem se, e somente se, $b<0$. Exemplo: $y=2x-1$. Se $x=0$, temos $y=-1$. A reta não passa pela origem e interceptaria o eixo vertical em um ponto em que y é negativo. Nesse caso, se $x=1$, temos $y=1$; e se $x=2$, temos $y=3$. Dobramos o valor de x, e o valor de y triplicou. Os retornos são mais do que proporcionais (são crescentes).

7.4 Retornos decrescentes de escala de escala (*Decreasing Returns to Scale* – DRS): ocorrem se, e somente se, $b>0$. Exemplo: $y=2x+1$. Se $x=0$, temos $y=1$. A reta não passa pela origem e interceptaria o eixo vertical em um ponto em que y é positivo. Notem que esse ponto é apenas uma referência teórica não observável na prática, pois não seria possível produzir a partir de quantidade nula de insumo ($x=0$). Nesse caso, se $x=1$, temos $y= 3$; e se $x=2$, temos $y=5$. Dobramos o valor de X, e o valor de Y menos do que duplicou. Os retornos de escala são menos do que proporcionais (são decrescentes).

No caso de produção múltipla, temos de multiplicar (ou dividir) todos os insumos por uma mesma constante positiva e ver o que acontece com todos os produtos. Nesse caso, ao invés de uma equação da reta, em que vemos o intercepto (b) da reta, temos de ver o que acontece com o intercepto do dito 'hiperplano', que é o sucedâneo da reta, no caso multivariado. A fronteira terá retornos constantes de escalas se, e somente se, o intercepto for nulo.

8. Indicadores de Performance (*Performance Indicators* – PI): embora alguns indicadores sejam medidas simples de eficiência, utilizados tradicionalmente, eles não representam uma avaliação de desempenho em saúde, não são, necessariamente, indicadores de eficiência nem de produtividade. Usualmente, os PI envolvem alguma razão matemática do tipo enfermeiros/leitos; consultas /médicos; médicos/habitantes etc. Note que enfermeiros e leitos são, ambos, insumos de processos produtivos em saúde. Consultas, por sua parte, não são bens finais em um sentido estrito, conforme já discutimos, e médicos costumam ser insumos. Habitantes, por seu turno, também não costumam ser bens finais, embora indicadores importantes de cobertura possam usar essa variável em diferentes recortes.

Nem sempre os PI têm algum significado econômico claro. Entretanto, eles podem, eventualmente, fornecer alguma noção restrita sobre a razoabilidade relativa de combinação de insumos ou produtos em uma amostra. Nesse caos, eles revelam valores muito discrepantes de uma média considerada razoável por algum critério administrativo, histórico etc. (exemplos: quantidades de enfermeiros/quantidades de

médicos; quantidade de médicos por habitantes). Tais indicadores têm algumas vantagens e desvantagens. Descrevemos alguma delas a seguir. Vantagens: são medidas simples, práticas e eventualmente tradicionais. Desvantagens: podem não ter conexão com a teoria econômica, podem misturar livremente insumos com insumos; ou produtos com produtos; ou variáveis endógenas (têm os valores determinados no modelo: por exemplo, a infecção hospitalar em hospitais) com exógenas (têm seus valores determinados fora do modelo: por exemplo, a escolaridade dos pacientes dos hospitais). Um problema adicional ocorre quando resultados parciais são conflitantes. Por exemplo: um hospital tem 'melhor' quociente de quantidade de enfermeiros/ quantidade de leitos (suponha – apenas para simplificar – que menor seja melhor). Então, o hospital deveria reduzir a quantidade de enfermeiros ou aumentar a quantidade de eleitos, ou ambas as coisas. Outro hospital tem 'melhor' quociente do tipo médicos/enfermeiros (suponha – também para simplificar – que menor seja melhor). Agora, o hospital deveria aumentar a quantidade de enfermeiros, ou reduzir a quantidade de médicos, ou ambas as coisas. A primeira medida recomenda reduzir a quantidade de enfermeiros, enquanto a segunda medida sugere aumentar a quantidade de enfermeiros, o que gera um impasse. Além disso, como vamos dizer qual o 'melhor' hospital (se for esse o desejo) dado que cada um deles supera o outro em um critério e é superado no outro critério? Outro óbice na utilização dos PI é que eles admitem, implicitamente, a presença de retornos constantes de escala na amostra. Por exemplo: *ceteris paribus*, um hospital que possui apenas dez leitos, e faz dez cirurgias em um mês (uma cirurgia por leito), terá o mesmo escore que um hospital que possui cem leitos e faz cem cirurgias (uma cirurgia por leito) no mesmo período. Entretanto, um hospital com apenas dez leitos é, claramente, pequeno demais para ser economicamente eficiente. Isoladamente, o indicador é incapaz a de avaliar a adequabilidade da escala de operação (tamanho do hospital).

Neste capítulo, de leitura que reconhecemos algo árida para não economistas, apresentamos um conjunto de conceitos econômicos fundamentais para a compreensão precisa do desenvolvimento e das possibilidades de aplicação do conceito de eficiência econômica, de seus desdobramentos (ET, EA, EE) e de conceitos correlatos, como produtividade, custos, efetividade, eficácia, dominância, entre outros. O leitor cujo interesse resida apenas na compreensão do conceito de eficiência econômica e de sua importância na formulação e na avaliação de políticas poderia abdicar da completa apreensão do que está exposto neste capítulo. Ainda assim, a nosso ver, eventuais candidatos a avaliadores de eficiência deveriam se familiarizar bastante com o conteúdo aqui exposto neste livro, ou alhures, sob risco de cair em armadilhas do senso comum, ou de ser cooptado por interesses políticos ou econômicos não explicitados, mas que possuem fortes apoiadores e difusores no Brasil e no exterior.

Referências

1. Varian HR. Microeconomic Analysis. 3rd ed. New York: W.W. Norton & Company; 1992.
2. Arrow K. Uncertainty and the Welfare Economics of Medical Care. Am Econ Rev. 1963;53(5):941-73.
3. Ocké-Reis CO. O mercado de planos de saúde no Brasil: uma criação do estado? Rev Econ Contemp. 2006;10(1):157-85.
4. Bahia (2009)
5. Andrade (2012)
6. Marinho A. A crise do mercado de planos de saúde: devemos apostar nos planos populares ou no SUS? Planej Polít Públicas. 2017;49:-55-84.
7. Brasil. Constituição (1988). Constituição da República Federativa do Brasil. Brasília, DF: Senado Federal; 1988.
8. Silveira DS, Santos IS. Fatores associados à cesariana entre mulheres de baixa renda em Pelotas, Rio Grande do Sul, Brasil. Cad Saúde Pública. 2004;20(suppl 2):s231-s241.
9. Rocha NFF, Ferreira J. A escolha da via de parto e a autonomia das mulheres no Brasil: uma revisão integrativa. Saúde Debate. 2020;44(125):556-68.
10. Medici AC. Financiamento e contenção de custos nas políticas de saúde: tendências atuais e perspectivas futuras. Planej Polít Públicas. 1990;4:83-98.
11. Baumol WJ. The cost disease. Why computers get cheaper and health care doesn't. New Haven: Yale University Press; 2012.
12. Butler JRG. Hospital Cost Analysis. Dordrecht, Netherlands Boston: Kluwer Academic Publishers; 1995.
13. Koopmans T. Activity analysis of production and allocation. New York: John Wiley & Sons, New York; 1951.
14. Fried HO, Lovell SS, Schmidt CAK. The Measurement of Productive Efficiency. Techniques and Applications. Oxford: Oxford University Press; 1993.
15. Bogetoft P, Otto L. Benchmarking with DEA, SFA and R. New York: Springer; 2011.