|  |
| --- |
| IU |
| Health Economics |
| DLMIHMHE01 |

# 

# Unit 1 – The Peculiar Market for Healthcare

**Study Goals**

On completion of this unit, you will be able to …

… recognize the need for health and the demand for healthcare.

… analyze the contributions of healthcare.

… identify supply resources for healthcare.

… describe typical market failures in the healthcare market.

# 1. The Peculiar Market for Healthcare

## Introduction

The healthcare market has characteristics that differentiate it from a perfect competitive market. With that in mind, this unit will discuss different economic concepts, starting from the production function of health and healthcare. This is followed by an explanation of the concepts of supply and demand, which is developed further in the specific context of the healthcare market. The analysis of these concepts will be done from both individual and societal perspectives.

Then, the nature of market failure will be explored. A variety of international case studies will be used to explain the concept, accompanied by useful figures. This unit emphasizes the way that government-regulated healthcare markets can improve efficiency.

## Demand and Need

The demand for healthcare comes from the need for health, which is a basic need of any individual. In this section, we will discover more about why the need for health exists and how it leads to demand in healthcare. Both individual and societal perspectives will be discussed. The concept of demand in economics will be briefly explained in the context of healthcare in particular.

### Health and Health Needs

Health is defined by the World Health Organization as “a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity” (WHO, n.d., p. 1). This is a wide-ranging definition that involves various characteristics that, combined, achieve the state of healthiness; this is not only limited to healthcare. In Maslow’s (1943) well-known hierarchy of needs, health is fundamental and therefore lies on the basic level of needs that creates a foundation for the higher levels, such as psychological and fulfillment needs.

As an essential need, the production of health at both the individual and societal levels is a central concern for healthcare providers and health economists. They desire meaningful, measurable ways to capture the health status. The methods should also be comparable among individuals and population with the general purpose of indicating when intervention is needed to improve health. Different measures of health have been adapted throughout human history alongside advancements in economies, technologies, living standards, and especially medicines. One empirical indicator that is meaningful at the societal level is the **mortality rate** (or death rate), which has been systematically available since 1841 in England and Wales (Folland et al., 2017). With a long historical development, the mortality data are now highly reliability and accurate. and researchers use it as an important indicator to evaluate, for example, the effectiveness of health-related interventions.

**Mortality rate**

The ratio between number of deaths caused by the health event, scaled to the population size, per unit of time

Although mortality rate is an important concept, other measures of health should also be considered to capture the full picture of health. A more recent measurement concept is morbidity, which is the state of being symptomatic or unhealthy due to a disease. This is also further integrated into a broader, more complicated concept: quality of life. Some efforts have been invested to measure the health-related quality of life (HRQoL), including the development of generic (e.g., EQ5D and SF36) and disease-specific instruments (e.g., EORTC and QLQ-C30; Pequeno et al., 2020).

#### The production of health: Individual perspective

In a production function, outputs are presented in relation to the potential inputs. Two different types of inputs, individual and societal, contribute to the health output at various levels. The former is explored in this section, while the latter is discussed in the next.

At the individual level, the benefits of being healthy are obvious. They include the following:

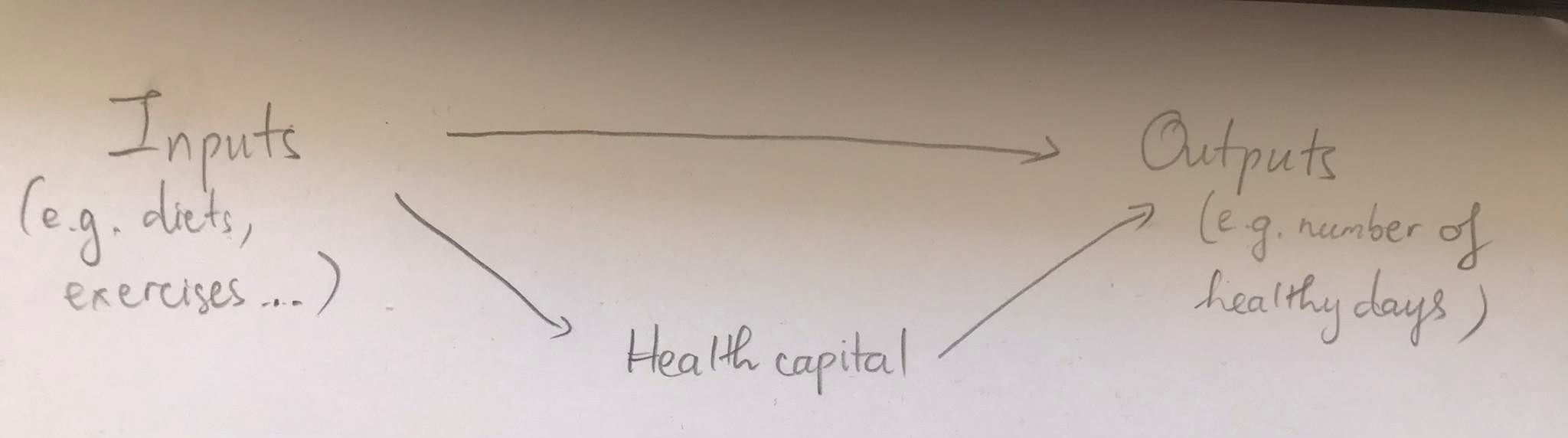
* We feel better when we are healthy.
* We have more time to work when we have less time lost due to illness.
* We work more productively when we are healthy.
* We may live longer by maintaining a healthy lifestyle.

Therefore, it is beneficial for individuals to invest in health. The economic concept of investment in health has been expanded based on the influential work by Michael Grossman (1972), who used human capital theory to explain the contribution of variables such as age, education, health status, and income to individuals’ health. Grossman (1972) describes how the demand in health is different from other demands using the following essential concepts:

* Health is what the consumer wants. Healthcare inputs are demanded to produce it.
* Individuals invest time in health-related activities and purchase medical inputs to improve their health. They cannot buy health directly from the market.
* Health does not depreciate instantly; it lasts for more than one defined time period. Health can be considered as a capital good.
* Health can be regarded as both an investment and a consumption good. People invest in health to increase the number of days they are available to work and to earn extra income. Consuming health also makes people feel better.

Now, we will dive deeper into the economic concepts of health capital investment to discover how and at what level the investment in health stock would help improve an individual’s health status. For better understanding, we will simplify the investment in health capital stock to produce a single output of healthy days, as shown in the figure below. The natural maximum number of healthy days for a year is 365, which is also the maximum output of the yearly investment in health an individual can get.

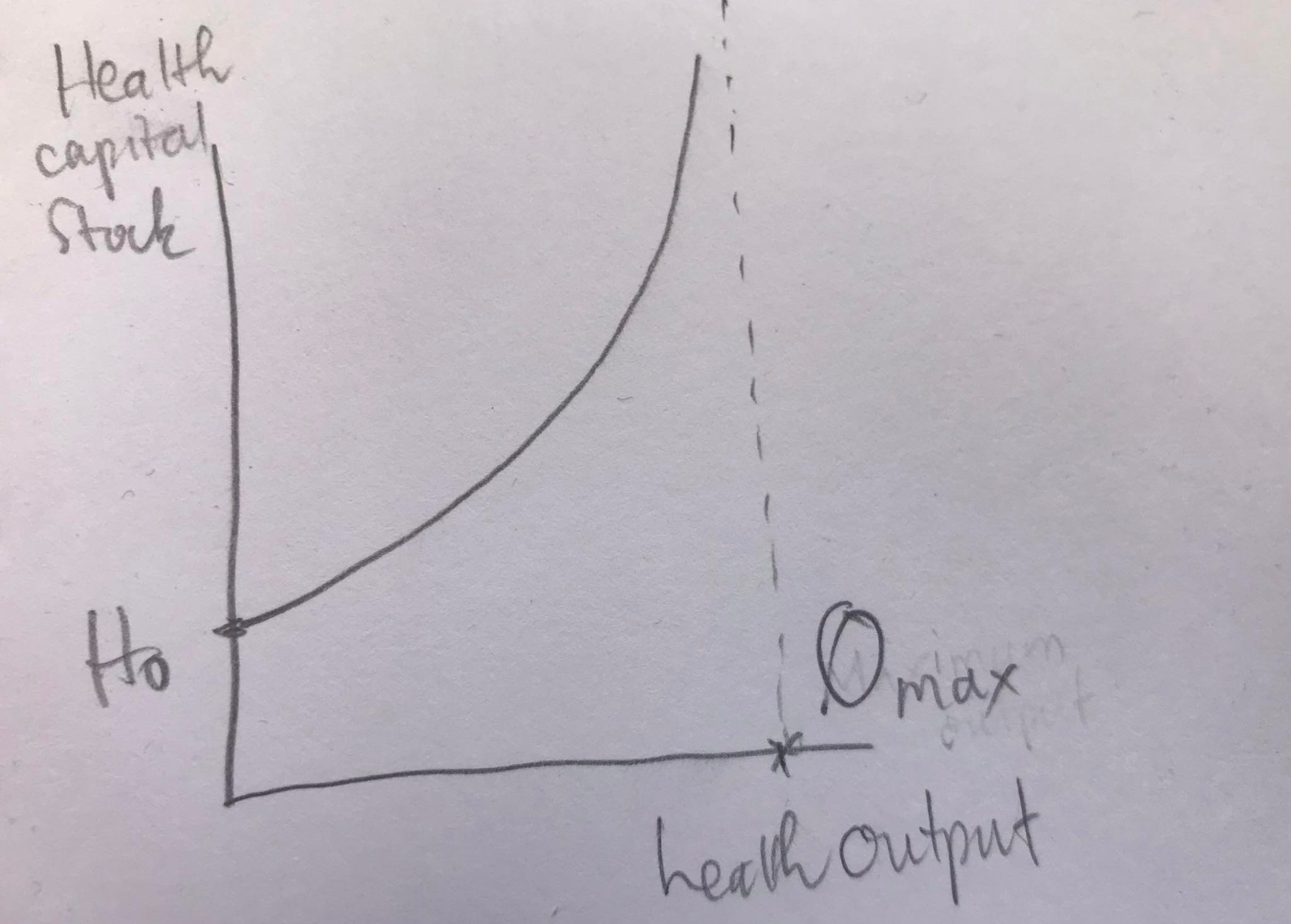
Health Capital Investment



Source: Duy Pham (2022).

The figure below further illustrates the relationship between health capital stock and the outputs. It should be noticed that the bowed shape of the graph indicates that when the amount of inputs increases, the marginal impact on the output decreases. This concept is known as the law of diminishing marginal benefits (Folland et al., 2017). As shown in the figure, represents the minimum health stock where the investment is positive, but zero health output is produced (i.e., zero healthy days). This means that the health output only starts to increase after reaching the minimum amount of investment. In the figure, represents the maximum amount of health outputs an individual can get; for example, the maximum number of healthy days an individual can get for a year is 365 days.

Relationship between Health Capital Stock and Health Output



Source: Duy Pham (2022).

If we call H the health capital stock, an individual gets H by investing time in health-related activities or buying medical care . Since health is a capital good, health capital will be depreciated over time at the depreciation rate . The function of health capital at a certain time can be formulated as:

Assuming time point is today, so yesterday is . This function can be explained by stating that the health capital you get today depends on three things:

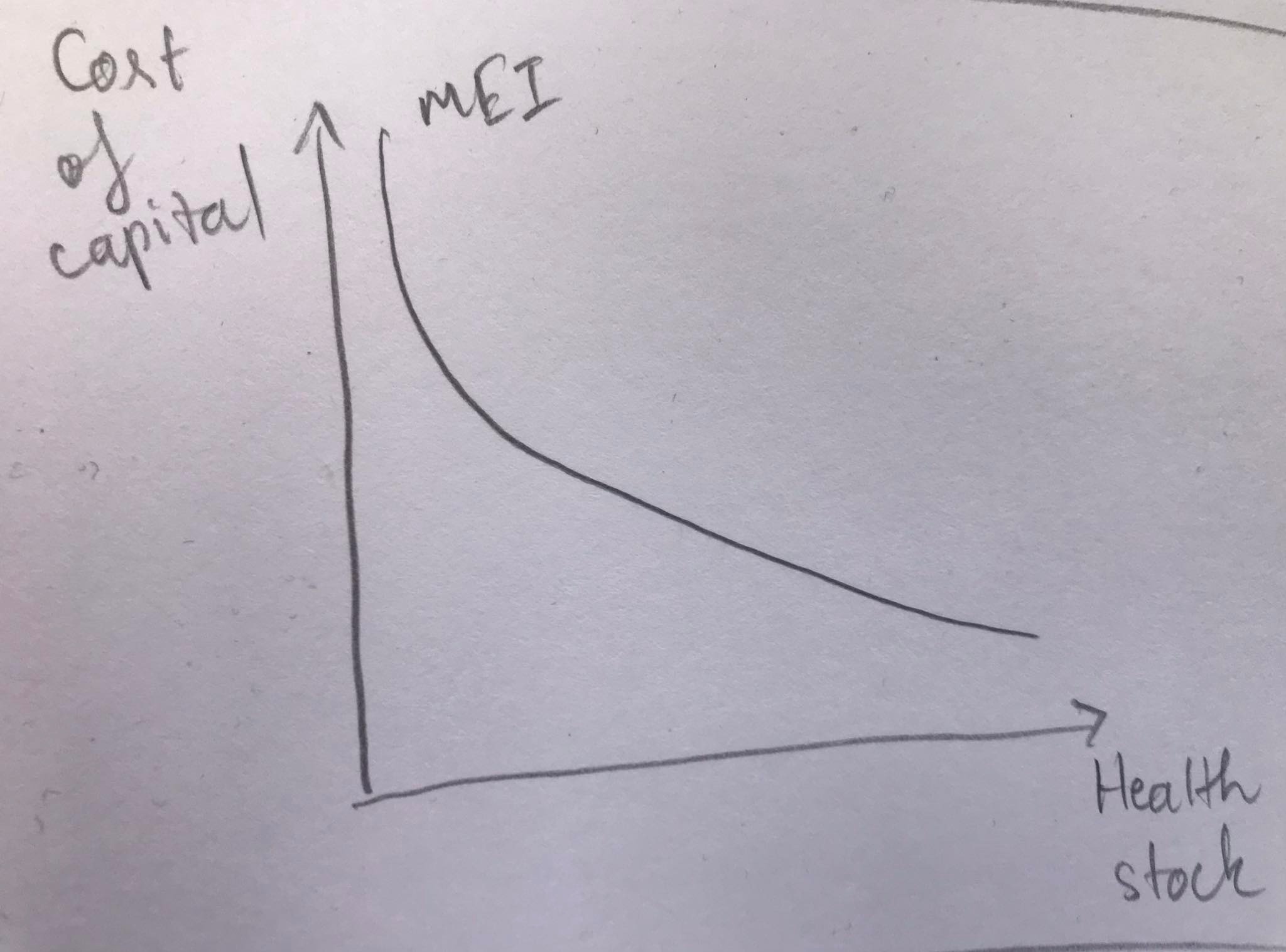
1. Your health capital yesterday (taking into account depreciation)
2. Time invested in health-related activities today
3. Medical care consumed today

**Rate of return**

The rate of return is a percentage of the net value gain or loss over the initial investment cost.

The amount of invested resources corresponds to the **rate of return**, which is calculated as a percentage of the net value gain or loss over the initial investment cost (Folland et al., 2017). The rate of return of investment in health, like any capital good, decreases as the amount of investment in health stock increases, which results in the downward-sloping curve of marginal efficiency of investment (MEI) in relation to the cost of the capital as illustrated in the figure below.

Cost of Capital versus Health Stock



Source: Duy Pham (2022), based on Folland et al. (2017).

Further detail on how age, wage, education, and uncertainty affect the changes in equilibrium (based on Grossman’s model) has been discussed by several other health economists in literature related to the field (Folland et al., 2017).

#### The production of health: Societal perspective

In a societal perspective, population health is the focus. Thus, we need to look at a bigger picture to deduce which inputs contribute to the population health status. Two main components should be considered when dealing with population health: the health outcomes of a pre-defined population and the distribution of these outcomes within that population.

There are many determinants of population health, which Kindig et al. (2008) categorized into five major groups:

1. Health (medical) care
2. Individual behavior
3. Social environment
4. Physical environment
5. Genetics

Medical care does not make a major contribution to determining population health. If we include public health measures in medical care, its contribution to population health becomes greater. Among those determinants, the largest mortality rate decline in history is credited to improvement in the social environment, particularly the greatly increased nutritional supply that resulted from the agricultural and industrial revolutions (Folland et al., 2017). However, the above statement does not mean that the importance of healthcare should be neglected.

### The Contributions of Healthcare

Here, we will look at how healthcare contributes to health. First, we will explore the marginal product of healthcare, which refers to the benefit the health status that comes from the increase in one healthcare unit. Subsequently, the relationship between healthcare expenditure and life expectancy is discussed.

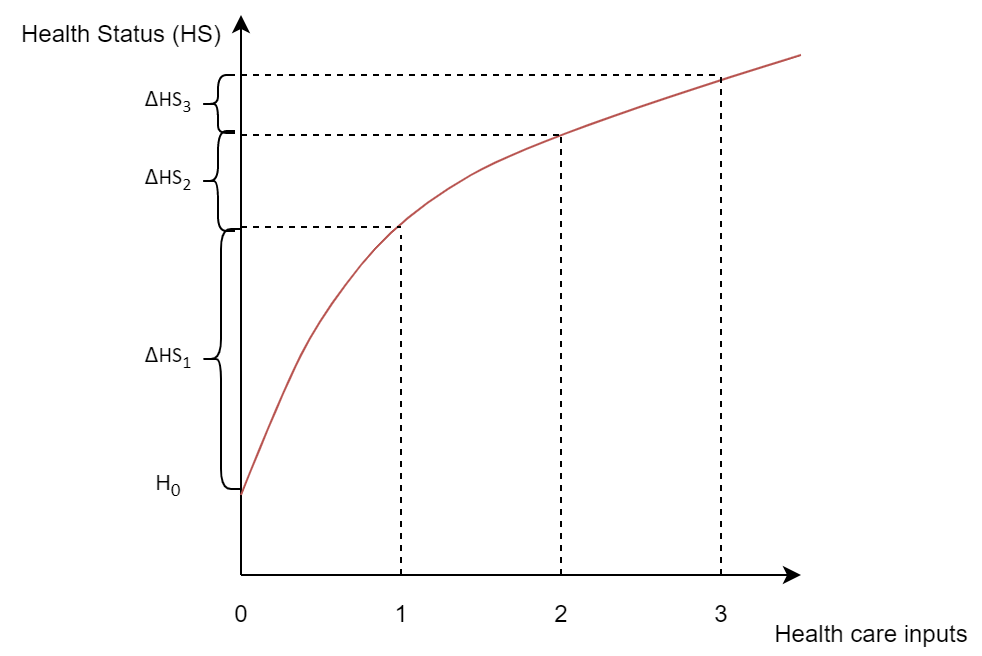
#### The marginal product of healthcare

The figure below illustrates the production function of population health by simplifying the healthcare (HC) inputs to be numerical units on the horizontal axis, while health status (HS) of the population is the output on the vertical axis. In practice, HS can be indicated by, for example, the mortality rates, number of healthy days, or morbidity indicators such as disability days. We can conclude from the figure below that the more HC inputs we invest in, the better the HS. This bowed curve looks very similar to the production of healthy days at the individual level as illustrated previously. The only difference is that, in this figure, the baseline point indicates the health status of the population independent of healthcare interventions.

The marginal product of healthcare inputs is calculated as the increment of HS resulting from one additional unit of HC while other inputs remain constant. In the figure below, ΔHS1 the marginal product of adding the first unit of healthcare input, ΔHS2 the marginal product of adding the second, and ΔHS3 the marginal product of adding the third. The bowed shape of the curve indicates the diminishing value of marginal products when adding an increasing amount of HC inputs, explained by the decreasing values from the ΔHS1 to ΔHS3 (ΔHS1 > ΔHS2 > ΔHS3, etc.).

The principal concept of the marginal product in economics is very useful when setting priorities in healthcare interventions. By evaluating their marginal product on the population health, along with their costs, informed decisions can be made more easily.

Production of Health

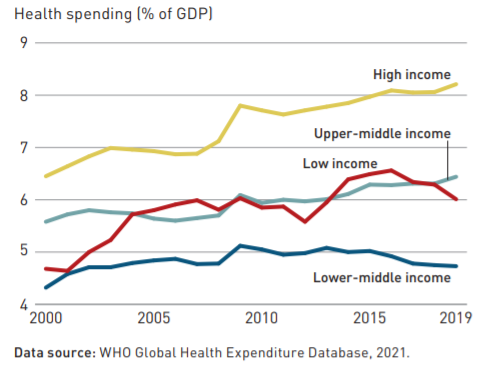


Source: Duy Pham (2022).

#### Healthcare expenditure versus life expectancy

In a recent report by the World Health Organization (WHO, 2021), heath spending in 2019 varied between an average of 4.9 percent of gross domestic product (GDP) in low- and middle-income countries and an average of 8.2 percent in high-income countries (see the figure below). The gap is even bigger when looking at the absolute value of health spending, where high-income countries spend 80 times more than low-income countries (WHO, 2021).

Global Health Expenditure

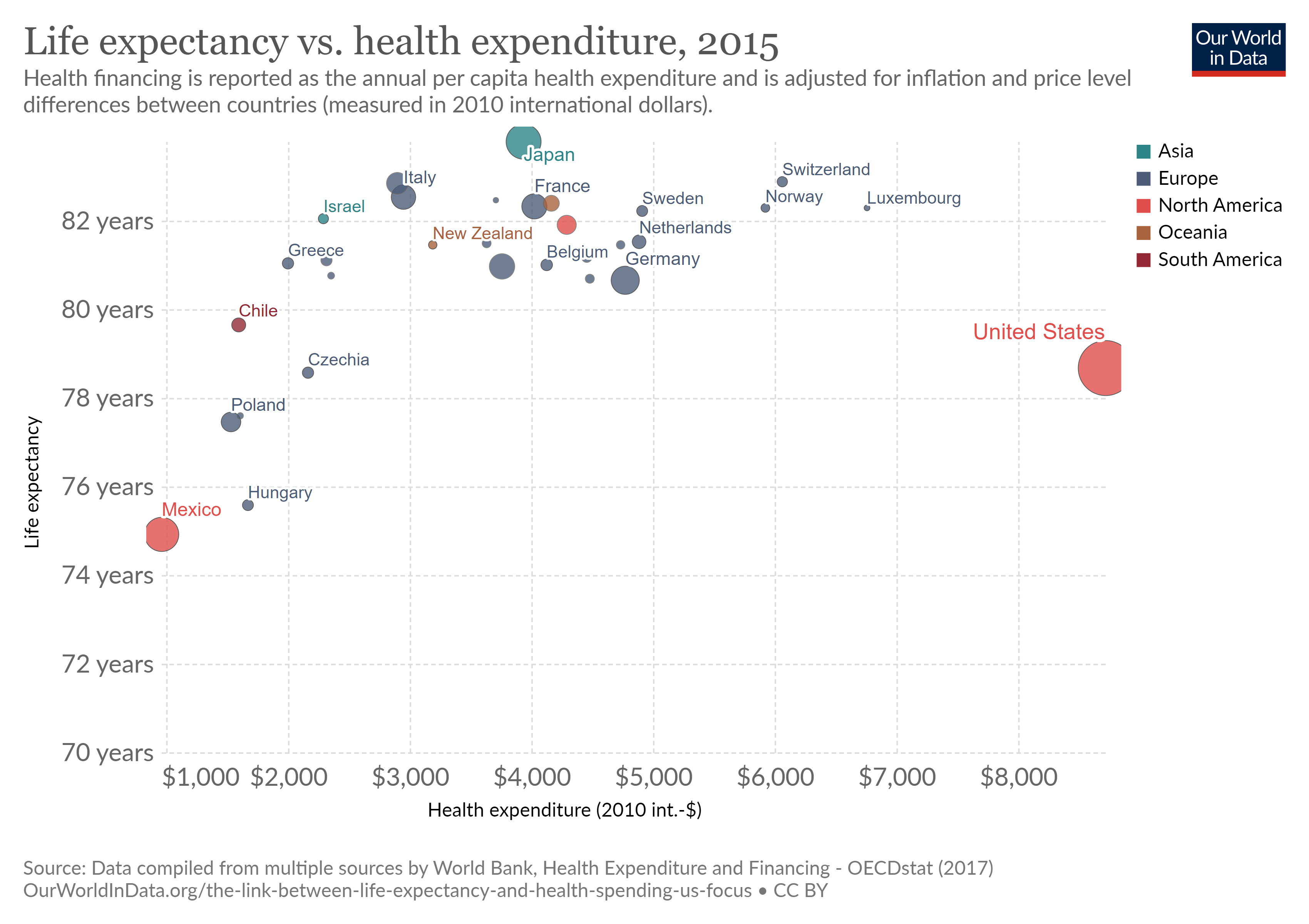


Source: World Health Organization (2021). CC BY-NC-SA 3.0.

The relationship between health expenditure and life expectancy among the Organization for Economic Cooperation and Development (OECD) countries in 2015 is shown in the figure below (Our World in Data, 2017). Generally, a higher amount of health expenditure correlates to higher life expectancy, although there are some outliners, such as the U.S. and Japan. Even though the amount of health expenditure in Japan is lower than in the US, their life expectancy is higher. There are three main explanations:

1. There are many contributors to life expectancy outside the healthcare scope, for example, lifestyle and environment.
2. Life expectancy is only one of many indicators of health; higher life expectancy does not mean that the health of the population is better.
3. The health system and health financing in these two countries are very different.

Life Expectancy versus Health Expenditure 2015



Source: Our World in Data (2017). CC BY 4.0.

### Levels of Healthcare

Healthcare provides different types of care, e.g., preventive, curative, and palliative, with the main purpose of maintaining and improving the population’s health status. Accordingly, healthcare involves all different levels, from the upstream prevention concept of public health to end-of-life (palliative) care.

First, we need to be clear about the difference between levels of prevention and levels of medical care. The concept of prevention is usually heard in the public health area, where one of the highest priorities is to protect the population from risk factors of a disease. There are three levels of prevention: primary, secondary, and tertiary (WHO, n.d.). Primary prevention refers to upstream interventions to improve the overall health of the population, such as vaccination or health promotion strategies (e.g., smoking and alcohol regulations). Secondary prevention includes the early detection of the disease for a better treatment approach, for example, through disease screening programs. Tertiary prevention integrates patient-centered disease management, which helps protect patients from complications, morbidity, or death.

Levels of medical care normally refer to delivery in healthcare, which can be categorized into four levels: primary, secondary, tertiary, and quaternary care. Primary care refers to a patient’s initial consultation, which is normally with a general practitioner. Secondary care involves the short-term treatment of patients who are acutely ill or injured, and usually, but not always, takes place in the emergency department of a hospital or the intensive care unit. The attendance of medical professionals during childbirth and medical imaging services are also categorized as secondary care. Tertiary care refers to long-term treatment, often on an in-patient basis, with medical specialists, and usually requires referral from a lower level of healthcare delivery. For instance, cancer management, surgeries, and palliative care are counted in this level. The final level of medical care is quaternary care, normally considered an extension of tertiary care, but focusing on uncommon diagnostics or rare diseases that are not widely spread.

### The Demand for Healthcare

The demand for healthcare is derived from the need for health. Here, we will address the concept of demand in economics and further look into how the demands for healthcare are similar and different from demand for other goods.

#### The concept of demand in economics

Demand is individuals’ wish to consume a certain good to fulfill their needs. They make choices based on their knowledge about what goods they need, how much to consume, and what is available at what price.

There are four main factors affecting the demand of a good that also influence choices made by consumers: (1) the price of the good, (2) the price and availability of the alternatives, (3) the price and availability of the complementary goods, and (4) the income of the consumer.

Depending on the specific goods, there are some additional factors that may also affect choices among consumers. For example, fashion trends may affect clothing choices and personal habits may affect the choices of healthy or unhealthy foods. Brand impact may also be another factor, especially concerning high-end quality products.

### The Concept of Elasticity

The concept of elasticity refers to the responsiveness of demand to changes in contributing factors, for example, price, income, or availability of alternative goods. The most common and useful type of elasticity is **price elasticity**, which shows how changes in price affect the demand for a good. This indicator also reflects the necessity of the goods; some goods are so essential that consumers are willing to buy them no matter the price. The normal way to calculate price elasticity is

**Price elasticity**

This is the measure of how responsive demand is to price.

Other types of elasticity include cross-elasticity, which reflects the sensitivity of the demand for one good depending on the price of other related good (i.e., substitute or complementary), and income elasticity, which reflects how changes in income affect the demand. These concepts are well-described in the basic reading by McPake et al. (2020, pp. 13—24).

#### What makes demand for healthcare special?

The aforementioned principles of economics still apply to the healthcare market at a certain level. However, there are some characteristics that make the healthcare market special.

First, in many cases, the consumers of healthcare are not the decision-makers for their health services; for example, patients usually follow advice from doctors or take their opinion into consideration when decision-making. Since healthcare is complex, personnel with advanced healthcare knowledge (i.e., physicians) are better equipped to deal with health problems. This is known as information asymmetry in healthcare market. Moreover, in many cases, the outcomes of treatment are very uncertain, which makes it difficult for the patients to make choices, so they have to depend on doctors’ recommendations to determine the best choices.

Second, demand depends not only on the preferences of the consumers, but also the ability to pay for the goods. In healthcare, health services may not be paid for directly by the consumers (i.e., patients) themselves, but via, e.g., the government or social/private insurance firms. Situations like this might reduce the price elasticity of healthcare (i.e., via government price regulations) compared to other goods, and therefore substantially affect healthcare demand.

Another important aspect that makes healthcare different from other goods is the uncertainty in healthcare needs. For instance, today we are healthy, but we do not know whether we will be healthy or ill tomorrow. This uncertainty increases in specific groups, such as elderly people or people living with comorbidities. This explains the demand of health insurance mentioned in the previous paragraph.

Unlike some other goods and services that we enjoy using in a short-term manner (e.g., eating foods or buying a car), healthcare is more of a long-term (possibly uncertain) investment at the expense of short-term pain or inconvenience. Therefore, there may be reluctance toward the demands for healthcare if consumers are not willing to endure the short-term discomfort in exchange for long-term improvement.

### Self-Check Questions

1. Please describe one way the demand for healthcare differs from the demand for other consumption goods, according to Grossman’s model.

*Health is what the consumer wants, and healthcare inputs are demanded to produce it.*

*Individuals cannot purchase health directly from the market. They produce health by investing time in health-improving activities and purchasing medical inputs.*

*Health does not depreciate instantly, it lasts for an extended period. Health can be considered a capital good.*

*Health can be regarded as both an investment good and a consumption good. People invest in health to increase the number of days they are available to work and to earn extra income. Health consumption also makes people feel better.*

1. Please define price elasticity.

*Price elasticity is the measure of how responsive demand is to variations in price.*

1. Please fill in the blanks to complete the following definition.

The concept of *elasticity* refers to the responsiveness of demand to changes in contributing factors, for example, *prices*, *incomes,* and the price and availability of alternative goods.

## 1.2. Supply: Resources, Production and Costs

Previously, we looked at many of the different aspects of demand in healthcare. In this section, we turn to the supply side. First, the various resources of healthcare, including financial, human, and structural, are presented. Then, the production and costs of healthcare are enumerated under both the short- and long-term.

### The Resources of Healthcare

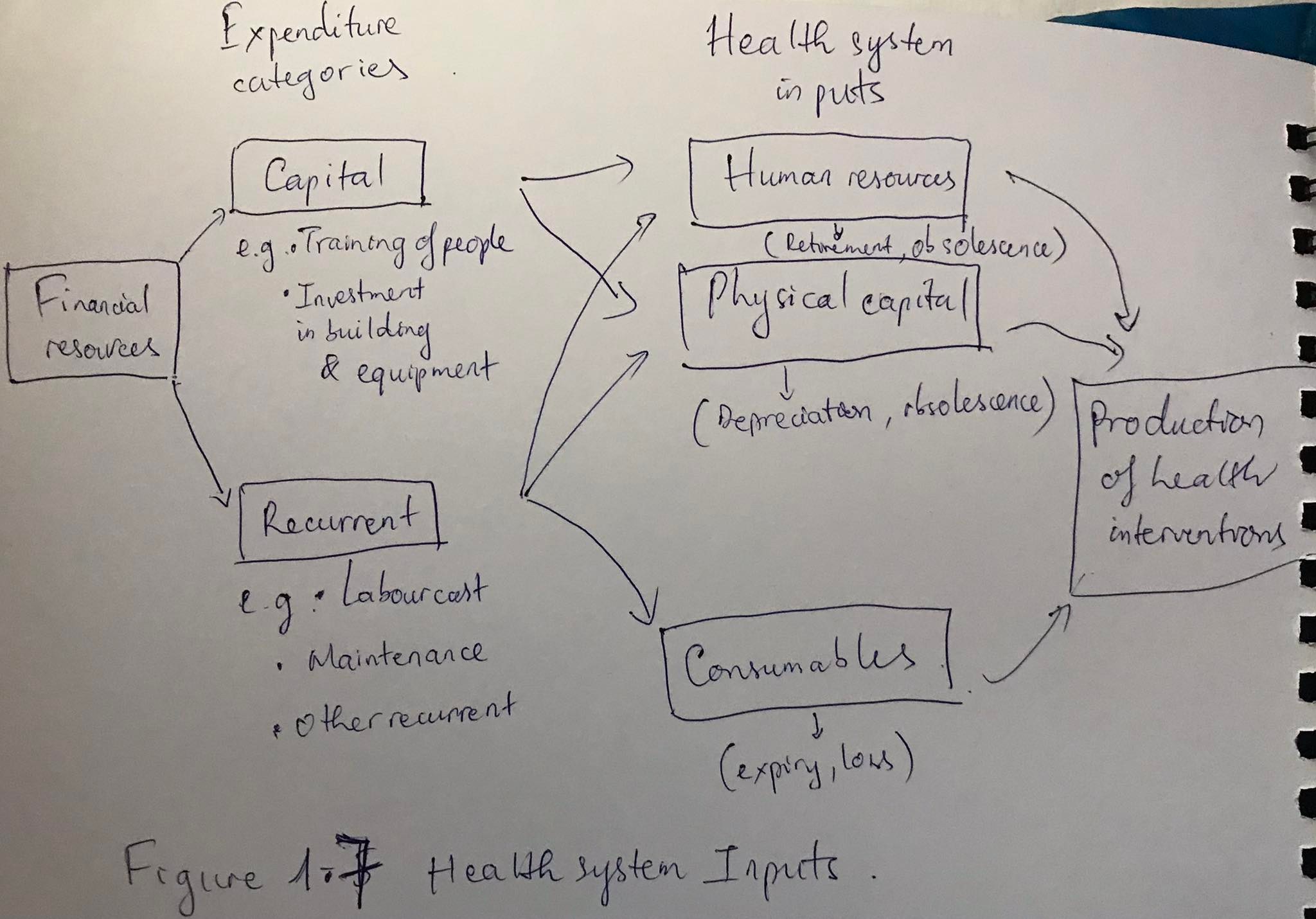
Because there are many inputs used to supply health services, they need to be balanced in order to have an effective healthcare delivery. For example, without the adequate medical equipment, physicians cannot perform their jobs effectively. Therefore, a hospital should not only invest in attracting human resources, but also in new medical equipment and maintaining the existing infrastructure. This balance of resources should be maintained across time and geographical borders (Hernández et al., 2006). In practice, there are usually mismatches between investment and recurring expenses, as well as between different types of inputs, which create roadblocks to optimal performance. To minimize future disparities, new investment decisions must be carefully evaluated, and the present input mix must be reassessed on a regular basis. Clear legal guidelines and incentives are essential if purchasers and providers are to adopt efficient practices in response to health requirements and expectations. The mixture of resources is varied contextually due to the resources available and the investment in healthcare.

#### Resource categories

How are financial resources allocated to create healthcare inputs? The investment decisions in healthcare, similar to other industries, are generally irreversible; a significant amount of money is invested in a variety of locations and activities that are difficult, if not impossible, to cancel or change. In addition, given the scarcity of financial resources in the healthcare sector, the investment should be critically evaluated before its implementation.

Healthcare is a complex system, consisting of various types of resources generated from top-down investments. The financial resources of healthcare can be split into two categories: capital and recurrent expenditure (see the figure below). Some examples of capital expenditure include investment in buildings and equipment (e.g., clinical units and other health facilities) and the training of people. When talking about the government’s role in healthcare, people may only think about the responsibilities of the Minister of Health. However, in reality, certain healthcare investment decisions are outside the Ministry of Health’s control, making maintaining the overall balance of inputs even more challenging. For example, the training of health professionals (e.g., physicians and nurses) is frequently the responsibility of the Ministry of Education rather than the Ministry of Health, and private investments in infrastructure and investment, such as private hospitals, may upset the balance. To maintain the healthcare activities, recurrent expenditure is also required. This consists of the expenditure in laboring costs; maintenance costs; and other costs, such as the costs of pharmaceutical productions.

Health System Inputs



Source: Duy Pham (2022).

As seen in the figure above, many types of financial investments contribute to three major types of inputs: human capital, physical capital, and consumables. These inputs contribute directly to health interventions at various levels of health prevention and treatment. These three inputs are inextricably linked. It makes no difference if the facilities, diagnostic equipment, and medications are not functioning, even when the knowledge, skills, and personnel levels are high. Balancing the mixture of resources is critical and contextual-based. We will discuss the relationship between these principle inputs and provide examples of their current mixture in different contexts in the following section.

#### Balancing the mix of resources: Real-world examples

One of the most essential inputs into the health system is human resources, which includes many types of clinical and non-clinical workers who enable both individual and public health action. The knowledge, skills, and motivation of individuals who provide services ultimately influence the success of healthcare systems. Moreover, the human resources’ recurring payment is often the single greatest item in the health recurring budget. Human resources expenditure accounts for nearly half (49.4 percent) of total global healthcare spending reported by the World Health Organization (WHO) in 2006 (Hernández et al., 2006). People, on the other hand, would be unable to perform services efficiently without the other inputs, such as physical capital (e.g., equipment and hospital) and consumables (e.g., pharmaceutical and services), both of which play a significant role in enhancing human resource productivity.

Healthcare systems are labor-intensive, necessitating the use of skilled and experienced employees in order to work correctly. A balance must be struck between the various types of health promoters and caretakers, as well as between health personnel and physical resources. Recruiting physicians to perform the most basic tasks would be a waste of money. The added value of additional units of health input tends to decrease as the value added by each subsequent unit of input grows, which is known as the law of diminishing marginal benefit in economics. For instance, if there are too few physicians, adding another physician will enhance healthcare; yet, if there are currently too many physicians, adding another physician will likely raise expenses rather than improve treatment. This concept is similar to the original supply curve in economic theories. Despite the oversupply of physicians, Kaitelidou et al., (2012) pointed out that the geographical distribution of human resources is highly unequal. Moreover, while there is an oversupply of some specialists (e.g., cardiologists), the number of general practitioners is far below what is needed.

Limited wages and benefits, as well as inappropriate working conditions ranging from labor in conflict zones to insufficient infrastructure and shortages of vital drugs and consumables, are commonly identified as the most serious issues affecting the healthcare workers in developing nations (Willcox et al., 2015). In certain nations, the great majority of government hired physicians visit private-paying patients to augment their normal employment income. For instance, informal payments (also known as thank you payments) are one important factor of creating corruption in the health sector in Vietnam (Vian et al., 2012).

Another contrast between human and physical capital that affects how people are managed is that doctors, nurses, and other healthcare employees are not only driven by their existing working circumstances, salary, and management. Additionally, they are impacted by their predictions for future situations based on their experiences, other people’s perspectives, and present trends. If eligible workers feel that their future compensation, benefits, and working conditions will deteriorate, their employment choices and motivation will reflect this perception.

To address the aforementioned issues, the first step is to establish a sustainable balance between the various kinds of resources, which are often context-dependent. Three tactics have been tried with varying degrees of success (WHO, 2000):

1. Maximizing the efficiency of available workers with a more geographically distributed workforce
2. Wherever possible, relying more on multiskilled individuals
3. Ensuring that talents and functions are more closely aligned

These strategies have been used differently in different countries and have recorded some successes. In Canada and the Scandinavian countries, the government provides multiple incentives to attract health professionals in the Northern regions to reduce the gap of healthcare inequality between urban and rural areas (WHO, 2000). In another example, some countries have successfully recruited foreign health professionals to increase the physician-population ratio, such as Saudi Arabia, Fiji, and Oman (WHO, 2000).

A number of health professionals have been provided with additional trainings to meet the needs. For example, public health schools established in Hungary and Jamaica to develop skills in epidemiology, statistics, and management to stratify different levels in the healthcare workforce (WHO, 2000). Additionally, globalization promotes increased mobility and offers the possibility of studying abroad. This, however, raises worries about brain drain, particularly when moving from low- to high-income nations. In many situations, health professionals educated overseas in industrialized nations want to remain in their own country.

### The Production and Costs of Healthcare

Several healthcare options are accessible. While certain components need highly skilled workers, sophisticated technology, and costly consumables, others require compassion and understanding. Some are fully mechanized, others resemble handicrafts, and while certain tasks need teamwork, others require experts to work alone. They are provided in a number of settings, and often, a single intervention is supplied as part of a broader package of related services. Healthcare professionals have an understandable aversion to considering their work as just industrial, especially when the optimum delivery of services necessitates not only technical expertise but also a caring attitude. However, there is a lot to be gained from other types of product and service development.

The following focuses on the production of healthcare and their related costs. The economic concept of production function will be revisited along with a more detailed focus on healthcare deliverables. Different economic perspectives on efficiency will also be discussed and illustrated.

#### Input and substitution

A production function is a mathematical expression that describes the relationship between inputs and outputs. In other words, production is a process of transforming or arranging matter (inputs) in order to make it more suitable for meeting a demand (Marshall, 1920). The inputs (or resources) in healthcare have been described previously, and categorized into three types: human resources, physical capital, and consumables. As a result, the production function of healthcare (Q) will be as follows:

where H stands for human resources, P for physical capital, and C for consumables.

**Substitution**

This is the interchangeable amount of inputs to produce the same output.

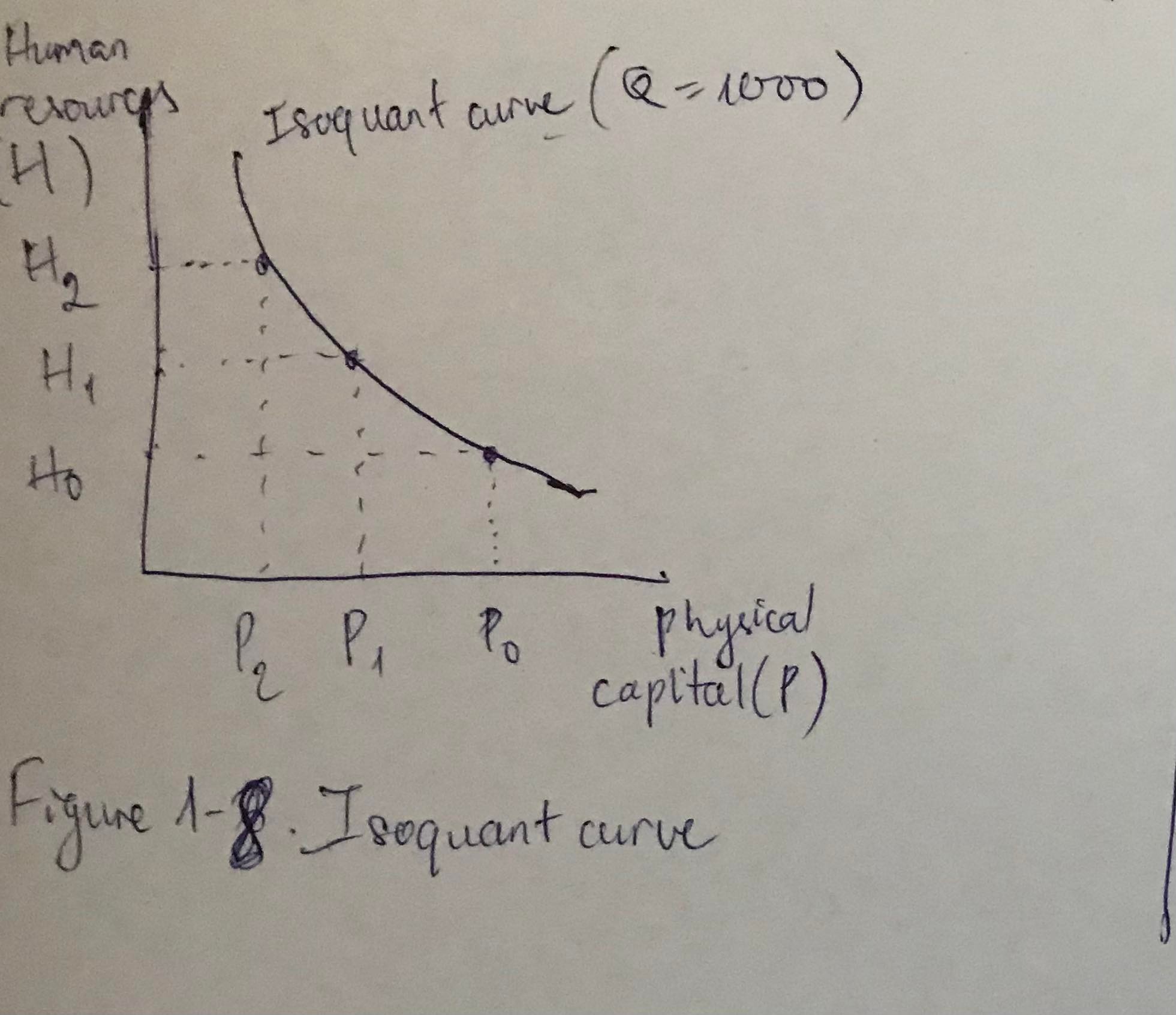
The interchangeable amount of inputs to produce the same output is the **substitution** in the view of economists. For example, to achieve a certain level of health output, we may either increase the amount of equipment (which is highly capital-intensive) or increase the number of personnel (a more labor-intensive choice). We can hire more experienced employees, employees with less training but more supervision, or employees that closely follow protocols. In the following, we will simplify the analysis by considering only two inputs: human resources (H) and physical capital (P): .

Let us take the example of a breast cancer screening program for 1,000 women that requires two inputs: physical capital P of mammography units (e.g., X-ray machines) and human resources H (e.g., clinicians). An **isoquant** (indicating “equal quantity”) can be expressed by identifying all combinations of H and P that create a given quantity of output. This isoquant here is the screening result of 1,000 participated women. The horizontal axis in the figure below represents the number of mammography units, P, and the vertical axis represents the number of clinicians involved, H. The mammography unit and clinician combination that yields the output of screening for 1,000 women are represented as points on the isoquant curve. The isoquant’s downward slope indicates that if we use less of one input factor, we must compensate by using more of the other. Because of diminishing marginal productivity, the isoquant is convex. When we consume a large amount of one element, its subsequent units are less productive. It is also important to notice that the isoquant curve does not cross either the x- or y-axis, which means that only one input will not be able to create the desired output in this case. It is understandable that if we have hundreds of X-ray machines but no one to read the results, we cannot have the output of screening results. On the other hand, if we have many clinicians but no X-ray machine, we also cannot produce mammography screening.

**Isoquant curve**

This is a contour line drawn across the set of points at which the same amount of output is generated when the amounts of two or more inputs are changed.

Isoquant Curve



Source: Duy Pham (2022).

In the simplified example above, we can see that the substitution of different inputs can achieve the same level of output, as indicated by the isoquant. So, why is this isoquant concept important? Firstly, there is more than one way to treat a disease and more than one way to improve health. Although the number of substitutions in healthcare is normally limited compared to other markets, in most cases, it is still feasible to make a choice between alternative technologies or different combinations of inputs. Secondly, the concept of the isoquant and its further application helps to identify two levels of efficiency. These efficiency concepts are used as the foundation for decision-making in healthcare.

#### Technical efficiency

Technical efficiency is achieved when an isoquant combination of inputs is selected. This means we do not waste input resources. On the isoquant, all points reflect technically efficient combinations. We would need more input variables than necessary if we choose a point to the north-east of the isoquant to create the same quantity. It would thus be able to lower the amount of labor and/or capital used while maintaining the same level of output.

However, it is easier said than done given the complexity and variety of both healthcare inputs and outputs. We have previously seen the complexity of health inputs and their possible combinations. Moreover, the problem might come from how we define the output. In the example of the breast cancer screening program above, we can define the output as the number of women screened, the number of breast cancer cases detected, or even death and disability prevented. The selection of output directly reflects the efficiency of the program. If we simply consider the number of women screened in the program as the output, we might not be able to evaluate the impact of the program on population health. This problem is not rare in health intervention evaluation, where the intermediate outputs or processes are evaluated instead of the appropriate health outputs.

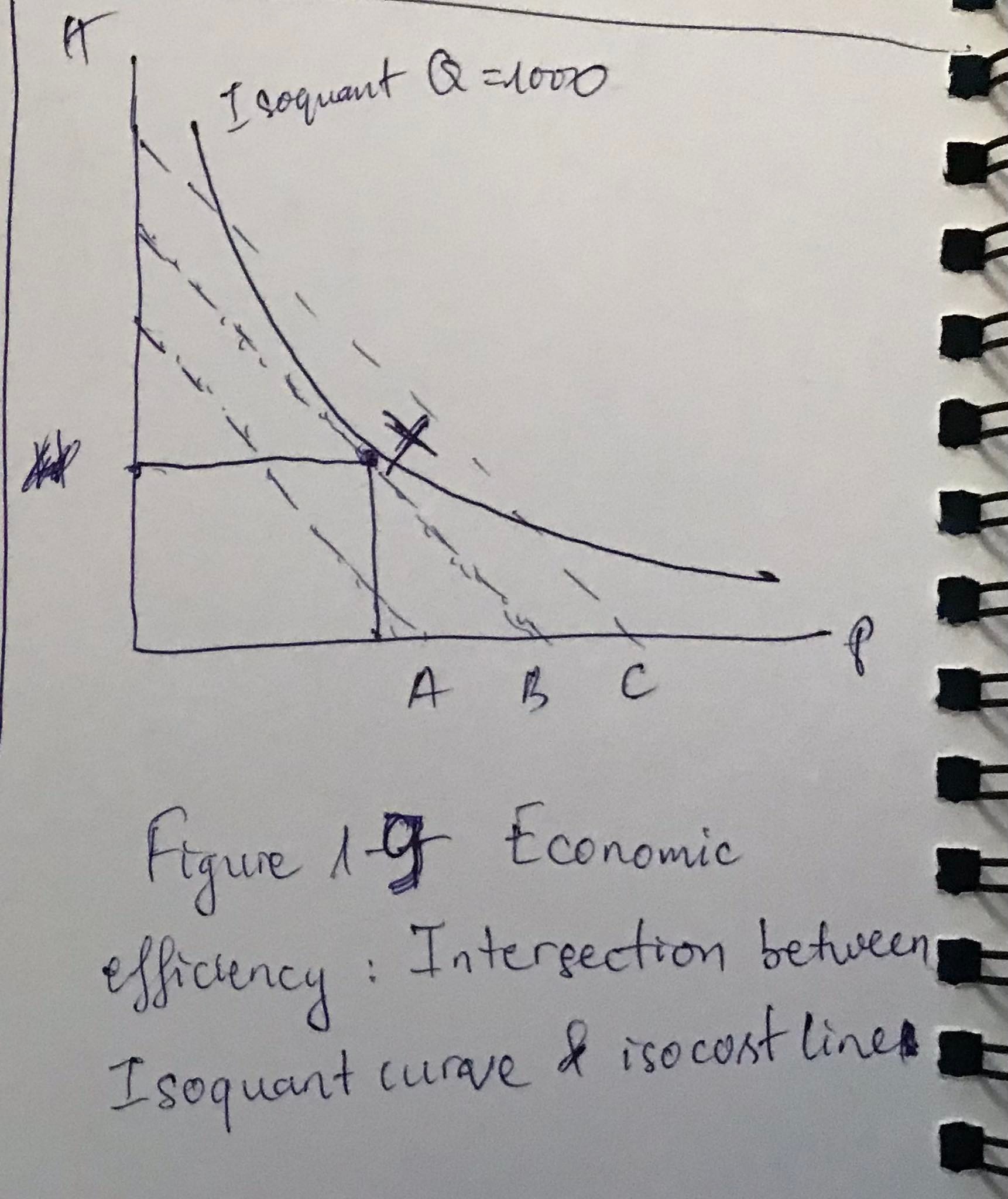
#### Economic efficiency

**Isocost line**

Points on an isocost line are points of equal cost of production.

We don’t just want to know if production is technically efficient; a further step is to reduce the cost of service production. A method for determining the most cost-effective technical combination of manufacturing elements is needed for this. This concept refers to economic efficiency (or cost efficiency), which is accomplished by including **isocost lines** in the analysis. The figure below added isocost lines to the isoquant curve in the example of the breast cancer screening program. The goal of economic efficiency is to minimize the cost of production of any given quantity (and quality) of services. In the breast cancer screening example, if we know the average wages earned by a clinician and the cost of an X-ray machine (ignore the depreciation cost for now) to perform mammography screening, we can draw the isocost line for any given budget (A, B, and C in the order of budget increase). At the budget line A, we do not have enough money to perform 1,000 mammography screening; whereas it is unnecessarily generous at the budget line C. The budget line B is feasible enough to perform the needed quantity (1000). The economic- or cost-efficient point is where the budget line is tangential to the isoquant of the quantity (and/or quality) of a given output.

Economic Efficiency: Intersection between Isoquant Curve and Isocost Line



Source: Duy Pham (2022).

In our example, it is reflected at point X, where the combination between the number of clinicians and X-ray machines can perform 1,000 mammography screenings at the lowest cost.

Remember that the phrase “highest quantity for the lowest cost” is inappropriate. Either we maximize output for a given budget or we try to minimize costs for a given volume of output. Both refer to the concept of economic efficiency.

#### The cost function in healthcare

**Opportunity cost**

This is the value of a resource’s best alternative use.

The concept of opportunity cost, which is typically understood as the value of the resource’s best alternative use, is one of the key concepts in economics. This notion is not as straightforward as it seems. There are several difficult problems that must be addressed in order to assess it, such as what influences the price of inputs, the possible alternatives, which one is the best, and how to determine the best alternative use of resources?

In the following, we mainly focus on the financial part of the cost, or the monetary value for production, by investigating its determinants and patterns, based on economics principles adapted to the healthcare market.

#### Economies of scale and economies of scope

Isoquant and isocost analyses give a framework for analyzing the expected links between healthcare inputs and their costs. An additional analysis of patterns of costs is essential for determining the most cost-effective method of providing services. One particular concern is whether there are economies of scale in healthcare services (i.e., does the unit cost of service production decrease as production scale increases?) If this is the case, it is more advantageous to have fewer large hospitals.

Another related concept is economies of scope, which is highly relevant for a multi-product firm, including healthcare. Economies of scope arise when it is more cost effective to manufacture two or more distinct items or services jointly than it is to create them individually.

Short- and Long-Term Cost Functions

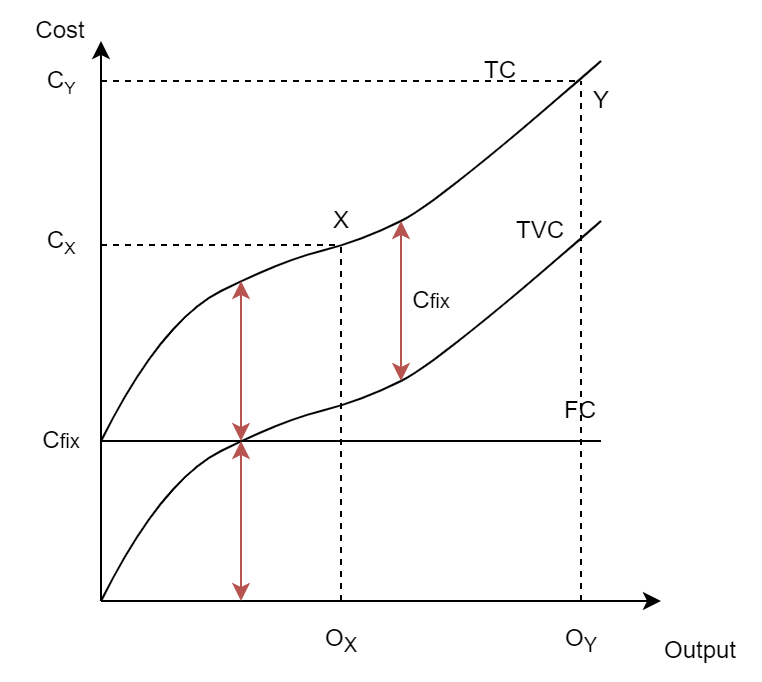
**Cost function**

This is a functional relationship between cost and output.

There are debates in research regarding the economies of scales and scopes of healthcare market. Since healthcare is an imperfect competitive market and very contextual, the research claims may be true in one particular place, but not others. Estimating **cost functions** for healthcare provision in a particular context can allow a better understanding.

There are two time periods of cost functions that should be noticed when doing cost analyses: namely short- and long-term cost functions. The short run is the time period during which a healthcare production unit (e.g., hospital) must continue to account for fixed costs, that is, inputs that cannot be modified. For instance, the short-run function would account for the majority of the fixed expenses associated with the number of beds. On the other side, the long run is accomplished when the hospital’s fixed obligations are terminated and expenses grow solely when the amount of production increases (Folland et al., 2017). This section focuses on the short-run cost function.

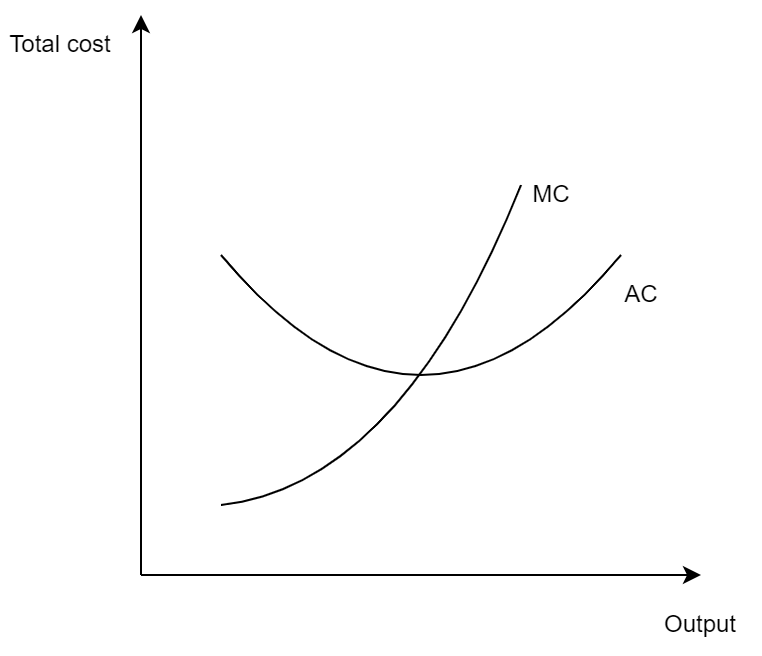
Short-Term Total Cost Curves



Source: Duy Pham (2022).

The figure above illustrates the short-run total cost curve, which is the sum of total fixed costs (FC) and total variable costs (TVC). We can determine the average cost using the whole cost curve. At lower levels of output (from 0 to OX), the average cost is decreasing, whereas at higher levels of output (From OX to OY), it is increasing. We may also estimate the marginal cost from the total cost, which is the additional cost of generating one more unit of output (McPake et al., 2020). The figure below depicts the average cost (AC) and marginal cost (MC) curves. At the lowest point of the AC curve, it crosses the MC curve, which is logical. When the average cost of a service decreases, then the cost of one additional unit (the MC) remains below it. On the other hand, the expense of adding another unit is higher than the average cost of a service when the latter increases. This is why the two curves intersect at the lowest point of the AC curve. The contours of the cost curves in this short-run research indicate “returns to a factor” rather than “returns to scale,” since they show the variation in cost when one component of production (or input) is adjusted while the other stays unchanged (McPake et al., 2020).

Marginal and Average Cost Curves



Source: Duy Pham (2022).

### Self-Check Questions

1. Which of the following is defined as a contour line drawn across the set of points at which the same amount of output is generated when the amounts of two or more inputs are changed?

* isocost curve
* isoquant line
* *isoquant curve*
* isocost line

1. Please name one of the three principle health system inputs and give 2—3 examples.

*human resources (e.g., doctors, nurses, and community health workers)*

*physical capital (e.g., hospitals, department of health building, and clinics)*

*consumables (e.g., medicines, medical devices, and supplements)*

1. Please differentiate between technical and economic efficiency

*Technical efficiency is achieved when an isoquant combination of inputs is selected. This means we do not waste input resources.*

*Economic efficiency is the combination between technical and cost efficiency. It means that we do not waste the input resources and, at the same time, the cost of production needs to be minimized.*

## 1.3. Asymmetric Information and the Agency Relationship

In a perfect competitive market, the customers should be provided with enough information about costs, quantities, and the production function from the inputs to the outcomes. Economists in perfect competitive market looked at how people make decisions in a model that assumed customers were provided with perfect information (Folland et al., 2017). As the healthcare market is imperfect, such an assumption cannot be made. A more thorough understanding of the health economy necessitates a deeper understanding of the effects of various informational difficulties in healthcare marketplaces.

**Adverse selection**

This is a phenomenon in which insurance attracts patients who are more likely than average to utilize services.

As a result of asymmetric information and agency connections in many healthcare markets, such as insurance markets, there is a high degree of uncertainty. Patients who are more likely than average to use services are drawn to insurance because of asymmetric information, for example. In the opinion of the majority of healthcare specialists, prospective beneficiaries know more about their health status and need for treatment than their insurer does (Folland et al., 2017). Over-insured patients will be enticed by reduced rates for high-risk patients and higher premiums for low-risk patients. The efficiency of the health insurance market is reduced by the financial transfer from healthy to more vulnerable individuals. Adverse selection is defined as “a phenomenon in which insurance attracts patients who are more likely than average rate, results from asymmetric information” (Folland et al., 2017, p. 196).

In this section, we will discuss asymmetric information specifically in healthcare in three areas, from individual to societal level. First, the insurance market and its adverse selection characteristics will be addressed. Second, agency relationships will be discussed, including relationships between individuals (doctor-patient relationship) and a wider application at societal level (social welfare). Finally, the effects of asymmetric information on prices and quality of healthcare will be examined and illustrated.

### Health Insurance: Adverse Selection

Information asymmetry is anticipated to occur as a result of prospective insureds knowing more about their projected health expenses in the future than the insurance company (Folland et al., 2017). Consider the following scenario: A prospective insured knows precisely what they will spend in the future, while an insurance company only knows how much money will be spent by all of the individuals who will be insured in the future. It is crucial to understand the difference between information asymmetry and poor information in this context. We can consider the imperfect information as a broader term that covers **information asymmetry**, which should be understood as the unequal level of information between two groups. However, imperfect information can also cover the incomplete information, where both sides have limited information, which is not the case in health insurance. Adverse selection would not occur if patients were no better at anticipating their health costs than the insurer.

**Information asymmetry**

This is the unequal level of information between two groups (e.g., insurers and insureds in health insurance).

What inefficiencies are caused by adverse selection in health insurance? In most cases, people cannot predict exactly how much money they will spend in the future due to the unexpected outcomes in healthcare. The primary reason for insurance is to protect against risk. When lower-risk individuals are matched with higher-risk individuals and everyone pays identical premiums, the lower-risk individuals pay higher rates and are more likely to be underinsured, which is the case with most insurance policies. A loss of well-being results from their inability to get insurance at prices that are appropriate for their level of risk. Higher-risk individuals, on the other hand, will pay a lesser premium and, as a result, will over insure; they will cover hazards that they would not have otherwise insured against. This is inefficient. The income will be transferred from low-risk to high-risk customers, which will result in both inefficiencies and revenue shift.

### The Agency Relationship and Social Welfare

An agency relationship describes a situation where an individual who, in healthcare, is often the patient grants the power to make decisions to someone else (the agent). The patient (principal) has delegated power to the doctor (agent) who, in many cases, will also provide the therapies that were ordered (Folland et al., 2017). This delegation of agency is primarily motivated by the patient’s understanding that the physician is the one with the education, training, and experience to make the most appropriate judgements. Thus, delegating this responsibility to the doctor, as a competent agent, is the most effective strategy to remedy the patient’s lack of knowledge. Consequently, information asymmetry and agency are closely relevant.

#### The doctor–patient relationship

Doctors have differing perspectives on what patients expect from them: Some argue that it is their responsibility to tell patients what treatment they should have, while others argue that it is their responsibility to provide patients with information so that they can make their own decisions. In principle, this agency connection isn't an issue because the agent’s (doctor’s) utility purpose is the same as the principal’s (patient’s): to maximize each patient's utility. However, the special characteristics of healthcare make this difficult. Take the following example: A treatment decided by a doctor causes side effects that dissatisfied the patient, resulting in their decision to stop the treatment without observing the long-term effects on their health outcome. Thus, the information asymmetry does not allow a doctor to act as a perfect agent for their patient.

From a reversed perspective, to make the right decision, the patient should give the doctor all the necessary information, after which the patient is expected to adhere to the decision that the doctor makes. The situation is that, in many cases, the patients cannot fully describe their health problem. Therefore, doctors sometimes have to rely on other medical technologies before making decisions. Again, even though the problem is diagnosed correctly, treatment choices are still a concern.

#### Social welfare

According to discussions about the individual doctor-patient relationship, a perfect agent is a doctor who gives the patient the combination of services that the patient prefers. This is in conformity with the medical code of ethics to the extent that the patient’s own interests are centered on their own health, which is a good thing. The ideal agent, in the event of a conflict, focuses on the wishes of the patient rather than their own. However, the patient’s wishes may differ from those of society. This brings up the question of who the doctor represents in the end: the patient, a group of patients, healthcare funders, or society at large? Doctors, like the rest of us, are unable to please everyone all the time; thus, the answers to these questions will define the ideal agent.

It is claimed that if doctors supply fewer non-health-enhancing treatments than their patients want, they will be excellent agents for their patients (Olsen, 2017). The total healthcare budget is increased by limiting waste, allowing doctors to serve more patients. This model assumes that our willingness to cross-subsidize healthcare is conditional on that care’s efficacy in improving health.

### The Effects of Asymmetric Information on Prices

A study on the impact of consumer knowledge and competition in primary healthcare services (Pauly & Satterthwaite, 1981) asserts that basic medical care is a reputation good for which an increase in the number of providers would result in an increase in the price of services. The reasoning that underpins this unexpected assumption is sound. It is reasonable to assume that the average client seeks information from other customers about their supplier experiences. Therefore, when the number of doctors increases, the average number of friends who visit any one provider declines, reducing the average amount of information that is available to the public. The consumer’s ability to respond to pricing and other practice aspects is determined by their understanding of—that is, information about—the available choices. As a result of the reduced knowledge, the firm demand curves’ price responsiveness (i.e., elasticity) decreases, leading equilibrium prices to rise. Reduced information, according to economic theory, tends to give each business more monopolistic power.

However, the role of imprecise pricing information in the development of monopolistic power should not be overstated. Remember that every buyer of a product does not need complete price knowledge to get moderately competitive pricing circumstances. Most customers lack comprehensive pricing knowledge for many of the items and services they purchase (they have no idea what rival sellers charge). Nevertheless, the competition between producers affects prices. Given an example in a grocery shop, the difference in prices between the similar products is not quite substantial. There is only slightly different pricing due to the value of the brand, quality of the products, or even attractive packaging. In fact, the customers do not have perfect information about the price of many of the products. This shows that insufficient pricing knowledge does not contribute to a great impact on the price. In the healthcare sector, where many services are totally or partly covered by insurance, there are extra variables to notice. While a patient’s sensitivity to price levels and price differentials in the choice of providers may decline with time, third-party payers, such as health insurance companies, have taken on the responsibility of monitoring prices (Folland et al., 2017). The actual reimbursement is frequently less than the provider’s rates due to selective contracting and other fee arrangements. This statement is particularly relevant to the publicly-funded healthcare systems.

### Self-Check Questions

1. The unequal level of information between two groups is referred to as …

* … adverse selection.
* *… information asymmetry.*
* … imperfect information.

1. Please complete the following sentence.

*Adverse selection* occurs when insurance enrolls individuals who may be *more* likely than average to use their services.

1. Please complete the following sentence:

According to Pauly and Satterthwaite (1981), primary medical care is a *reputation good*, and a rise in the number of practitioners may result in an increase in pricing.

## 1.4. Externalities

Externalities are the next market failure in healthcare that must be taken into consideration. It is important to remember that market failures are not necessarily a negative thing, as may be shown in the case of positive externalities, for instance. The idea of externalities may only be understood when seen from a social viewpoint in which there are interactions between people and organizations. In other words, an individual’s actions may have an impact on the conduct of others in the community. This section will begin by discussing the fundamental notion of externalities before delving into the many forms of externalities that exist in healthcare.

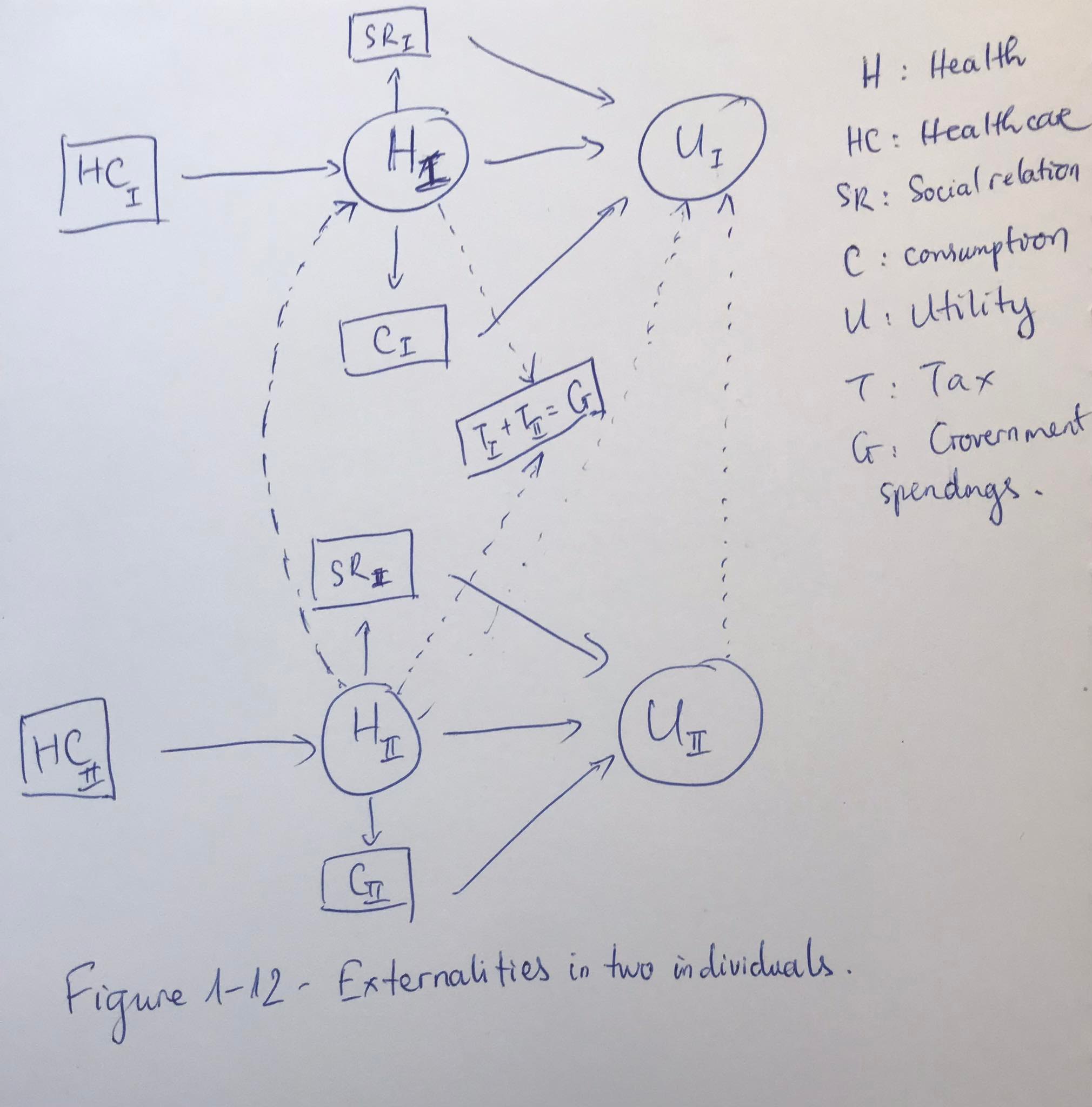
### The Concept of Externalities in Healthcare

An externality is a cost or benefit that a producer incurs or receives but does not pay for. Externalities may be beneficial or detrimental, and can result from the production or use of a product or service. Negative or positive effects may be personal (to an individual or an organization) or societal (for the entire community).

According to one illustration of the many forms of interpersonal interactions in health, there are four distinct ways in which a person’s improved health because of their healthcare usage may impact the usefulness of another person (Olsen, 2017). The figure below depicts externalities in the usage of healthcare. Individual I’s utility through their personal use of healthcare, as well as indirectly through individual I‘s health (coming from B’s use of healthcare) can be formalized inside an extended utility function:

This function can be explained as follows: The utility of an individual I not only depends on their own social relations (SR), consumption (C), and health (H), but also on four partial links with individual II, in terms of i) contagion H­­­­I (HII), ii) economic contributions from II, G (HII), iii) paternalistic altruism (caring for health) HII, and iv) general altruism (caring for general well-being) UII (Olsen, 2017).

Externalities among Two Individuals



Source: Duy Pham (2022), based on Olsen (2017).

### Positive Externalities

The externalities that are most likely to have a significant impact on healthcare markets are those that are positive. The most important of these are herd immunity and caring, which are explained in more detail below.

#### Herd immunity

Positive externalities exist in the prevention and treatment of various infectious disease. When a dominant percentage of population is vaccinated or infected, the remaining minority of unvaccinated people are unlikely to become infected. This is the concept of herd immunity. In the case of the coronavirus (COVID-19) pandemic, the World Health Organization (WHO, 2020) recommends that herd immunity against COVID-19 be developed by protecting individuals via vaccination rather than by exposing them to the virus that causes the sickness. It is necessary to vaccinate a substantial percentage of a community to successfully build herd immunity against COVID-19, hence limiting the quantity of virus that may potentially circulate throughout the population. The purpose of achieving herd immunity is multifaceted; one goal is to keep vulnerable individuals secure and protected from sickness when they are unable to get vaccinations (for example, owing to health issues that may cause adverse events or vaccine allergies). Currently, there is still uncertainty about what proportion of the population must be vaccinated to achieve herd immunity in the case of COVID-19. There might be a reason that this virus develops new variants quickly and each variant has its own rate of infections and severity.

#### Caring externality

The caring externality is the second form of positive externality that may be significant in health. It appears that we are significantly more concerned about whether others have access to healthcare than we are about whether they drive an expensive car or dine out frequently. Referring to the previous figure on the externalities between two individuals, it is clear that the utility and health of an individual will be improved if surrounding individuals have good health.

### Negative Externalities

Negative externalities, such as pollution, are well-known in the healthcare market. They are often (but not always) produced because of manufacturing. Dyes are released into a river by a dyestuff producer, such as a clothing manufacturing facility, for example. As a consequence, financial obligations are incurred. The water industry must apply additional filtering procedures in order to produce water that is fit for human consumption. As a consequence of the damage done to the river’s ecosystem, fishing has decreased. The aforementioned clothing manufacturer is not automatically affected by any of these charges; their marginal cost curve is untouched and the resulting costs are not absorbed into the sales and purchasing agreements between the clothing manufacturer and their customers (Folland et al., 2017).

Particularly in health, there are some negative externalities that can be pointed out. For example, one individual smoker can affect others through secondhand smoke. Another example is the transmission of COVID-19 to others from a person who get infected.

### Self-Check Questions

1. According to WHO (n.d.), what is herd immunity?

*Herd immunity is the indirect protection from an infectious disease that happens when a population is immune, either through vaccination or immunity developed through previous infection.*

1. Which of the following is an example of positive externalities?

* passive smoking
* pollution
* disease transmission
* *caring*

## 1.5. Market Failure and Its Consequences

Usually, while discussing economic theories, the ideal competition model is brought up. This should be seen as an analytical tool rather than a depiction of reality, particularly in the context of the healthcare market. Remembering that variations from perfect competitive markets are not necessarily negative is essential; positive externality is a perfect example. Differential outcomes are merely an indication that if the market is permitted to decide on price, output, and the distribution of supply and demand, efficient outcomes would not be reached. Results that are perfectly efficient, on the other hand, seem unlikely to be within our grasp in the near future. As soon as a divergence is detected, the goal at hand is to develop a strategy that takes advantage of the resulting inefficiency and outperforms the market as a whole (Folland et al., 2017). Because the government controls and distrusts the healthcare market, it is probable that they bear some blame.

So far, the main market failures related to healthcare have been discussed, including asymmetric information and externalities. There are other failures that can be named in the healthcare market, including public goods and the monopoly (or oligopoly) market. We will briefly touch on the definitions and examples of these terms in this section.

### Public Goods

Whenever someone eats an apple, no one else will be able to consume that fruit. This is rivalry. If they choose to sell the apple to someone else, they will sell to someone who will pay the amount of money they want, and they will refuse to sell them to anybody who will not pay the amount desired. Apples are excludable. An apple is an excellent example of a private-sector good. Public goods, on the other hand, are distinguished by the absence of rivalry and non-excludability.

Non-rivalry refers to the notion that one person’s consumption of a product does not restrict another person from doing the same or similar thing (Olsen, 2017). A theatrical performance is unusual in that once it has been provided for one person, it is completely free to be distributed to as many people as there are seats available in the theatre at that time. However, a theatrical play is not a public good unless it is free for everyone to access. Non-excludability refers to the fact that non-payers are unable to consume. It is impossible to prohibit everybody who uses the roadway from benefiting from it once it has been cleaned. Non-excludable products are almost always non-rival (not the case of a roadway), whereas non-rival goods are almost always excludable.

Environmental healthcare services and health awareness campaigns are among the healthcare services that can be referred to as public goods (Folland et al., 2017).

### Monopoly and Oligopoly

There are natural monopolies in many parts of the healthcare industry, which contains the features restricting the number of providers in the market. This is in contrast to the perfect competitive market, where there is the need for small competitors to vie for a portion of the market’s business, as well as the possibility for small businesses to establish themselves in the marketplace (Folland et al., 2017). There are a variety of factors working against this in healthcare market. Because of the large, fixed costs associated with starting a firm, only those organizations who can access considerable resources from the capital markets are capable of making the first investment necessary to get their venture off the ground. When it comes to market entrants, when investments cannot be transmitted from one to another market, which is known as sunk costs, it is an extremely unattractive type of investment. This is because, in the event of a new entry’s failure, most of the investment will be lost (particularly in cases where the lack of success indicates that the market has limited potential to expand). This is especially true in cases where more complex and high-tech services are involved, for example, advanced medical technologies like X-ray or cancer screening machines, as they necessitate significant capital investments (Olsen, 2017). A more general statement is that this limitation has an impact on the provision of hospital services in general. Only extremely big marketplaces (such as metropolitan areas) allow for more than a few hospitals to compete against one another. This creates a barrier to other companies from entering. Despite the fact that the dominant business’s pricing is high, indicating its high level of profit, the dominant firm can always drop the price for a sufficient length of time to push a new rival out of the market. The new rival will be unable to join the market unless it is able to immediately begin functioning at the same size as the existing company. A similar level of protection is offered by substantial fixed costs in terms of entry barriers.

### Self-Check Questions

1. Please name two critical characteristics of public goods.

*non-rivalry and non-excludability*

1. Please fill in the blanks.

There are natural *monopolies* in many parts of the healthcare industry, which contains the features *restricting* the number of providers in the market.

Summary

The demand for healthcare derives from the need for health, which lies on the basic level of needs of an individual. Health can be measured by different indicators; however, there is no single indicator covering all aspects of health dimensions. One reason is that there are many inputs contributed to the health output, of which healthcare is only one small component. This statement does not mean to underestimate the importance of healthcare since its contributions are recognized throughout history, for example, in improving life expectancy.

Although healthcare is a special market, principle economic concepts are still applicable to some extent, e.g., demand and supply concepts. From these principles, a further step comparing the differences in healthcare market to the perfect competitive market would help to discover the contextual-based healthcare market in different settings. Consequently, the concept of efficiency needs to be addressed in healthcare delivery for its optimal contributions. Other economic concepts were also discussed, including economics of scale and scope, to analyze the cost functions in the short- and long-term.

There are some market failures that distinguish healthcare from perfect competitive market, for example, information asymmetry, externalities, public goods, and the monopoly/oligopoly market. These failures are closely related to each other, together affecting the efficiency level of healthcare delivery. Understanding them can help develop interventions to overcome them and adapt to context-specific health systems.

# Unit 2 – Government Intervention in Healthcare Markets

**Study Goals**

On completion of this unit, you will be able to …

… explain the rationale for government interventions in healthcare.

… identify different forms of government intervention.

… analyze the government involvement in healthcare using case studies of China and India.

… evaluate the rationale for competitive strategies and their application using a case study of the Netherlands.

# 2. Government Intervention in Healthcare Markets

## Introduction

The failure of the market serves as the economic justification for government regulation. Perfect competitive markets, as we have seen time and time again, provide efficient results in the long run. Therefore, when inefficient outcomes are discovered, they must be the product of a market failure. With the purpose of correcting market failure, government interventions enact laws based on the idea that they would remedy any existing market inefficiency and result in an increase in overall efficiency of the economy. Given the number of market failures, it should come as no surprise that the healthcare industries are strictly regulated by every government around the world.

Government interventions may not only deal with market failure, but also address equity, although sometimes it can be hard to distinguish between inequity and inefficiency in the healthcare market at the time of implementation. Most of the time, the unequal healthcare distribution occurs for some groups of people, due to the inefficiency in the market. In this unit, we will begin with a discussion on the economic rationale for government interventions to solve the market failures in healthcare. This is followed by different forms, involvements, and strategies of government interventions using case studies to provide an international perspective.

## Economic Rationale for Government Intervention

In this section, the reasons we need government interventions in healthcare market will be discussed. The government has a role in correcting the market failure, as well as addressing equity in the healthcare market. Considering the situation of poverty in some countries, government involvement is critical to overcome this situation by providing essential healthcare for the poor. This discussion will be followed by the role of correction of market failures, including the monopoly and oligopoly market, public goods and externalities, and information asymmetry. Some other rationales will also be summarized.

### The Roles of Government in Combating Poverty

One significant rationale for government involvement in health is poverty reduction. To eradicate poverty successfully, two steps are necessary: (1) effectively using labor resources, and (2) improving health status by providing basic access to education, nutrition, and healthcare. It is important to note that the investment in the poor’s health also helps improve their educational and economic prospects. Thus, it supplies them with the necessary conditions to get them out of poverty, apart from the immediate benefit of pain relief. Additionally, in the majority of cultures, delivering health and education to the poor requires political support. Investment in the poor’s health is considered a cost-effective and politically acceptable strategy to eliminate poverty.

Healthcare might be out of reach for many of the poor. Giving them a small sum of money would have a minimal impact on the demand for healthcare services. However, providing access to free or low-cost healthcare may lead to a considerable rise in the use of healthcare by the poor, who are more sensitive to price and have a greater illness burden compared to the non-poor (World Bank, 1993). Strict restrictions on the types of healthcare that are covered by public funds may be necessary to ensure that low-income patients get free or reduced-cost treatment. It is common for free healthcare to everyone to be rationed depending on location or quality. The poor may not be able to benefit from such universal interventions. They may, however, have more political support and be better able to deal with insurance market challenges. The poverty incidence and the capacity to pay in a specific setting will determine who should get free or discounted healthcare. It does not mean that spending more money results in more healthcare for the poorest. More importantly is how the government can provide for the people really in need using appropriate strategies. It is understandable that extremely poor nations need more attention in that sense since they do not have sufficient resources to offer free healthcare for all.

### The Roles of Government in the Correction of Market Failures

The following looks at the roles of government interventions in correcting healthcare market failures. Each type of market failure will be discussed in turn, including the monopoly and oligopoly market, public goods and externalities, and information asymmetry. Other rationales will also be pointed out.

#### Monopoly and oligopoly market

Monopoly power is the typical illustration of market failure in healthcare. In contrast to the perfect competitive market where start-up companies are encouraged to enter the market to increase the level of competitiveness, the healthcare sector contains some characteristics that are against this phenomenon (Folland et al., 2017). Because of the large, fixed costs associated with starting a firm, only those organizations who can access considerable resources from the capital markets would be capable of making the first investment necessary to get their venture off the ground. This limits the competitive level in healthcare market, and if the monopolist does not provide appropriate control over the market (which is mostly the case), inefficiencies will occur.

In healthcare markets, the term “monopoly power” does not have to be connected with single monopolistic firm. It also reflects the market where one or a few vendors are dominant (known as oligopoly). This is common when looking at pharmaceutical companies, where the industry is commonly controlled by large vendors in the field, at least for the high proportion of market revenue. Patent protection also contributes to the relative oligopoly pharmaceutical markets. Since the cost of research and production of a medicine is very high, only some large firms with a substantial revenue can invest. The monopoly power also appears in hospital services in areas containing few hospitals. Various health insurance markets, controlled by governmental organizations, are also a valid example of monopoly power. Even in marketplaces with a large number of vendors, such as those for doctor and dental services, the possibility for monopolistic power persists. Some professions are barred from entry due to licensing laws and other forms of regulation. As can be seen, the market failure of monopoly power exists in many areas of the healthcare industry.

Numerous concerns occur in the context of monopolistic power. First, certain entrance restrictions are the consequence of government involvement. Licensing and patent laws are examples of this. Licensure aims to establish minimum quality standards, whereas patent laws aim to encourage new activity in research and production. Second, monopolistic power may be unavoidable in some situations. For instance, in a small market, such as rural areas where only one hospital is needed based on the low demand level, the monopoly power exists. Even when the demand falls, this single hospital may not be functioning if it does not receive sufficient subsidy from the government or reduce its own costs. Third, in some cases, government failure by providing inappropriate suggestions to reduce the monopoly power makes things worse than ignoring it. In particular, some public provision or price restriction regulations may be failed due to the lack of suitability (Folland et al. 2017).

Generally speaking, the involvement of the government in resolving the market failure of monopolistic (or oligopolistic) healthcare market is critical. However, it is easier said than done.

#### Public goods and externalities

Public goods are distinguished by the characteristics of non-rivalry and non-excludability. In other words, public goods include things or services that one person can use or profit from without limiting the consumption or benefit of others. This, however, influences individuals’ willingness to pay since, as long as only one pays, some others have benefits, so who is willing to pay for the good of others? For example, public health upstream primary prevention such as radio education campaigns and disease vector control are considered pure public goods, which can only be provided by the government. Another public good, new scientific knowledge has made a significant contribution to the twenty-first century’s fast advances in health. Its continuing development will be dependent, at least in part, on governments. The optimal degree of provision of any public good, as well as the right choice of interventions, necessitate a comprehensive economic examination in terms of examining the cost-effectiveness ratio. Cost is one part of the analysis; however, cost alone is not enough to determine the value of benefits, especially in the case of public goods, where the benefits need to be considered from a societal perspective. Some non-governmental organizations (NGOs) may provide such recommendations, but evaluation guidelines at governmental level are vital.

An externality is a cost or benefit that a producer incurs or receives but does not pay for. This leads to situations where there is too much or too little supply in the market if the producers are not aware of this type of market failure. As a result, pricing for products and services may not be representative. If we want to achieve efficiency in the presence of externality, the societal perspective is needed, where the social benefits (or losses) are equal to the marginal social costs. With this view, the marginal benefits (or costs) are a combination between marginal private benefits (or costs) and marginal external benefits (or costs). Curing tuberculosis (TB), for example, also stops the illness from spreading. However, a person’s desire to be cured of TB is unlikely to be influenced by the prospective danger to others. If the externality is not considered, treatment will be overpriced in private markets and provided insufficiently.

Positive externalities should not be confused with personal benefits. One of the justified reasons for health subsidies is that the beneficiaries will be better and provide benefits for the society by working more effectively. However, this type of benefit is relatively personal. Therefore, the improvement in working productivity could not be considered an externality of providing healthcare.

#### Information asymmetry

Customers who are aware that they are at greater risk seem to be more motivated to get additional insurance and are therefore more inclined to use it as a result of this realization. This means that insurers have an incentive to identify high-risk customers and either remove them from coverage or charge higher premiums to compensate for their higher level of risk. Taking precautionary steps to gather relevant information about hazards increases the cost of insured healthcare while having minimal impact on overall outcomes.

Health is an unpredictable outcome. Although the risk of one health issue is usually already evaluated by insurance companies before adapting the price strategy, an unexpected event, such as a pandemic, may create an excessive amount of reimbursement. Another issue is that when consumers are insured, they tend to use more services than they really need, which leads to over-consumption problems. This results from the lower financial cost of illnesses that people perceive when they have insurance. This situation not only causes over consumption of treatments but may also create health problems for the insured people as a result of their neglect of health. This concept is known as “moral hazard,” where one is passing on the expenses to others since they are not fully responsible for their behavior. It emerges as a result of ambiguity and the inability of insurers to completely monitor their customers.

The challenges in insurance markets have a direct impact on healthcare markets. Again, the matter of over consumption is recognized, which may create overburden on healthcare capacity. The difficulty of measuring healthcare risks, as well as the difficulty of placing a price on a living body, make it impracticable to determine how much is “too much” in terms of medical and health insurance expenditures. These problems all come from asymmetric information in healthcare.

The government might take on two separate duties. Firstly, the government can help publicly disseminate existing information, either directly or indirectly via private sector activities. Secondly, the pool of knowledge should be developed by actively engaging in scientific research, either directly or indirectly via financing for private sector research.

**Merit goods**

These are the commodities that are deemed useful to someone regardless of the individual’s own preferences.

### Other Rationales

There are many further arguments in favor of government intervention. The federal government plays an important role in stabilizing the economy via the implementation of macroeconomic policies (Folland et al., 2017). These policies may have a massive effect on healthcare programs, assuming their changes in funding, taxation regulations, or even interest rates.

Another distinct area is the government’s involvement in promoting high-quality products consumption. **Merit goods** or products refer to the commodities that are deemed useful to someone regardless of the individual’s own preferences. Some prevention interventions targeting unhealthy behavior, such as alcohol consumption control, tobacco control, and illegal drug control, are linked to the concept of merit goods.

Governments may improve market functioning even further by publishing the information on cost, outcome, and assessing guidelines for healthcare quality. These are not cost-effective needs to be pointed out by identifying essential therapeutic packages. This distinction may therefore have an impact on how to structure commercial or social insurance plans, as well as the appropriate response from the affected persons, such as doctors or patients. It is also significant if the government provides information on the cost-effectiveness of several treatments; this helps insurers and providers adopt the suitable costs for attracting their customers. There is no general rule about how much the public sector should control healthcare delivery, as opposed to funding it. In some situations, for example, in highly populated or impoverished areas, there is a need for government interventions to provide critical health services since it is impractical to leave it to private care. Subsidizing an NGO is an alternative technique of delivering similar services in many regions of the developing world. However, in most cases, the major goal of governmental policy may be to encourage competitions. Competition can be generated between public and private sectors, between private providers, and even between non- and for-profit organizations. Competition generates better choices for consumers while decreasing prices via increased efficiency. In a competitive context, government supply may increase quality or limit prices, while normally, non-competitive public provision is likely to be inefficient or of poor quality.

### Self-Check Questions

1. Please name two typical market failures in the healthcare market that require government intervention to correct.

*monopoly and oligopoly market, public goods, externalities, and information asymmetry*

1. Please fill in the blank.

*Merit goods* are commodities that are regarded as beneficial to someone, independent of the person’s own preferences.

1. Which of the following is not rationale for government intervention in healthcare?

* *making healthcare a perfect competitive market*
* combating poverty
* correction of market failures
* stabilizing the health economy

## 2.2 Forms of Government intervention

Government intervention in healthcare markets can take a number of forms. Regulations may be introduced that address resource distribution, and taxes may be levied, along with or instead of social insurance mechanisms. The last part of this section considers the public provision of healthcare.

### Regulations

Governments impact resource distribution by enacting rules and regulations. In the most severe cases, governments might outright ban particular commodities or activities, such as the manufacturing and use of illicit substances. Governments more typically control the manner in which things are manufactured or consumed. Legislation often appeals to policymakers because it looks to address issues without incurring significant government expenses. However, the legislation imposes considerable financial burdens on others.

The following are the primary roles that regulation may play in the health sector, according to Allsop and Mulcahy (1997):

1. Regulating how a firm can enter and exit

2. Regulating competition in the market

3. Regulating market organization

4. Regulating reimbursement and renumeration

5. Controlling quality and standards

6. Ensuring the safety

Referring back to the role of correction of market failures, the first three roles relate to the correction of monopolistic power in healthcare market. That is, they rectify the particular failures of a monopolistic market structure. Next, regulated reimbursement strategy is responsible for remedying various market failures. It may help incentivize providers by rewarding certain desired qualities, or it may use monopolistic power to keep salaries and healthcare expenses low. Lastly, by controlling standards and quality, as well as maintaining safety, the information asymmetry issues are addressed.

Another model considering three principal questions is what, whom, and how in regulation (Kumaranayake et al., 2000). Accordingly, the author pointed out various variables for regulations, including the entry, quality, quantity, price, distribution, and competitive practices. This adds a quantity variable, which is also an important aspect to consider, even though this might already be reflected in point three of regulating market organization in the previous model by Allsop and Mulcahy (1997). There are many aspects of quantity regulation. Firstly, it may relate to controlling the quantity of health equipment or facility to manage the volume of inputs for a specific need. Secondly, the quantity of organizations also needs to be regulated to improve efficiency in healthcare. For example, in a small area, the number of clinics need to be limited, otherwise it may create an oversupply situation. On the other hand, in a metropolitan area, regulations for increasing the number of providers would help improve efficiency, as well as limit monopolistic power. Further elaboration of this particular model can be found in the literature by McPake et al. (2020).

### Taxes and Social Insurance Mechanism

Private out-of-pocket payments and voluntary health insurance premiums, as well as state sources, such as taxation and social health insurance contributions, can all be used to fund healthcare. As previously noted, no society allows unrestrained market forces to dictate the delivery of all healthcare (although some elements of a system may operate under more or less free market circumstances); the reasons for this have been investigated under the topic of monopoly and oligopoly market failure of healthcare market. Controlling the financing mechanism is one approach for policymakers to exercise direct influence on the volume of funds available for healthcare and the allocation of those resources within the system.

There is opportunity to define what amount of money is to be made available to health and how those resources should be allocated to fit with established goals if the government pools tax income to pay for healthcare services. Priorities might include allocating resources based on need, evenly distributing the financial burden of healthcare payment throughout the community, and reaching adequate levels of preventative health measures, among other things. In contrast, if healthcare is paid for by patients at the point of use (i.e., through the market), the government has no direct control over the resources, which may or may not be distributed in accordance with policy goals. As a result, taxes, social health insurance, and other mandatory payments exist in the financing mix of most healthcare systems. However, the mixture of these components is context-based and dependent on the available resources in different context.

Until the late 1980s, several systems could be classified based on a dominating source of funding; however, in recent years, health systems have been funded by an increasingly complicated mix of resources (McPake et al., 2020). A system that has historically been classified as a social health insurance-based system may include a significant proportion of tax-based support. Similarly, while the United States is commonly regarded as a private-based system with a large number of individuals insured, public taxes and private health insurance account for a sizable share of overall healthcare resources (e.g., to fund Medicaid and Medicare; McPake et al., 2020).

Given this rising tendency in complicated finance systems, evaluating the strengths and limitations of specific funding mechanisms makes sense. It is important to remember that individuals in society are the ultimate source of healthcare funding. Individuals bear the final burden regardless of how resources are gathered (i.e., through taxes or out-of-pocket payments). There are, however, options for who pays, how much, when, and for what. Decisions vary depending on the finance structure, with various implications for how and how effectively a healthcare system operates.

#### The difference between taxes and social health insurance

Public tax and social health insurance have many similarities but also differ in crucial ways that should be highlighted. Some of the differences between these two funding methods are related to practical uses of the mechanisms rather than intrinsic characteristics. It is also worth noting that the language used to explain various funding strategies might be misleading. The term “insurance” is occasionally used even when the health funding is from the general taxation, which can be found in the case of the National Health Service (NHS) in the UK.

**Tax contributions**

Direct revenues (e.g., income tax); indirect revenues (e.g., consumption taxes); and other government revenues (e.g., licences) are all tax contributions (McPake, 2020).

**Tax contributions** come from different sources of revenue. These contributions are pooled by the governments and distributed to different sectors via a financing process, including the healthcare sector. Conversely, **social health insurance** is imposed on the payroll and normally paid by both employees and employers, with contributions for the jobless, retired, and other non-working dependents covered in various ways (e.g., contributions are paid by the government; McPake, 2020). Depending on the context, social insurance funds can be claimed by one or a group of different insurance companies (either public or private). The social insurance will pay for healthcare benefits of those who contribute to it.

**Social health insurance**

Contributions to social health insurance are imposed on the payroll and are normally paid by both employees and employers.

### Public Provision

Governments provide a diverse variety of goods and services, for example, education, roads, water sanitation, fire, and police protection. However, many of these are not considered to be pure public goods. Pure public goods have to be both non-rival and non-excludable. This is also true in the majority of public healthcare, for instance, free vaccinations. Although free vaccination has positive externalities, it is excludable and the limited amount makes it not available for everyone, thus creating rivalry. A common example of public goods is national defense.

In any specific community, the public provision in healthcare should answer the questions of what, how, and for whom appropriately since, with limited resources, it is impossible to provide everything for everyone in the community. First, what types of healthcare should the government provide? Depending on the economies and governance priority, the government may offer different ranges of health services, from free vaccinations for infectious diseases to reimbursed treatments for rare diseases. The “what” question refers to both the quality and the quantity of these services. The next question is how the government provides such services. They may be funded by taxes, or the government may allow the private sectors to provide them via a contract. In other words, the question is whether healthcare should be reimbursed primarily by taxation or social insurance funds (universal coverage), paid by the consumers themselves (private), or a mix of both (most of the cases). Lastly, “to whom” is the critical question that needs to be answered in relation to the redistribution of heath care services. The concepts of equity, fairness, and justice needs to be considered in this case of distribution in healthcare.

### Self-Check Questions

1. According to Allsop and Mulcahy (1997), what are the six primary roles that regulation may play in the health sector?

*regulating how a firm can enter and exit, regulating competition in the market, regulating market organization, regulating reimbursement and renumeration, controlling quality and standards, and ensuring the safety*

1. Please fill in the blank.

Direct revenues (e.g., income tax); indirect revenues (e.g., consumption taxes); and other government revenues (e.g., licences) are all *tax contributions.*

1. Please fill in the blanks.

Vaccinations are not considered a public good since they are both *rival* and *excludable*.

## 2.3 Government Involvement in Healthcare

Every state recognizes that the government is responsible for ensuring the provision of some essential public health services. Only a small number of interventions achieve this goal, even in the instance of vaccination, which is arguably the health initiative that has gained the greatest attention from the government and private donations, as well as having the most clearly verifiable outcomes. Many places continue to be neglected by the healthcare system. Another important component of an effective public health program is a lack of other available resources.

In terms of clinical services, the main government failure in most nations is the endeavor to supply everything to everyone, with no differentiation between more and less important treatment and more or less needy patients. There are certain health services provided by the public sector that are inefficient, regardless of what remedies are implemented, they will not be cost-effective. Due to the widespread and sustained criticism of these inefficiencies, it is now clear that they can only be resolved by fundamental improvement in health organization, such as refocusing government resources away from treatment and toward financing it by encouraging competition among healthcare providers. For these reforms to be successful, a clear distinction between required and discretionary expenditure will need to be made, as well as a renewed governmental commitment on delivering value for money in the delivery of healthcare. There is no equitable alternative strategy to regulating government expenditures that can be used instead of this one. Most governments also do badly when it comes to regulating markets for commercial services, such as insurance. Another problem for which governments are unprepared is the fast rise and almost absolute absence of regulation of private insurance.

In the following, we will discuss the case studies in two highly populated countries, China and India, as the examples of the government involvements in healthcare. The information is collected from Bloom et al.’s 2008 study.

### China: The Transition from a Rural Health System to a Market Economy

When it comes to the healthcare sector in China, many of the government’s old capabilities have been retained (under the Communist Party; Bloom et al. 2008). Pricing is managed in this way, for example, by keeping consultation and hospital day rates low, while allowing hospitals to gain profits from prescription sales and laboratory test costs. As a result, there is greater diagnostic screening and medication usage, resulting in large cost increases for all parties involved. As a consequence of these concerns, patients have a low degree of trust in the healthcare system (Bloom et al., 2008). Accordingly, farmers resisted joining voluntary schemes, and the Ministry of Agriculture opposed the implementation of mandatory rural health insurance. Newspapers have both exposed healthcare professionals’ self-serving conduct and published allegations of physical violence against health workers (Du, 2000).

As a result, in 1970s, there was a weakening in the medical profession’s power in rural China due to the revolution of civil society groups during the Cultural Revolution, which led to the regulatory mechanism in healthcare not being controlled by the appropriate personnel in rural areas. Many non-professionals become employees to provide healthcare at rural health centers, working as township level “barefoot doctors” (Bloom et al., 2008).

#### New regulatory procedure

Since the late 1990s, a new innovative regulatory procedure was developed by the government to fit the healthcare market in China (Fang, 2008). It started with procedures by the Communist Party’s anti-corruption department to decrease the incentives among health professionals to receive informal payments (under the table money) from the patients and bribes from pharmaceutical companies. The reform also conducted the role of improving the quality of front-line providers and developing new rural health insurance systems (Bloom et al., 2008). We will focus on the procedure of improving the quality of front-line providers in the following.

#### Transformation of “barefoot physicians”

In the period from the 1960s to early 1970s, in rural China, the “barefoot physicians” were educated to lead the public health interventions and deliver general primary care at the lowest level of township healthcare facilities (Bloom et al., 2008). In the market reform of the 1990s, these poorly trained professionals either retired or were gradually trained to ensure a basic educational level to provide a certain level of quality of services. The licensing system was developed with the requirement of obtaining annual examinations in most Chinese provinces; those who failed the test would lose their right to practice (Bloom et al., 2008). With the periodic campaigns, the quality of care in the township facilities improved significantly. Most of the clinics, which were run by the unlicensed physicians, were gradually forced to close down or were transferred to the licensed physicians. The government policy also asked the township facilities to have regular meetings among the physicians for a quality check and to monitor their performance. However, it seems that in real practice, this is limited due to the resource constraints (Bloom et al., 2008).

#### The training of village physicians

Similarly, the training of village physicians has also been restructured. In many counties in China, two- or three-year college programs have been set up to prepare village physicians for the profession (Bloom et al., 2008). Because of the disadvantage of distance, the village clinics usually attracted young graduates, especially married couples, whose work is at a similar level to that of village clinics.

Moving forward, the government planned to make greater use of these village physicians by engaging them in public healthcare interventions (Bloom et al., 2008). For example, some village doctors were asked to follow up on patients with hypertension to monitor their risks and healthcare needs.

#### Health insurance

Since 2003, the federal government has provided large economic transfers to disadvantaged communities with the goal of improving access to healthcare (Bloom et al., 2008). In particular, the state’s strong preference for demand-based agreements over budget aid for government health institutions has been a standout feature of these transfer payments thus far. Essentially, it represents the belief that actions must be done to make healthcare institutions more accountable to the general public. The most significant of these arrangements is the New Cooperative Medical Program (NCMS), which is the most widely used county-level volunteer health insurance scheme (Bloom et al., 2008).

Additionally, the Ministry of Civil Affairs is responsible for providing a safety net for very impoverished areas (Bloom et al., 2008). A more common option, however, is for the government to pay qualified families’ NCMS contributions and a portion of their copayments. In 2003, the government announced a plan aimed towards expanding the NCMS (Bloom et al., 2008). Programs were created to facilitate the distribution of public expenditure to support healthcare services. As long as there is an equal level of commitment between the local and federal governments, poorer counties receive a pre-determined amount of beneficiary from the federal governments (Bloom et al., 2008). Households are asked to pay into the schemes, as well. Initially, the plans covered just a small percentage of the cost of inpatient treatment. The new initiative started gradually, with a modest number of experimental projects. The administration quickly expedited implementation, and most counties now have a program in place. Simultaneously, the central government quadrupled its contribution to social welfare initiatives from ten to forty yuan per capita (Bloom et al., 2008).

A number of factors will affect the impact of these initiatives, including how well the managers concerned are able to implement their regulatory roles, how much of an emphasis is placed on them by local government officials, and the extent to which the poor and other beneficiaries’ interests are considered (Bloom et al., 2008).

### India: The Growth of the Private Sector without Proper Regulation

According to the National Sample Survey Organization (2006), private-sector providers in India are quickly expanding their presence not only in inpatient, but also in outpatient sectors. In 2004, 58 percent of rural and 62 percent of urban hospitalizations were covered privately (NSSO, 2004). Private providers are increasingly prevalent in the outpatient care sector, where around 80 percent of patient treatment occurs, usually by untrained practitioners (Bloom et al. 2008).

#### Front-line providers

In India, regulating the front-line provider market is an unclear mechanism, involving the role of a variety of relevant institutions, including (1) the Medical Council of India (MCI), (2) the Departments of Health in different states, (3) some physicians associations (e.g., Indian Medical Association or Indian System of Medicines; Bloom et al., 2008). The MCI is responsible for maintaining standards in medical qualifications, as well as postgraduate education, designing code of conduct for health professionals, and provider registrations. Other institutions have various roles in the health system at different levels from the federal to perspective states; however, none of them can successfully solve India’s fundamental regulatory problems on their own (Bloom et al., 2008).

In India’s healthcare system, quality of treatment provided by front-line clinicians is a critical concern due to inappropriate regulation. As a result, the inpatient care sector saw unregulated development in private providers, resulting in facility duplication in metropolitan areas. The problems of different levels of quality, the presence of non-licensed practitioners, and corruption are also presented (Radwan, 2005). This creates problems specifically regarding the outpatient care, even for treating a minor condition since they have not obtained the appropriate license or knowledge to be able to do that. The problem is even worse since many patients themselves are unaware of the lack of qualifications among their practitioners (Mukherjee et al., 2007).

Additionally, it seems that there is no proper regulation to control the quality of private care. There might be a clinical code of conduct that is provided by the MCI; however, no consequences are specified for misconduct. Even in the face of substantial complaints, the institution is rendered helpless as a result of this (Bloom et al., 2008). Associations of inexperienced providers, on the other hand, have proven to be ineffective self-regulatory organizations. Consequently, the Consumer Protection Act (COPRA) may be the only way for harmed consumers to protect their rights in certain instances (Verma, 2002). Medical malpractice is defined as a criminal act under the COPRA, which was established in 1986 with the aim of safeguarding the rights and interests of patients (and consumers more generally; Verma, 2002).

#### Inequity in the costs of health services

The current regulatory structure does nothing to assure the right value for money. This leads to the huge difference of charging for the same services, for example, in the case of cesarean section, which costs between 3,000 and 50,000 Rupees (Lal, 2005). There is no pricing regulation to manage the right level of payment for a particular health service in India. This is both inequitable and inefficient in the view of an economist.

When it comes to equity, the concern centralizes around the subject of access equality as well as the quality of care for communities with lower socioeconomic status. Data indicate that the cost of private, high-quality medical treatment in the front-line sector is often prohibitively high. As a result, poor consumers are compelled to obtain outpatient health services from unqualified providers and inpatient treatment at poorly-funded public hospitals (Bloom et al., 2008). Because public hospitals are underperforming in many places, a large number of people living below the poverty line (about 44 percent) are forced to seek care at private hospitals, where they must carry a significant financial burden for the duration of their stay (NSSO, 2004).

So far, the government has little focus on regulation to achieve equity in healthcare distribution. By means of equity in distribution, state governments should offer relevant strategies and provide subsidies for healthcare to the poor and most vulnerable groups.

#### Health insurance

Another critical concern in India is the regulation of the health insurance industry, owing to the absence of financial health protection available to the majority of the population. Although there is a mix between public and private insurance, with a variety of providers ranging from mandated social health insurance to private voluntary scheme to community-based plans, the market is still small, as only around three to five percent of Indians are insured (Bloom et al., 2008).

India’s insurance market was opened in 1999, and the Insurance Regulatory and Development Authority (IRDA) Act was passed, which formed a regulatory authority and encouraged the participation of the private sector (Bloom et al., 2008). Prior to passing this act, the government had covered all healthcare provisions. There was no mention of health insurance in the IRDA, but it was included in the non-life insurance section. Politicians don’t seem to understand how the healthcare industry is changing and how this affects patients (Bloom et al., 2008).

### Self-Check Questions

1. When did the market economy reform happen in China?

*late 1990s*

1. Please explain two ways in which the Chinese government improve the quality of health service of village physicians?

*Most Chinese provinces have instituted a licensing system in which rural physicians must pass yearly exams. Those who fail these exams may lose their license to practice. Local health agencies conduct efforts to shut down clinics run by unauthorized staff. Many township hospitals hold frequent meetings with village physicians, and doctors from township institutions may visit villages to check their performance.*

1. Which authority permitted the private sector to enter the Indian insurance market and when?

*the Insurance Regulatory and Development Authority (IRDA) Act of 1999*

## 2.4 Government Failure

In some cases, government interventions may be sound in the view of an economist, but government failures can still occur due to specific contextual characteristics.

Firstly, the effectiveness of an action may be overestimated. Governments only have a limited amount of influence over how private entities respond and have the potential to undermine the desired objective. For instance, physicians’ fees have been determined by discussion with provincial governments in all provinces of Canada since 1971, and rates are no longer growing faster than the general price level (World Bank, 1993). Physicians performed a larger number of procedures to safeguard their salaries, notably during the inflationary period between 1971 and 1975 (World Bank, 1993). This reaction was most prominent in areas where actual fees had fallen the most. As a result, the savings in government spending were substantially lower than expected.

Secondly, governments may lack the competence to administer and implement programs effectively. Indeed, they may be victims of corruption, as well as incompetence (World Bank, 1993). This problem is exemplified by two donor-funded public hospitals in two Latin American countries, each having 500 to 600 beds. One was just too big to manage and run, and hence could not be used at greater than 60 percent capacity. The other was so poorly built that it could only hold one third of the anticipated number of patients (World Bank, 1993).

Thirdly, the healthcare system is controlled by special interests, both inside and outside the healthcare market itself (World Bank, 1993). By providing financial assistance for the training of redundant physicians, covering the costs of low-value discretionary treatments for wealthier patients, and ensuring the survival of local businesses, governments contribute to the development of incentives that restrict sound policymaking. When public action fails to overcome the opposition of those who would suffer as a result of it, even when society as a whole would benefit, it is referred to as a failure of the public action strategy (World Bank, 1993).

The critical question is how decisions should be made in the healthcare market. It seems that most government policies have not been specific enough in healthcare delivery due to thinking only about providing healthcare to as many people as possible; this creates the funding problem. Considering the resource constraints, governments should prioritize what to fund, include, and exclude. Health economics should play a critical role here, as it helps to evaluate the proper strategies from the government. Moreover, the government should be aware of the behavior of insurers, providers, and patients when making decisions about healthcare. Before making any decisions, the advantages and disadvantages of each alternative should be taken into consideration. For example, in the field of health insurance, the government should determine how much support they should provide for social insurance (as an insurer themselves), or how much they should allow the private sector to jump in. This raises a further question of how to govern the private sector if there is an incentive to promote competition in the healthcare market. In many cases, the government have to consider the trade-off between the societal benefits as a whole and individual care. To sum up, it is easier said than done; although the strategies seem to be sound, market failure is always a risk. It should be noted that since every healthcare system is unique, government involvement should be adapted for suitability.

### Self-Check Questions

1. Please provide two reasons for government failure in healthcare, according to the World Bank (1993).

*Governments may overestimate the effectiveness of an action. Governments may lack the competence to administer and implement programs effectively. Governments’ interest in healthcare lies both inside and outside the optimization of the healthcare system (e.g., satisfying voters).*

## 2.5 Competitive Strategies

Given the market failures in the healthcare market, in many countries, there are incentives to encourage competition within the health systems in order to achieve better efficiency. Competition may arise among healthcare providers, private and public health sectors, or health insurers. In this section, we will start by introducing the rationale of competition in general economics, in particular the healthcare market. A case study of healthcare in the Netherlands will be presented as an example of healthcare reform by applying the regulated competition strategies.

### The Rationale of Competition

In an industry, companies need to compete against each other to maximize their profits. The nature of the perfect market would let the market itself function with the competition among the firms to deliver the market efficiency. In order to compete with others, a company is expected to develop and maintain their competitive advantage. With the competitive advantage, a company can distinguish themselves from their competitors and then make a better profit, by either (1) lowering the costs of production or (2) delivering differentiated products into the market.

If a company can maintain lower costs than its competitors, more profits can be generated in two ways:

* 1. Selling the same quantity of products as their competitors, with the lower costs of production
  2. Selling a higher quantity of products by lowering the market price of that product to become dominant distributor in the market

There are some common strategies that can be used to lower the costs of production, including mass production and applying new technologies.

A company can produce unique products that distinguish them from their competitors. This allows them to decide on the price; either they choose to have a higher price and sell at the same quantity as their competitors, or keep a similar price to their competitors and sell a higher quantity thanks to the uniqueness of the products. In both ways, their profit will increase.

In a dynamic industry, competitive strategy can be created in different ways, whether as a result of the firm’s direct operations (proactive planning) or as a result of the firm’s response to external changes (reactive strategies). Competition responds to the establishment of a competitive advantage by striving to neutralize or eliminate the advantage. In order to do this, one may either imitate those with a competitive advantage or behave in a way that results in the development of a brand-new advantage.

As a result, we should integrate the protection of competitive advantage as an indispensable part of a competitive strategy. It is at this point that we get to the crucial sustainability concept of competitive advantage. The better the preservation of an advantage, the longer the period the holder can use in order to achieve a higher profit based on the hold of that advantage. Once an advantage has been created, there is also a possibility for the holder to control the time that their competitors need to produce a similar or new advantage of their own. The goal in this situation is to maintain a competitive advantage. Let us take the example of the car market, where a company creates and sustains their competitive advantage in producing and using electric cars in the global market. Although they did not invent the electric car or even the luxury electric car, the advantage comes from the successful business model with a vision of bringing compelling electric cars to the market (Zucchi, 2021). To maintain their competitive advantage, one strategy they apply is building up a network of charging stations, which is the largest barrier to adopting the electric vehicles into the market (Zucchi, 2021).

In summary, there are two main elements of a competitive strategy: how to establish the advantage and how to maintain that advantage. In a company, in order to create any competitive advantage, available resources need to be utilized. Resources may range from human, finances, facilities, or the organization of the particular industry.

#### Competition in the healthcare market

Traditionally, in healthcare, competition appears when considering prices, quality of services, advanced technology, or the combination of two or more of these measures. Similar to other markets, competition in healthcare also has the incentive to decrease costs and improve efficiency in health services. As a result, the competitive strategies would help to deliver a lower price to the end-market users (i.e., patients).

There are a variety of characteristics that contribute to positive-sum competition in healthcare, which are explained by Porter and Teisberg (2004). These characteristics vary in different healthcare and prevention levels, from prevention to diagnosis to primary and tertiary treatments. However, although this concept of positive-sum competition contains various potential benefits, the complicated needs of overall adjustments make it difficult to be applied.

Rivers and Glover (2008) pointed out three essential parties contributing to healthcare competition:

1. Individuals offering healthcare
2. Organizations providing healthcare services
3. Organizations offering healthcare finance, plans, and insurance

Firstly, there might be competition among healthcare professionals (e.g., physicians) who can offer healthcare for patients and get the value back by direct out-of-pocket payments or reimbursement from third-party insurers (Rivers & Glover, 2008). Many competitive strategies can be conducted, such as focusing on the quality of patient-centered care to maintain the relationship with the patients and benefit from their recommendations. Physicians may also compete for a place on the ranking list of a preferred provider organization (PPO) affiliation to attract the patients. Another strategy is to perform a market analysis to obtain the demand in a particular region and provide services particularly for that region.

Secondly, the organizations that provide healthcare services (e.g., clinics and hospitals) are also a source of competition in the healthcare market (Rivers & Glover, 2008). Normally, hospitals will compete with each other for qualified physicians, patients, and contracts with third-party payers. The competition for physicians mostly happened in the past, when hospitals attracted qualified physicians with advanced technology, a supportive working environment, or further educational opportunities. A successful business would thoroughly plan the recruitment strategies ahead of time by considering its staff’s age distribution. With the development in health education, the competition for physicians has become less dominant (at least in countries with an oversupply). In recent years, hospitals are more likely to compete for patients by providing better health services at lower rates (Rivers & Glover, 2008). These are only examples of competitive strategies; however, it really depends on the context and health systems when determining which should be used.

Thirdly, competition occurs among the organizations that offer healthcare finance, insurance, and plans (e.g., health management organizations [HMO], preferred provider organizations [PPO], and different insurance firms; Rivers & Glover, 2008). The competition among these organizations is less in the publicly-funded health systems since the government has the monopoly power, or a limited number of dominated insurance firms have oligopoly power. Considering the information asymmetry between insured and insurers, it is likely that the customers would be aware of available packages that they would prefer. The competition in the health insurance market should research factors such as access, price, benefits, and quality before making competitive strategies. In addition, it is important to notice that, because of the complexity of the plans offered by insurers on selection or benefit or coverage services, it is difficult for patients to compare (Rivers & Glover, 2008).

### Case Study: Regulated Competition in the Netherlands

After World War II, the Netherlands developed a mixed health system combining the private healthcare delivery with state-funded healthcare and controlled a variety of health insurance plans (Harrison, 2004). The concept of solidarity has been spread out to all the political parties, together supporting the universal and equal access in healthcare regardless of financial ability. Although this brings a certain success to the Netherlands’ health system in terms of cost containment, during the early 1980s, the health expenditure was significantly rising due to recession (OECD, 2001). This lead to tighter and more direct control of the state over insurance payments, as well as provider expenditures. A movement emerged in the late 1980s that advocated for less state and corporatist engagement in health finance while preserving governmental pledges to solidarity, supported by representatives from major national groups and government health officers (Harrison, 2004). As a result, the government introduced regulated competition among the insurers and services providers in the late 1980s in the Netherlands (Harrison, 2004). This section will discuss the reforming process in the Netherlands used to adapt the regulated competition strategy.

#### Background: Prior regulation

After World War II, the Netherlands’ competition policy and legislation underwent significant changes. However, despite the fact that the competition legislation was approved in 1958, the infamous failure to enforce its purported bans on pricing agreements, market sharing agreements, and collusive tenders contributed to the increasingly widespread perception of the Netherlands as a country that was tolerant of cartel activity. According to a 1992 report, the Netherlands was responsible for almost 40 percent of the significant cartel proceedings brought by the European Commission (de Jong, 1992).

There were over 245 agreements to partition markets and almost 270 agreements to regulate prices on the government’s cartel record, which was kept secret. Furthermore, the registry listed almost 50 agreements to exclusive trade and more than 200 pacts to stifle competition in distribution (OECD, 1993).

With inadequate enforcement and widespread private agreements, as well as significant use of public and private regulation to regulate input costs, the intensity of market competition is deemed to be relatively low in many industries, particularly those shielded from import competition. Meanwhile, the European Commission has brought several actions against market-wide bid manipulation and exclusive partnership arrangements that impede competitions in Dutch companies (OECD 1999). The excessive control resulted in a high percentage of healthcare expenditure during this period, at an average of 17 percent between 1972 and 1982 (Harrison, 1995).

In the early 1980s, direct government interventions had a substantial impact on reducing increases in healthcare expenditures by a large margin (OECD, 2001). Consequently, total health expenditures as a percentage of gross domestic product (GDP) started to decline. Despite the fact that hospitals were more resistive to rules than other aspects of the healthcare system, general hospitals achieved great strides in improving outpatient treatment, cutting inpatient expenditures, and shortening inpatient stay lengths throughout this period (OECD, 2001). According to Harrison (2004), these developments in hospitals were primarily driven by three factors: government-imposed hospital budget caps, new technologies that enabled less invasive forms of diagnosis and surgery, and government policies and budget incentives for hospital mergers and reduced bed capacities in general hospitals (long-term care facilities were excluded). However, expenses for hospital personnel were exempt from government oversight in most cases. Because specialists were paid on a fee-for-service basis by insurers, their activities were not included in hospital budgets and were not controlled by budgetary constraints (Harrison, 2004).

The increase in the quantity of specialized treatment contributed to driving up the hospital and overall health expenditures in the second half of 1980s, despite the cost-controlled regulation from the government (Harrison, 2004). The repetition of high healthcare expenditures over multiple years can be observed as the disadvantages in regulation strategies at that time by many politicians and economists. Furthermore, the issue, combined with the long-term conflicts between the government and health providers (e.g., medical specialists), resulted in protests against the existing regulatory procedure (Harrison, 2004).

#### Proposal for regulated competition

Deriving from the dissatisfaction with the existing regulatory mechanism in the healthcare sector, the insurance system reform to encourage competition among statutory insurers was suggested in the late 1980s by the Dekker Commission (van de Ven, 1997). This reform was built on the political ideology in the late 1970s and early 1980s in the Netherlands, where there were criticisms of the social welfare system after the shift in political party (Harrison, 2004). Despite the largest economic downward situation in the whole of Europe after the World War II, the advocate moved towards the reform in the health system during this time. It is believed that by encouraging competition in healthcare, government expenditure on healthcare can be significantly reduced, and the private market can help to restore the nation’s economy and improve its competitiveness globally (Visser, 1990). This view is supported by the present members of the ruling center-right coalitions, several economists, an Advisory Commission on Industrial Policy, and some provider groups (Harrison, 2004). As a result, significant privatization of governmental services occurred. Members of the opposition Social Democratic Party also campaigned for a “new realism” in areas such as social benefits and wage policy. For example, in 1983, the Dutch trade union confederation agreed to cease salary indexation in exchange for employers’ promise to create job opportunities. Wim Kok, who was elected Prime Minister in 1994, led the confederation in taking this action (Harrison, 2004).

The committee expected that competition among insurers and providers would serve as a powerful incentive for both patients and providers to use resources efficiently, to substitute hospital care for less expensive outpatient and community care, and to improve healthcare coordination (Harrison, 2004). A more market-oriented approach was also supposed to improve patient choice and deliver treatment suited to the individual’s needs. To guarantee unity and justice, as well as ensure that providers (notably physicians) accepted responsibility for ensuring the quality of treatment, regulations would be required (van de Ven, 1987).

To make this a reality, the Dekker Committee proposed establishing a single mandatory insurance system that would cover around 85 percent of overall health and social care expenditures, as well as integrating the following distinct financing streams (Ministry of Welfare, Health, and Cultural Affairs, 1988):

1. Social care is funded by the government (e.g., aging care and care for people who are disabled).

2. Long-term care, including nursing homes, and care for people who are disabled or mentally ill, is provided under the Exceptional Medical Expenses Act (AWBZ) by the compulsory insurance scheme.

3. Compulsory sickness fund insurance for low-income employees (at a defined income level) is implemented.

4. Civil servants and other public employees are provided with insurance.

5. For the self-employed and with income higher than a certain level, optional private insurance can be chosen.

Public sickness funds, as well as for-profit insurers, are obliged to offer the basic package of statutory coverage under the new mandatory regulation. In addition to that, both of them could also offer an optional supplementary package to cover services that are not mandatory, for example, dental, optic care, or over-the-counter medications (Harrison, 2004).

These regulations boost the competition among the insurers on both the quality and price of services. Thanks to that, the patients can choose the most suitable insurers for themselves. This competition would contribute the potential for efficiency and innovation in health insurance market (Ministry of Welfare, Health, and Cultural Affairs, 1988). The proposals of the Dekker Committee were finally adopted by the administration and approved by Parliament in 1988 (Ministry of Welfare, Health, and Cultural Affairs, 1988). Following the 1989 elections, the newly formed alliance of Social Democrats and Christian Democrats carried on the reform agenda (Harrison, 2004). In that regard, the government implemented similar strategies that were recommended by the Dekker Committee , but emphasized compulsory insurance. For example, the AWBZ program has been implemented, covering 96 percent of overall expenses, which is higher than the recommendation from the Dekker Committee (Harrison, 2004).

Further interesting reads about the healthcare market reform in other countries, such as Sweden or the United Kingdom, can be found in the literature by Harrison (2004).

### Self-Check Questions

1. Please name the overall benefit of having a competitive advantage.

*Competitive advantage helps a firm to make a higher profit.*

1. Please name two elements to assess a competitive strategy of a firm.

*the establishment of a competitive advantage*

*the protection of that competitive advantage*

1. Please fill in the blanks.

In the Netherlands, the Dekker Commission proposed establishing a single mandatory insurance system that would cover around *85 percent* of overall health and social care expenditures. After that, the AWBZ program was implemented, covering *96 percent* of overall expenses, which was higher than the recommendation from the Dekker Commission.

Summary

Two main economic rationales for government interventions in healthcare are the roles in combating poverty and the roles in the correction of market failures. If the regulations address the effects of market failures successfully, there should be an improvement in efficiency. Given many market failures existing in healthcare markets, all countries around the world have their own regulations to deal with the problems. These regulations are unique and contextual.

A variety of interventions in healthcare market are conducted by the government. They have the essential role in controlling market entrance and exit, containing competitive practices, controlling market organization, controlling remuneration, regulating quality and standards, and ensuring safety. Additionally, the government provides funding for healthcare through taxes and social insurance mechanisms. Controlling the financing mechanism is one approach for policymakers to exercise direct influence on the volume of funds available for healthcare and the allocation of those resources within the system. Moreover, the government can provide public provision over healthcare. To combat main economic concerns regarding what to deliver, how to deliver it, and for whom should be considered critical in the public provision of healthcare.

The government involvement in healthcare was discussed by using case studies in the two most populated countries: China and India. It is noticed that every government has their own strategies to tackle the level of market failures in healthcare. The main government failure in most nations is the endeavor to supply everything to everyone, with no differentiation between more and less important treatment and more or less needy patients.

The rationale of competition in healthcare is also an important activity in healthcare to create the incentive to improve efficiency. A case study in the Netherlands was taken as an example of regulated competition in healthcare controlled by the government.

# Unit 3 – Equality and Fairness

**Study Goals**

On completion of this unit, you will be able to …

… understand the concept of equity in health.

… analyze different theories of distributive justice.

… identify social determinants of health.

… apply the concept of equity to policy actions.

# 3. Equality and Fairness

## Introduction

The healthcare system is complex and possibly contains some market failures, which require correction by the government. Distribution of healthcare is one of the most important actions that the government should pay attention to in the healthcare sector. This unit will focus on concepts in healthcare distributions and their theoretical and philosophical foundations. We will start to look at the rationale behind redistribution of healthcare via the distributive preferences, in which the concepts of welfarism and extra-welfarism will be explained. Then, we will discuss the concept of health equity and its differentiation from the equality concept. Equity in healthcare will also be addressed in this unit.

The link between equity, justice, and fairness will be discussed, and the theories of distributive justice will be linked to the concept of health equity. Influential theories will be addressed, along with developments and criticisms, including utilitarianism, liberalism, and egalitarianism. In addition, Amartya Sen’s idea of functioning and capabilities will also be presented.

The social determinants of health will be explained and a case study of Denmark will illustrate the importance of this concept. The unit will be concluded with policy lessons drawn from theoretical and philosophical backgrounds, with a particular focus on the recommendations from Whitehead (1992) and the conceptual framework for tackling inequities by addressing social determinants of health, proposed by Solar and Irwin (2010).

## 3.1 Distributive Preferences

As healthcare is a complex system, distribution in healthcare is also complicated. In this section, we will discuss the preferences of distribution. Firstly, the political ideologies underlining redistribution will be raised for discussion. Then, externality is given as a rationale for redistribution, which is also explained. Different theories behind distribution will be covered, including the concept of Pareto efficiency and its relation to welfarism and extra-welfarism in health economics. These principles are essential to the field of health economics, especially health economic evaluation.

### Ideologies of Redistribution

Why is there a need for redistribution? One possible ideology is that the poor have the political power of majority in the society, and as a result, the rich are obliged to redistribute the wealth to the poor through political processes. Another ideology assumes redistribution to be insurance for the ruling party, which reduces the likelihood of revolution, as it can be used to bribe the working class with things like a minimum wage, pension benefits, and accessible education and healthcare (Olsen, 2017). These interpretations seem to be cynical, considering redistribution resulted from the relationship with benefits between the rich and the poor in society; either the poor are seen as ripping off the rich, or the rich are seen as offering the bare minimum in exchange for abusing the poor. These two ideologies do not include those who genuinely care about each other.

The third hypothesis, which we will examine in greater detail, is founded on the belief that individuals care about other members of society. Voluntary redistribution between the rich and the poor in a society might result from altruism and can be explained using the framework of independent utilities functions (Olsen, 2017). Assuming a redistribution occurs in a two-person economy with a rich person (R), and a poor person (P), voluntary redistribution will continue until R’s marginal utility gain from P’s higher income equals R’s marginal utility loss from their own lower income (Olsen, 2017).

However, it seems to be less optimum if redistribution only comes voluntarily. When the model of interdependence between two individuals is extended to many individuals in a society, it may lead to the “free riding” phenomenon. To elaborate, in a group of rich people, one may wait and observe other individuals contributing to the poor. When the contribution is significant from many of the rich, the marginal benefit of contribution from one extra individual (the observer) would be tiny. At the end, the observer might not even contribute, which means that this rich individual takes a free ride on the contributions of others.

To prevent the free-riding phenomenon, it may be reasonable for the rich to vote for mandatory redistributive taxation in society. This explains why rich people may choose “voluntary compulsion” and support political parties whose taxation would restrict their own private consumption. This is an argument that only makes sense when the individual is distinguished as a citizen and a consumer (Olsen, 2017).

### Externalities as a Rationale

An externality is a cost or benefit from one individual’s action that impacts other individual(s) who did not intend to incur that cost or benefit. Externalities may be beneficial or detrimental, and they can result from the production or use of a product or service. Negative or positive effects may be personal (to an individual or an organization) or societal (for the entire community). A popular example in healthcare is herd immunity, resulting from immunization campaigns against infectious diseases. People outside the market transaction (those who are not already vaccinated, in this case) are less likely to be infected since the immunized individuals are less likely to become a carrier of the illness. Depending on the disease characteristics, relatively, a certain percentage of immunized people in a society would result in herd immunity. This is an example of a beneficial consumption externality.

A beneficial externality, however, is also a market failure in the health market, which creates inefficiency in healthcare production. In a fully competitive market, individuals will consume the goods until the price reaches equilibrium point, when their marginal benefits meet their marginal costs at . The figure below shows an efficient level of consumption reaching the maximum output for an individual at, when there is the absence of externalities.

In the existence of beneficial externalities, a positive marginal external benefit (MEB) will be present. As a result, the marginal social benefit (MSB) should be calculated by adding the MEB and individuals’ marginal private benefit (MPB). The figure below represents this concept. When considering the MSB, the optimal output in a societal perspective is achieved at point , which is higher than the maximum private output . At this point, the price is also higher, at point . Thus, society may be justified in subsidizing vaccines (or anything to increase the beneficial externalities) on the basis of efficiency. The subsidy amount to reduce the price from to can be seen in the figure below. It is noticed that the social benefits may happen at a certain quantity; therefore, the MEB curve starts at a higher quantity than the MPB curve. A similar pattern applies to MSB since it is the sum of benefits between the MPB and MEB curves. If negative externalities exist, the marginal external loss may exceed the marginal external benefit, resulting in a lower level of output.

Efficiency in society with the presence of a beneficial externality

Shape

Description automatically generated

Source: Duy Pham (2022).

As discussed, immunization is one rationale for subsidies for healthcare based on the role of beneficial consumption externalities. In a larger dimension in the healthcare market, such as a social insurance system, we need to justify using a similar but different concept of externalities, which is also called “altruistic externality” (Pauly, 1971) or “charitable externality” (Folland et al., 2017). Accordingly, whenever individuals believe that some group of society is receiving insufficient care to the point where a kind person would be ready to pay to assist these people obtain care, the willingness of the rich to help the poor obtain basic healthcare comes directly from the kindness. Such generous impulses are most likely prevalent in most communities. Again, this is the concept of altruism.

Consider the curve MEB in the previous figure to quantify this externality. The efficient level of production is then higher than the fully competitive market level of output, . Under some conditions, this efficiency may be sufficient rationale to act, for example, by providing a social insurance scheme or simply enacting redistribution in healthcare.

### The Pareto Efficiency Concept

According to the Pareto principle, social welfare can only grow unambiguously if the welfare of any member of society rises while the welfare of no one declines. When it reaches its optimal level, the point of **Pareto efficiency** is when it is impossible to continue doing this without negatively affecting someone else in society.

**Pareto efficiency**

This concept is achieved when it is impossible to benefit one individual without reducing the benefit or utility of another in a society.

To elaborate on Pareto efficiency, we can apply the Edgeworth Box by hypothesizing a scenario of two people, A and B, who are trapped on an island with no other human beings. There are only two goods available on the islands, and both are essential for their livings: food and fresh water. They have to distribute these goods among themselves to be able to maximize their utility (efficiency) and distribute fairly between themselves (fairness) to attain the maximum amount of survival time on the island until they are rescued. In this hypothesis box, we can only address the first point on efficiency, not the second point on fairness.

**Indifference curve**

Points on the indifference curve represent the combination of goods yielding the same consumer’s utility

In the figure below, the quantity of food in the market is represented by the vertical axis, while the total amount of water is measured by its horizontal axis. In the bottom-left corner (at point A), individual A has nothing, while B consumes all of the available food and water. In the opposite top-right corner (at Point B), individual A possesses everything in the economy while B has nothing. The **indifference curves** A1, A2, and A3 represent individual A’s preferences related to food and water for a certain utility level, with the notion that the utility at the A3 curve is higher than the A1 curve. Indifference curves B1, B2, and B3 represent individual B’s preferences, with higher utility on the B3 curve than the B1 curve. Looking closer at three scenarios of allocation (points U1, U2, and U3) where the food and water is allocated to the two individuals, at point U3, A is on their indifference curve A3, whereas B is on their indifference curve B1. This utility is not Pareto efficiency, since the utility of A or B can be improved without harming each other. This statement is illustrated when considering the indifference curve B\*1 and point U\*3. In this case, at point U\*3, while individual A maintains their utility level at curve A3, the utility of individual B has been improved from curve B1 to curve B\*1. Therefore, the allocation at point U\*3 achieves a better total utility for two individuals. Point U\*3 and point U1 (which can be explained similarly), are Pareto-efficient allocations of the available goods. However, it is critical to remember that Pareto-efficiency does not mean that the goods are allocated fairly. Assuming that both individuals had similar characteristics, fair allocation should lie somewhere in the middle of the box, for example, point U2. The fairness distribution is debated in different ways and sometimes, it happens that the distribution is fair but not Pareto-efficient.

The Edgeworth Box for a two-good and two-individual economy

Diagram

Description automatically generated

Source: Duy Pham (2022).

### Welfarism versus Extra-Welfarism

In health economics, welfarism and extra-welfarism are important concepts to keep in mind. Although there is a long history of debate, there are no clear definitions to distinguish between these concepts in literature (Coast, 2017). In the following, we will start to discuss welfarism and extra-welfarism based on existing literature and the related concepts in health economics. The theory behind them will be explored using influential research in the field and interrelate these concepts to economic principle, for example, Pareto efficiency. The discussion will analyze the key differences between these two concepts and their applications.

#### Welfarism and welfarist economics

According to Sen (1977), welfarist economics can be categorized into two traditions: classical and neo-classical. “Classical” welfarist economics, also known as the original welfarism, is under the classification of utilitarianism first promoted by the English philosopher, Jeremy Bentham in the late eighteenth century. Early utilitarian theory proposed the “greatest happiness of the greatest number,” i.e., the sum of all individual utilities, as the goal of any society. As a result, the optimum level of utility can be reached when this sum is maximized. This classical interpretation of welfarism has been updated to the neo-classical welfarist economics, including two further subdivisions (Brouwer et al., 2008). The Paretian tradition accounts for social interrelation by adding one more component to the classical one by implying the Pareto principle in the allocation for social optimum level of utility. Another subdivision is the Bergson-Samuel social welfare function, accounting for preferred distribution on a certain welfare frontier, that may not satisfy the Pareto efficiency. Some economists also consider non-utility information (e.g., Sen’s capability approach) or the characteristics of the originators and the receivers of utility (Brouwer et al., 2008).

Welfarism, according to Sen (1993) and Hurley (1998), can be interpreted narrowly so that the optimal goodness only accounts for the level of utility attained by individuals. This narrow definition lies on the fourth tenet of a dominant framework of welfarist economics, built on explicit normative principles (Hurley, 2000). This framework includes four key tenets:

1. Individuals maximize welfare by selecting the preferred option over the list of ordered alternatives (the utility principle).
2. Only individuals can judge what is best for their utility and how much it contributes for them (individual sovereignty).
3. The behavior and processes together derive value for utility, rather than each of them individually (consequentialism).
4. The goodness of any situation is based solely on the level of utility achieved by individuals in those circumstances (welfarism).

To summarize, the basic feature of welfarist economics is that it limits the evaluative space to individual utility maximization. This may drastically constrain the assessment, particularly when utility comparisons between people are deemed unattainable (Brouwer et al., 2008). As we shall show below, extra-welfarism broadens the evaluative space by allowing for consideration of factors other than individual value.

#### Extra-welfarism

Extra-welfarism is not a new concept; it has many historical roots. A concept appeared in the public finance sector named “merit good,” which is one example root of extra-welfarism. Musgrave (1959) defined it as the meritorious good that should be subsidized from the society, but not purely a public good. This concept is quite unclear on what should be considered meritorious and how it is different from those goods that create externalities for the society (Culyer, 1971). The second root considered health to be “basic” (Tobin, 1970) or “primary” goods (Rawls, 1971) that ought to be appropriately allocated in the society. An influential work by Sen (1980) maintained that the preoccupation of welfarism on individual utility was too limited and should be captured more broadly by considering capabilities of different groups in a society.

Based on these roots, Culyer (1991) described extra-welfarism as a vital “class of ‘extra’ welfare sources […] the non-goods characteristics of individuals […] Extra-welfarism thus transcends traditional welfare: it does not exclude individual welfares from the judgements about the social state, but it does supplement them with other aspects of individuals” (p. 67).

This concept does not mean to ignore the importance of individual utility, but it strongly states that individual utility should not be considered the only outcome of interest. This statement is the critical distinction between welfarism and extra-welfarism. So, what are the other outcomes besides utility?

The extra-welfarism stream adds extra components to utility-based welfarism. These add-ons are referred to as non-utility information including, for instance, capabilities, the quality of utility, equity weights, and other characteristics (Brouwer et al., 2008). In healthcare, rather than the individual preferences in health, many health economists believe that health should be treated as a multi-attributable value, accounting for both the utility and non-utility information. Health plays an essential component in human beings and should be valuable by itself.

In welfarist economics, comparisons between individuals in the sphere of utility might not be meaningful in some cases. For example, an optimistic individual may rate their utility higher than a pessimistic individual, even though they have a worse health condition. Extra-welfarist methods for public decision-making, on the other hand, tend to be based on direct interpersonal comparisons. Even if it’s impossible to tell whether a sick person has less utility than a healthy individual, it is clear that a sick person has worse health (Brouwer et al., 2008). Quantitative measures yield results comparing individuals and allow for testing and monitoring of programs aimed at improving health or reducing health inequities, even though some of them may be imperfect. The recent focused measure in health is quality-adjusted life years (QALYs). This measurement is a widely discussed topic, which is out of the scope of this section. However, it is critical to keep in mind that extra-welfarism can be regarded as the grounded principle of QALYs.

### Self-Check Questions

1. Please fill in the blank.

The concept of *Pareto efficiency* is achieved when it is impossible to benefit one individual without reducing benefit or utility of another in a society.

1. Please define the indifferent curve.

*The combination of goods on an indifference curve provides the same level of utility to the consumer.*

1. Please name one of the four tenets of a dominant framework of welfarist economics, according to Hurley (2000).

*Individuals maximize welfare by selecting the preferred option over the list of ordered alternatives (the utility principle); only individuals can judge what is best for their utility and how much it contributes for them (individual sovereignty); the behavior and processes together derive value for utility, rather than each of them individually (consequentialism); and the goodness of any situation is based solely on the level of utility achieved by individuals in those circumstances (welfarism).*

## 3.2 Concepts of Health Equity

We may hear the term “health equity” many times, especially from the target of universal health coverage nominated by the United Nations (UN). This section will start with a detailed discussion on this concept and explain how it is different from others, for example, equality. Then, determinants of health inequities will be explained, followed by the definition of health equity from the World Health Organization (WHO). Different categories of health equity will also be discussed, and this section will be concluded with an introduction to equity in healthcare.

### The Meaning of Equity in Healthcare

The concepts of equality and equity are not the same. The figure below illustrates the difference between these two concepts.

Equality and equity

A picture containing text

Description automatically generated

Source: Maguire (2016). CC BY-SA 4.0.

On the left part of the picture, everyone is provided an equal amount of good (a box, in this case). The two taller people are able to watch the baseball match, whereas the shortest person cannot see the match. This illustrates the concept of equality, which is the equal distribution in quantity of goods among the receivers. On the other side of the figure, the concept of equity is illustrated. Here, goods are distributed based on the need, so now everyone can see the match regardless of their height.

Inequality is also different from inequity. It is normally understood that inequality is an umbrella term under which inequity is included. In healthcare, health inequalities are not necessarily inequitable. To elaborate, there are three groups of determinants of population health that may result in health inequalities among individuals in a society, according to Evans et al. (1994) and Olsen (2012). Firstly, inherited diseases are explained by genetic differences between individuals. The second determinant of health inequalities is the social and physical environment. Thirdly, people’s behavior related to health also contributes to health inequalities. To some extent, the latter two determinants can be avoided. Therefore, those avoidable, unnecessary inequalities would be considered health inequities.

Whitehead (1992, p. 432) classifies determinants of health inequalities further in seven categories:

1. Biological (or genetic) differences

2. Health-damaging behavior from free-choice, for example, playing risky sports and pastimes

3. Temporary health advantage by early adoption of a health-promoting behavior by one group (if other groups will catch up)

4. Health-damaging behavior because of severely restricted lifestyle choices

5. Being exposed to stressful and/or unhealthy living and working situations

6. Deficient access to fundamental health and other public services

7. “Natural selection or health-related social mobility, involving the tendency for sick people to move down the social scale”

Whitehead (1992) claims that the first three categories normally do not count as inequities in the literature. One example of the biological variation (point one) is that, because of the gender difference, cervical and ovarian cancer only appear in women and prostate cancer only appears in men. This inevitable variation should not be considered health inequity. In the second point, some behavior, about which individuals have a great degree of choice (e.g., playing soccer), may have a higher risk of injury, but it should be viewed as a voluntary choice. For point three, it makes sense to give an example about a vaccination program. If there are not enough doses for a particular population, some individuals benefit from it before it reaches others. In this case, it is also personal choice since they already accepted any risk by taking the vaccines.

The remaining determinants of health (points four to seven) are avoidable, and sometimes unfair and unjust to some extent. The level of unfairness and unjustness can be judged by the extent to which individuals choose to do the health-harming activities, or whether these activities are out of their control. For instance, in an area where most available foods are unhealthy and the healthy options are expensive, people have less choice of healthy options if they cannot afford to pay for them. Another example is that many disabled people suffer from adverse health conditions, that may lead to fewer job opportunities. This would cause even a worse health status for them and repeat the cycle.

#### Towards a definition of equity in health

So, what is inequity? There are four characteristics that make inequality become inequity: unnecessary, avoidable, unfair, and unjust (Whitehead, 1992).

The level to which something is unfair or unjust is judged differently in different contexts and time periods, but the degree of choice involved is one concrete criterion to use for the definition. Therefore, Whitehead (1992) comes up with a working definition of equity in health as everyone having a fair chance of reaching their full health potential, and no one being at a disadvantage when trying to achieve this.

**Vertical equity**

Different levels of need should be allocated with different resources.

Based on that concept, the World Health Organization (n.d.) specifies that health equity can be achieved “when everyone can attain their full potential for health and well-being” (para. 1).

### Vertical and Horizontal Equity

**Horizontal equity**

This is the equal allocation of resources for equal need.

There are two concepts of equity that significantly influence the policy implications: vertical and horizontal, originated by a Greek philosopher, Aristotle. **Horizonal equity** is the equal allocation of resources for equal need. This concept is reflected in universal health coverage plans, which provide a basic level of healthcare for everyone in need. In healthcare, horizontal equity means that people with equal health need should get an equal level of treatment. Another concept, **vertical equity**, states that people with different levels of need should be allocated with different resources. The challenge of this approach is how to determine the appropriate level of differences in resource allocation according to the difference in need. In healthcare, this means that people with unequal health needs should receive different levels of healthcare. For example, this concept is applied in health interventions targeting the poor to reduce the health gap between the poor and the rich. In addition, vertical equity would also mean that individuals (or households) with different abilities to pay would have to make appropriately different contributions, particularly in taxation or health and social insurance schemes.

These two concepts of equity have also been applied in income taxation. Horizontal equity means that people with similar income should pay equal income tax and vertical equity claims that people with greater income should contribute more in income tax. Normally, the range of income respective to the percentage of income tax is defined by the state. Here, we are not going to discuss this topic in more depth; it is mentioned only for understanding the concepts of equity.

### Equity in Healthcare

Before discussing equity in healthcare, it is essential to note that healthcare is only one of a number of factors that result in differences in health between groups in a population, or between various populations. To close the gap, there is a need for integrated programs targeting various aspects related to health, including healthcare.

Whitehead (1992) states that healthcare offered to all groups of the population should focus on its accessibility, quality, and acceptability. The author also defines equity in healthcare as consisting of three components:

1. Equality in healthcare access for those with equal needs
2. Equality in healthcare utilization for those with equal needs
3. Equality in quality of care for all

#### Equality in healthcare access for those with equal needs

This is related to the concept of horizontal equity. To be able to achieve this, a fair distribution of healthcare services based on needs should be implemented in geographic areas and, at the same time, barriers preventing access should be removed. Inequity in healthcare access occurs if healthcare resources and facilities are not distributed fairly. This situation normally happens in rural areas where supply cannot meet demand. In terms of barriers, different types include financial, organizational, political, and cultural. An example of financial barrier among people with low income is transportation, which may be a great cost for them, but is necessary to get access to healthcare. Limited opening hours of a hospital or a clinic is an example of organizational barrier. In some contexts, there is a policy that immigrants are not allowed to use insurance-based services, which limits their access to healthcare. Language differences might also be a barrier to access to healthcare for immigrants or minority groups.

#### Equality in healthcare utilization for those with equal needs

Equality in healthcare utilization for those with equal needs is the second concept of equity in healthcare. Utilization is different from access. For example, in a mass vaccination program where a group of people with equal needs have equal access to vaccines, their utilization might be different. Since vaccination is voluntary, people may choose to opt out because of their personal or religious beliefs, which leads to the difference in healthcare utilization. In this case, it should not be considered inequitable. Unequal healthcare utilization is inequitable only when the use of health services is restricted among disadvantaged groups with a health need (Whitehead 1992). Giving an example of vaccination program again, vaccines being distributed based on positive discrimination is a sign of unequal utilization.

#### Equality in quality of care for all

In terms of equality in quality of care for all, in many communities, it is critical that all people have equal access via a fair system that is based on need rather than a person’s status. In times of scarcity or budget cuts, this problem is especially significant. The treatment of certain groups because of their race or ethnicity, for example, would look unjust in such an environment. Hyldgård et al. (2021) compare income- and education-related inequality in quality of care and conclude that income-related inequality was more common than education-related in Denmark. Unequal quality of care also occurs when comparing figures between different countries. For example, moving from Denmark and Sweden to Hungary and Latvia, the death rate of people that had a heart attack within 30 days of hospitalization doubled (Organization for Economic Cooperation and Development, 2014). This report shows that the survival rate of different cancer types also varies significantly among European countries. Therefore, efforts are needed to generate an equal quality of care throughout different levels of prevention, early screening, and early treatments of cancer, as well as other serious diseases.

### Self-Check Questions

1. Please give one example of determinants of health inequalities that are not health inequities.

*natural or biological variation, health-damaging behavior if freely chosen, such as participation in certain sports and pastimes, and the transient health advantage of one group over another when that group is first to adopt a health-promoting behavior (as long as other groups have the means to catch up fairly soon)*

1. Please define equity in health, according to Whitehead (1992)

*Equity in health implies that everyone should have a fair opportunity to attain their full health potential and, more pragmatically, that none should be disadvantaged from achieving this potential, if it can be avoided.*

1. Please differentiate between horizontal and vertical equities.

*Horizontal equity means the equal allocation of resources for equal need, whereas vertical equity means different levels of need should be allocated with different resources.*

## 3.3 Theories of Distributive Justice

There is a similarity between three concepts: equity, fairness and justice. In fact, they are closely linked to each other. In order to understand the concept of health equity, this section will go deeper into the philosophical theories of distributive justice. Although these theories can be applied in other sectors, they are highly relevant to the distribution of healthcare. Thus, in this section, we will start by discussing the link between equity, fairness, and justice. Then, popular theories of distributive justice will be addressed. We also will give existing criticisms of each theory after the discussion.

### The Link between Equity, Fairness, and Justice

In many cases, the terms fairness and equity are used interchangeably. For example, in the glossary of the National Library of Medicine (2020), equity is defined as “the degree to which some distribution or other is judged to be fair”. Equity in health has usually been understood in relation to fairness in healthcare distribution. In general, the concept of fairness refers to what is intuitively right, just, and acceptable (Olsen, 2012).

While fairness is a broad concept that can be used in many different areas in everyday life, the more philosophical concept, which is more professionally used in law and political science is justice. Rawls (1958) pointed out two principals of the concept “justice as fairness”:

1. Liberty principal. There should be equal rights to the most extensive liberties.
2. Difference principal. If inequalities work to everyone’s advantage, they are acceptable, otherwise not.

The liberty principal addresses procedural justice, whereas the difference principal addresses distributive justice and ultimately leads to the maximin solution (Olsen, 2012). In the field of health economics, the theories of distributive justice are primarily relevant.

### Distributive Justice Theories

#### Utilitarianism

The utilitarian theory originated from the work of Jeremy Bentham. He believed in the premise that people behave in a way that gives them pleasure and avoids pain, which can be valued and compared between individuals based on the utility (or happiness). The utilitarian essentially considers that the optimum social welfare should be the maximum sum of utilities of all individuals in a population. This sum is originally unweighted under the view of classical utilitarianism, which may cause a situation in which the small increase in total sum in a society would be favorable, even if it harmed others. The theory has been adapted by John Stuart Mill, starting to weight pleasure and pain levels and developing interest in the source of individual’s utility. However, the utilitarian model still relies on individual’s subjective judgement on their utility.

In healthcare, health maximization is a simplified and modified version of utilitarianism. Instead of the original concept of maximizing utility (or happiness), the focus is now on maximizing health. It seems to be simpler to measure health status than the broad concept of utility, using widely accepted available methods. Moreover, if the health scale is assigned a finite range (i.e., from 0 to 10), it would improve the preference comparison between individuals. Health gain can be treated as a social value and used for interpersonal comparisons regardless of the patients’ characteristics.

The idea of classical utilitarianism on the measurable and comparable utility was criticized as unscientific approach since one individual level of satisfaction cannot be added on the top of another person to obtain the sum. Although the revised version to weight the individual pleasure has been developed, it comes to the question of what domain should be accounted for the utility of an individual.

Another criticism is presented by Robert Nozick (1974), who raises the issue of whether or not there are any potentially malicious persons in society. Consider the following scenario: a person who, out of intolerance or pure malice, derives pleasure from the pain of another group of people in society. Is the nefarious usefulness of such a person also taken into consideration?

#### Liberalism

In the classical liberal tradition, individual freedom is valued strongly; thus, libertarians endorse strong rights to individual’s property and liberty. In other words, people have the freedom to enter an economy and choose what they believe is best for themselves and their families. However, some resources may not be appropriate for full liberty of ownership. For example, van der Vossen (2019) points out that if natural resources and artifacts are used by agents (e.g., the government), people require rights to access some of these resources. In this case, constraints exist on ownership and its appropriation. According to van der Vossen (2019), the stance on how natural resources can be owned could be put on a continuum from right- to left-liberalism. When moving from right to left, the level of preserving equality increases; hence, left liberalism comes closer to egalitarianism.

Although the ideology of libertarians features the importance of individual rights and freedoms, this does not mean that the state cannot put certain forces on the society. Friedrich Hayek (1960) states that it is permissible to enforce that people pay for public police services. Later, Robert Nozick (1974) claims that the responsibilities of state should be kept minimal to some narrow functions of protection, for instance, against force, theft, fraud, and enforcement of contracts. He states that individual rights will be violated if the state puts enforcement on other aspects than these. It seems that social programs in healthcare belong to the other aspects that a minimal state should not pay for in the view of Robert Nozick (Folland et al., 2017).

Criticism of liberalism derives from the assumed libertarian constraint itself. In many cases, some degree of liberty can be traded off to gain efficiency. For example, Pauly (1978) argues that if the cost of spreading information about a potentially unsafe drug to the whole society is large, it might be more efficient to simply restrict access to that drug.

#### Egalitarianism and Rawlsian justice

The primary (or strong) egalitarianism emphasizes equal distribution of a particular good for everybody as justice (Folland et al., 2017). For example, when we have four candies and two children are waiting, egalitarian will give two candies for each child (equal distribution). When the number of candies increases to five, “primary” egalitarian will still distribute 2­–2 among the children, even though we can give two to one child and three to another (if the candy cannot be halved). The latter distribution improves the benefit of the second child without reducing benefit of the first. This simple example of distribution is related to Rawls’ (1971) maximin (or difference) principle, which claims that “social and economic inequalities are to be arranged so that they are both (a) to the greatest benefit of the least advantaged […] and (b) open to all under conditions of fair equality of opportunity” (p. 65).

Rawls (1971) included his principle specifically in what he called primary goods, including: “(1) basic liberties such as freedom of thought, (2) freedom of choice of occupation, (3) powers and prerogatives of office, (4) income and wealth, and (5) social bases of self-respect” (p. 65). However, Rawls only briefly discusses how to value these goods, so it is little evidence on how these primary goods should be prioritized, or what trade-offs can be made among them. Since the maximin principle is based on Rawls’s work, it mainly concerns with the aforementioned primary goods, rather than health or utility. However, if we consider health as a measurable good, it still is applicable. In health, under Rawlsian justice, given that there are two people with illnesses and the available resource can only treat one, it is still better to have one healthy than two unhealthy people. This example does not say which one should be selected for treatment, which may result in discrimination. However, if we are able to evaluate who is more disadvantaged as a result of the illness, the resource should be used to treat this disadvantaged person under the view of Rawlsian justice.

In the view of Rawlsian justice, the distribution of resources, such as healthcare, would depend on the severity of the disease. There are criticisms on this pointing out the risk of mass ignorance for the rest of society due to the betterment of one individual. For example, Arrow (1973) criticizes a situation in which a disease needs expensive treatments to keep the patient barely alive with little satisfaction and puts the rest of population at risk of poverty. In general, the action should be weighted carefully and the setting of priorities is contextual- and resource-based.

#### Functioning and capabilities approach

Amartya Sen earned the Nobel Prize in economics in 1998 for his outstanding work on social choice theory. As with the multitude of impartial principles of justice that can withstand rigorous inspection, social choice theory posits that a range of competing principles may factor into our judgments of alternative social orderings. While it may appear impossible to satisfy all or even the majority of these conflicting principles simultaneously, such impasses are frequently resolved by incorporating additional information about interpersonal comparisons of well-being and relative advantages. The given set of principles of justices were named as the transcendental route by Sen (2009), and he points out two problems:

1. The feasibility of choosing an agreed transcendental solution and
2. The redundancy in the search for a transcendental solution.

Sen (2009) claims that a given set of principles of justice, such as in Rawls’ two principles, would be hard to apply to identify a nature of justice in a society. Rather, he advocates a “comparative approach” to define the desirability of particular society based on various social realizations on the continuum investigation.

The second part of his approach is considered “realization-focused,” which means that instead of focusing only on institutions and rules, it is essential to give attention to actual realizations in the involved societies. Sen (2009) differentiates the “arrangement-focused” and the “realization-focused” views of justice. The former term refers to the arrangements of some organizations, regulations, or behavior rules that indicate the view of justice in a society. The question is whether these organizations are delivering justice or not. Instead of arranging in the top level, Sen argues that the justice should include what emerges in a particular society, and how the people actually live under the circumstances of pre-defined arrangements and regulations. From this point, Sen (2009) emphasizes the importance of the realization-focused comparative approach (involving functioning and capability) over the concept of utilities and happiness (from the utilitarian viewpoint).

Gore (1997) criticizes the capability approach for its individualism. He argues that the social arrangements in Sen’s view are only evaluated by considering how good or bad it is for an individual’s well-being and freedom. This would limit the number of goods that cannot translate their values to an individual level, for example, a set of social norms or traditions in a particular society or a shared language. A similar criticism also argues that Sen’s approach fails to consider the externalities of one individual’s freedom, and that the concept of individual freedom is still vague (Nussbaum, 2003). There are different works on the capability approach, and they contain some substantial variations. One example is Nussbaum (2011)’s list of central capabilities, pointing out different categories motivated by the concept of human dignity, which is not in line with the freedom emphasis in Sen’s approach.

### Self-Check Questions

1. Please name one example of distributive justice theory.

*utilitarianism, liberalism, and egalitarianism*

1. Please name and define one out of two principles of “justice as fairness” by Rawls (1971)

*The liberty principle means there should be equal rights to the most extensive liberties, and the difference principle means that if inequalities work to everyone’s advantage, they are acceptable, otherwise they are not.*

## 3.4 Exogenous Determinants of Health

Health is influenced by various factors, including biological factors, individual choices, and the environment in which an individual lives. In this section, we will focus on the social determinants of health to further investigate how to address these factors in order to improve health equity. A conceptual framework that helps health-related institutions to recognize and reduce social determinants of health will also be discussed. Then, a case study from Denmark will be given as an example of early life conditions determining the future health status of an individual.

### Social Determinants of Health

Social determinants of health are broadly defined by the Commission on Social Determinant of Health (Commission on Social Determinants of Health [CSDH], 2008, p. 9) as “the poor health of the poor, the social gradient in health within countries, and the marked health inequities […] caused by the unequal distribution of power, income, goods, and services, globally and nationally, the consequent unfairness in […] people’s lives” (p. 9). These circumstances should be considered in the health equity concept since they mostly are avoidable and can be improved to a certain degree. The Commission on Social Determinants of Health (CSDH) was established in 2005, led by Prof. Michael Marmot who has contributed substantially to the field, to tackle the social factors that lead to health inequities around the world.

In 2010, the CSDH establish a conceptual framework (see the figure below) to point out the different levels of social determinants of health and how they can be addressed in policy action (World Health Organization, 2010). Accordingly, the framework includes two levels of social determinants, namely structural and intermediate determinants, influencing the equity in health and well-being of a population. Structural determinants of health inequities consist of broad contextual socioeconomic and political structures in a particular setting, with variety of categories. These contextual structures are reversibly related to the socioeconomic position and social class in the society, as well as the level of education, occupation, or income (see the figure below). The structural determinants lead to intermediate determinants of health, consisting of the following factors:

* material circumstances (living and working conditions, food availability, etc.)
* psychosocial factors (social support, social stigma, etc.)
* behavior (smoking, alcohol consumption, etc.)
* biological (genetics, etc.)

This is an overview of how the social factors influence equity in health and well-being of an individual. Interestingly, the social factors may contribute to the biological factors, which are normally regarded as unavoidable and are thus not considered in health inequity. To some extent, it is true, but biological differences, for example, caused by air pollution or radiation toxicity, are avoidable.

The Commission on Social Determinants of Health conceptual framework

Diagram

Description automatically generated with medium confidence

Source: Solar & Irwin (2010). CC BY 2.0.

Moving forward, the Commission on Social Determinants of Health also point out three broad approaches for policy action: “(1) targeted programs for disadvantaged populations; (2) closing health gaps between worse-off and better-off groups; and (3) addressing the social health gradient across the whole population” (World Health Organization, 2010, p.7). These three approaches are closely linked to the previous discussion on distributive theories. The first two approaches are more related to the second principle (difference principle) of Rawl’s theory of justice in reducing health equities. The third approach is specifically connected to Sen’s capability approach of looking into particular social functioning and meeting human needs. World Health Organization (2010) emphasizes that there is the need for contextual and intersectoral policy approaches at different levels of society to reduce health inequities.

### Case Study in Denmark – Early Life Conditions and Mortality Later in Life

This section is based on a case study from Denmark by van den Berg et al. (2009), who point out many early life determinants of health having a causal effect on mortality later in life. The data were collected based on different sources, such as the Danish Twin Registry, data on macro-economic outcomes, demographic indicators, and meteorological conditions (van den Berg et al., 2009). The study shows that the birth method significantly influences the life expectancy of Danish people. This study also claims that macro-economic fluctuations, which they refer to as the business cycle, substantially effects people’s socioeconomic conditions and result in a shorter life for the most disadvantaged. The business cycle contributes to the changes in at least four aspects of early life conditions, including nutrition, disease load, living conditions, and stress (van den Berg et al., 2009). This statement is in line with the social determinants of health conceptual framework, proving that the contextual factors have a strong effect on individual health and well-being.

Self-Check Questions

1. Please name two levels of social determinants in the framework by the World Health Organization (2010).

*structural and intermediate*

1. Please name one of the three broad approaches for policy action on social determinants of health proposed by the World Health Organization (2010).

*targeted programs for disadvantaged populations, closing health gaps between worse- and better-off groups, and addressing the social health gradient across the whole population*

## 3.5 Policy Lessons

So far in this unit, we have discussed different concepts of fairness in distribution, from the political ideologies to philosophical approaches, equity, and justice theories. The question is how we can apply these concepts to appropriate actions in healthcare sector. With that in mind, this section will discuss the classical seven WHO principles of action on health equity (introduced by Whitehead [1992]) and follow-up with the specific framework for action on social determinants of health by the World Health Organization (2010). We will see the link between these principles and their roots in the philosophical distributive justice theories.

### Seven Principles for Action – Whitehead (1992)

Whitehead (1992) is one of the most significant works on the efforts of defining equity in health and healthcare. In this paper, the author also comes up with seven principles for action on health equity (Whitehead, 1992, pp. 8–14):

1. “Equity policies should be concerned with improving living and working conditions.”

Whitehead (1992) states that in order to show the problems of inequities in health, we need to act on its roots, and living and working conditions are the prominent cause. She claims that this might be a more sufficient approach than providing tertiary care of illnesses rooted in these conditions. This statement should be supported by public health professionals.

Many health interventions that benefit the population at large, in fact contribute a more dramatic impact on the least advantaged group, based on the concept of diminishing marginal benefit. These interventions would also help to reduce the gap between the rich and the poor.

1. “Equity policy should be directed towards enabling healthier lifestyles”

The second principle is again related to the public health perspective that the local and national agencies should provide healthy choices in society or at least close the potential barriers preventing people from making healthier choices.

1. “Equity policy requires a genuine commitment to decentralizing power and decision-making, encouraging people to participate in every stage of the policy -making process”

By involving people in policy-making processes, they can contribute their voice and represent their needs in particular. This statement means to support the “realization-focused solution” instead of “transcendental solution,” where policies are only decided by a group of institutions. This has been criticized in the work of Sen (2009).

1. *“*Health impact assessment together with intersectoral action”

Inequities in health do not only come from healthcare sector, but from all of the sectors in any society. Whitehead (1992) highlights the need for cooperation between different sectors to assess the impact of their intended actions on the health of the population, especially for the most vulnerable groups. Moreover, the health policies in particular need to be appropriate for other sectors in the society.

1. “Mutual concern and control at the international level”

The political adaptation in one country may affect other countries. Given one example among an economic group of countries, some of them are encouraged to produce agricultural products contributing to better health, which then improves the nutrition for other countries. Therefore, the political control may need an international force in some areas.

1. “Equity in healthcare is based on the principle of making high quality healthcare accessible to all”

Equity in healthcare consists of accessibility, utilization, and quality. To improve the equity in healthcare, some of the policy actions may be resource allocation based on needs, appropriate geographical distribution of health services, quality check of health services, and examining the reasons for low healthcare utilization (Whitehead, 1992).

1. “Equity policies should be based on appropriate research, monitoring, and evaluation”

The last principle is evidence-based policies. Research on the real problems in a particular society needs to be conducted to be able to deliver the appropriate policies. Additionally, monitoring and evaluation of the interventions should be conducted before (via trial) and after implementations.

From these seven principles in action, Whitehead (1992) emphasizes that the problems of health inequity cannot be solved by one individual organization or sector; it has to be an intersectoral or even international cooperation. If we look back on these principles, many of them are related to the distributive justice theories, especially Rawl’s difference principle and Sen’s capability approach.

### Conceptual Framework for Action on Social Determinants of Health

With the efforts of the Commission on Social Determinants of Health (CSDH), the conceptual framework (see the figure below) for policy development and action has been constructed to help the analysts and policymakers target policy at different levels of determinants, including structural, intermediate, and individual (World Health Organization, 2010).

Framework for tackling inequities by addressing social determinants of health

Diagram

Description automatically generated

Source: Solar & Irwin (2010). CC BY 2.0.

This framework identifies three important strategies to tackle social determinants of health inequities: (1) contextual-specific strategies, (2) intersectoral action, and (3) social participation and empowerment. The author emphasizes the importance of intersectoral action and points out that the policy should not only focus on intermediate determinants of health but also structural determinants. These two foci, in combination with engagement of social participation and empowerment, will help facilitate the policy effectively. Additionally, these policies need to be followed up, as well as monitored and evaluated over time to ensure that they address the social determinants of that particular society. Evidence across countries can be used to implement the policy in another context; however, it needs to be carefully assessed before adaptation. Further detailed discussion can be discovered in the work of the World Health Organization (2010, p. 50–65).

Self-Check Questions

1. Please name three important strategies for policy action on social determinants of health proposed by the World Health Organization (2010)

*contextual-specific strategies, intersectoral action, and social participation and empowerment*

Summary

Equity and fairness in health and healthcare are the targets of all health systems around the globe. However, these concepts might be understood differently from various perspectives. We need to understand their philosophical roots to interpret them correctly and take appropriate actions.

We first need to understand the rationale for redistribution in the healthcare sector. This comes from the roots of political ideologies of generosity (or altruism) and beneficial externalities. From there, welfarism and extra-welfarism are developed in the field of economics and thus included in health economics. These two concepts have a long history of development and criticisms, and they influence the methodology of economic evaluation in healthcare.

The concept of welfarism and extra-welfarism are closely related to political philosophical theories of distributive justice, which are also applicable to healthcare distribution. Influential theories include utilitarianism, egalitarianism, maximin, and realization-focused comparative approach (or capability approach).

Based on these theoretical concepts, the avoidable determinants of health that are contributed to health inequities can be identified and addressed, for example, the social determinants of health proposed by the WHO. Altogether, these components help to concretely build up the appropriate policies in order to achieve health equity.

# Unit 4 – Delivering Healthcare

**Study Goals**

On completion of this unit, you will be able to …

… describe the role of physicians in healthcare delivery.

… analyze payment mechanisms in primary and secondary healthcare delivery.

… explain the concept of supplier-induced demand.

… understand the economics of hospital care.

# 4. Delivering Healthcare

## Introduction

Healthcare delivery is one of the most critical components in a functional health system. It ranges from the management of all services related to prevention, diagnosis, and treatment for the advancement, maintenance, and enhancement of health to the patient flow for efficient arrangement of resources. The key player in healthcare delivery is the physician, who is both an important member of the workforce and a decision-maker. In this unit, we will focus on the clinical healthcare delivery (i.e., diagnosis and treatment) by dividing it into two levels: primary and secondary healthcare. The delivery of, for example, public health intervention will be outside the scope of this unit. Thus, primary healthcare delivery will be referred to primary care physicians (also known as general practitioners) and secondary will refer to hospitals. We understand that there are other levels of healthcare delivery; however, those discussed in this unit are the core units of healthcare delivery in any healthcare systems.

This unit will start with a discussion about physicians and their behavior in relation to delivering healthcare, followed by the focus on supplier-induced demand behavior among the physicians and the rationale behind that. Then, the secondary healthcare delivery of hospitals will be analyzed, focusing on the economic aspect of hospital management.

## 4.1 The Physician as a Supplier of Medical Services

The physician is one of the most important human resources in healthcare; they are not only a service supplier but also decision-makers in many situations. Physicians have to pursue advanced educations and trainings before they can practice. This section will begin with a focus on physicians and their importance in the healthcare market. Then, a benchmark model to understand the physician’s behavior follows. Based on these, further discussion on how to pay primary care physicians will be conducted. Common payment mechanisms, including fee-for-service, capitation, salary, and pay-for-performance, will be pointed out, and summarized using real examples.

### Physician: A Fundamental Member of the Workforce and Decision-Maker

Healthcare systems are labor-intensive, necessitating the use of skilled and experienced employees in order to work correctly. A balance must be struck between the various types of health promoters and caretakers, as well as between health personnel and physical resources. Health system inputs can be grouped into three components: human resources, physical capital, and consumables. Considered to be a key worker, physicians belong to the most important input of healthcare: human resources. Certainly, there are also non-clinical health professionals, but physicians should be considered fundamental to the healthcare sector. In many societies, physicians are normally the most respected occupation because of their hard work and long-term investment (in education and practice) in improving health. In the healthcare market, physicians’ decisions affect other inputs as well, for example, consumables (i.e., medicines) and the costs of healthcare. Information asymmetry in the healthcare market also contributes to the power of physician in the decision-making process (Folland et al., 2017).

Physicians work in most units of healthcare, such as the primary care unit (the primary contact point where patients seek healthcare) and the quaternary care unit (where specialized skills related to rare diseases are needed). There are also a wide range of specializations in medical training, for example, dental, pediatric, ophthalmic, and surgical, that require a substantial investment to achieve. Despite the variety of medical areas, physician practices should be guided by some ethical principles. The four important ethical principles are as follows (Gillon, 1994; Taylor, 2013):

1. The respect for patient autonomy
2. Beneficence
3. Nonmaleficence
4. Justice

Further discussion on these principles is outside the scope of this unit.

### A Benchmark Model of the Physician’s Practice

In a special market such as healthcare, the profit-maximization concept may not fit since it involves basic medical ethics that the healthcare provider should follow. Rather, McGuire and Pauly (1991) developed a benchmark model to capture the physician’s practice behavior based on the concept of utility-maximization instead of profit-maximization. The author includes three determinants of the physician’s utility: (1) the net revenue , (2) leisure , and (3) the demand inducement . While the first two elements have positive effects on the utility of a physician, the latter contributes to physician’s utility negatively. To be effective, this model uses some essential assumptions (McGuire & Pauly, 1991). First, each physician is assumed to have a certain number of patients who paid a pre-defined amount of out-of-pocket money for the health services provided by the physician. This leads to a fixed price of each service delivered. It is also assumed that the physician is capable of inducing the demand by changes to the quantity of services to adjust their earning from the number of services delivered. Now, we look at the physician’s utility function based on the benchmark model:

There are three independent variables in this function that contribute to the utility of the physician. The physicians have to trade-off between each pair of these variables to achieve their desired utility. There are three pairs to consider:

1. Net income and leisure time
2. Leisure time and inducement
3. Net income and inducement

The physician can achieve their desired utility by trading off between net income and leisure time. These variables have a negative correlation with each other, meaning that to increase the net income, physicians have to trade off their leisure time. On the other two pairs, there are positive correlations, meaning that increasing one variable will lead to the increase of another to maintain the utility of the physician. In this model, inducement is considered a “bad” thing, as increasing inducement will reduce the physician’s utility. To compensate for that reduction, there should be an increase in other variables, such as net income and leisure (L). Depending on the preferences of the individual physician, relationships between these variables vary (McGuire & Pauly, 1991).

This Benchmark Model includes the net income variable, which points out that physicians also seek profit. This prevents the physicians from becoming a perfect agent. However, the profit incentives differ depending on reimbursement strategies.

### Paying Primary Care Physicians

In the following, we will focus on the payment mechanisms available for primary care physicians, who are the first point of contact when patients seek care. Normally, in other levels of care, there is a more complicated organization, such as hospitals. This section touches upon the primary care unit, which provides basic health services and is owned by a small group of physicians (also called general practitioners [GPs]). In this section, we use the term primary care physicians (PCPs) throughout for consistency. There are various payment mechanisms in different health systems around the world, and each of them has a different mixture of payment methods. We will discuss in turn the three classical payment mechanisms: fee-for-service (FFS), capitation, and salary, and we will also include some examples of the pay for performance (P4P) mechanism. Real-world examples will be given for each of the payment mechanisms.

#### Fee-for-service (FFS)

The term “**fee-for-service**” (FFS) refers to a payment structure in which a healthcare provider (i.e., PCPs) is compensated for each service performed on a patient. FFS serves a dual goal of paying the cost practice incurred while also providing a financial incentive for the physician (as an agent) to execute more of the duties requested by the purchaser (as principal). To accomplish the latter goal, favored jobs would be compensated at a greater rate per unit of physician work. FFS is considered the most complicated remuneration mechanism (Olsen, 2017) since it involves itemized payment when there are huge variations of healthcare services available. The reimbursement based on this mechanism can be paid differently depending on contexts. The fee can be paid directly by the patients (out-of-pocket expenditure) or by a third-party payer, either health insurers or public funds. Rafiei et al. (2022) claim that FFS is used in most healthcare systems in the world; however, it accounts for various proportions of PCP’s payment in different countries or even within a country. For example, while the FFS is used to reimburse 30 percent of the total PCP’s payment in Denmark, this number is up to 95 percent in Australia. Within Canada, it is even different for different type of PCPs, ranging from 45 to 85 percent (Rafiei et al., 2022). Rafei et al. (2022) also mention that physical examination, consultation, and vaccination are reimbursed using the FFS mechanism in most countries around the globe.

**Fee-for-service**

This is a payment structure in which a provider is paid per service performed on a patient.

As FFS reimburses for every service provided, it creates strong incentives for physicians to give patients necessary treatments; however, supplier-induced demand (SID) might occur in this case, as the physicians may provide unnecessary services to increase their profit. FFS encourages the PCPs to increase the production and quality of their services, which helps to reduce referral to higher levels, reducing the pressure on higher levels of healthcare. As FFS is complicated, one disadvantage is high administrative costs for both providers and governments.

#### Capitation

**Capitation** refers to the payment method in which physicians receive a set fee for each enrolled patient, which means that their income is based on the number of patients on their list. Under capitation, physicians benefit from investing in long-term relationships with their patients and providing quality services to gain patients’ trust. In practice, the number of patients per physician maybe regulated, i.e., there may be a minimum or maximum number of enrolled patients allowed (Olsen, 2017).

**Capitation**

Physicians receive a fixed amount of money for each enrolled patient.

This payment mechanism creates some negative incentives for physicians. For instance, capitation may encourage unnecessary referrals to specialists, giving the PCPs a “free-ride” and reducing their cost without affecting their income (Olsen, 2017). However, the physicians should be careful; if the patients are not satisfied with the services provided by them, they may register elsewhere. Unnecessary referrals could also overburden higher levels of care. Another possible form of undesirable physician behavior that is incentivized by capitation results from the possibility of satisfying patients by accepting sick leave requests even when unnecessary. Generally, capitation gives more power to the patients if they can choose their preferred physician with whom to register.

Generally speaking, capitation creates an incentive to reduce costs, thus sometimes leading to undertreatment. If possible in a given context, capitation may incentivize the physicians to choose patients in low- rather than high-risk groups, which is known as the “cream skimming” phenomenon. This would reduce the physicians’ costs, increasing their profits. To deal with this undesirable incentive, many countries apply the adjustment factor for capitation among the high-risk groups (Rafiei et al., 2022). For example, in Turkey, the coefficient of capitation is higher among children under four years (coefficient 1.6), elderly over 65 years (1.6), pregnant women (3), and prisoners (2.25). The coefficient is 0.79 for the rest of the general population (Rafiei et al., 2022).

Despite the negative effects, capitation also brings advantages. As reimbursement is based on number of patients registered rather than number of services provided, there are strong incentives for cost-saving among PCPs. Consequently, this mechanism promotes the avoidance of unnecessary treatments. Capitation also encourages doctor continuity (physicians maintaining long-term relationships with their patients), which helps to monitor the progress of the patients’ health problems.

In most health systems where capitation is used, this payment approach represents only one element of the physician's total remuneration.

#### Salary

**Salary**

This payment method is independent of the services provided or number of patients.

**Salary** refers to payment given based on time, usually monthly, regardless of the services provided and number of patients. This payment method intends to create security for physicians and improve accessibility for the patients (Rudmik et al., 2014). This is why the salary payment mechanism appears mostly in the public sector in many countries, such as Sweden, Thailand, and Iran (Rafiei et al., 2022). Rudmik et al. (2014) state that the salary payment mechanism also creates an incentive to reduce the quantity of care. This method is particularly suitable for recruiting physicians in underpopulated and undersupplied regions. However, salary may also create negative incentives, especially in the underfunded system. Low salary could result in physicians asking for informal payments from patients, which is a common situation in developing countries (Pekerti et al., 2017). Similar to capitation, salary may also create an incentive for referral since physicians do not receive remuneration for their services. Patients often complain about discourteous physicians under this mechanism (Dan & Savi, 2015). Salary payment also limits the incentives of physicians to strive for quality and patient satisfaction, as well as patient continuity.

#### Pay for performance (P4P)

**Pay for performance**

This is a payment mechanism that rewards physicians if they meet pre-defined targets for quality or efficiency in health services delivery.

Aside from the three classical remuneration mechanism for PCPs discussed above, **pay for performance** (P4P) has been used more in recent years to improve the quality of healthcare services and increase efficiency in healthcare delivery (Rudmik et al., 2014). P4P rewards physicians for meeting pre-defined targets related to quality or efficiency (Mathes et al., 2019). Although P4P Is derived from promising rationales, it is challenging to identify and implement an appropriate P4P structure that can motivate the physicians (Rudmik et al., 2014). For example, if the government cannot provide P4P for all categories of healthcare delivery, they will choose a select few. This leads to physicians only providing exclusively those services that will be paid by the P4P mechanism. There might also be a problem that cases that do not need a particular treatment are given it anyway because of the remuneration from P4P (Rudmik et al., 2014).

Rafiei et al. (2022) describe the objectives for the P4P mechanism implemented in many countries, including improving preventive care, improving efficiency, management of chronic diseases, patient satisfaction. The authors also mention that P4P mechanism accounts for a certain percentage of physicians’ income in many countries, such as Iran (50–70 percent of rural PCPs’ income), Turkey (20 percent), and the Netherlands (ten percent).

#### The application of blended mechanism

As discussed, each type of payment mechanism has its strengths and weaknesses. Depending on the setting, each country designs its own appropriate payment mechanism for PCPs. To maximize the advantages and reduce the disadvantages of individual methods, most countries apply a blended mechanism, mixing two or more mechanisms (Ma & McGuire, 1997). Wranik and Durier-Copp (2010) outline three commonly used blended payment methods:

1. Combining FFS with capitation to help physicians earn a small fee from the patients and be remunerated by FFS
2. Adding an FFS element into the capitation system by paying physicians for pre-defined services and also allowing FFS bills for services that are not listed
3. Combining the salary system with FFS, which means that apart from the secured salary, physicians can earn extra income as a certain percentage of the bills for services

It is important to note that there is no gold standard of payment mechanism for PCPs since it depends significantly on the contextual background, which is why many countries design their own mixture of mechanisms. Some countries’ examples of PCPs payment mechanisms can be found in the work of Rafiei et al. (2022).

The table below summaries the key characteristics of three classical PCP payment mechanisms and possible actions to reduce negative effects.

Primary care physician payment mechanisms

|  |  |  |  |
| --- | --- | --- | --- |
|  | Positive effects | Negative effects | Possible actions to reduce negative effects |
| Salary | No incentives:  - to deny access or treatment  - for excessive treatment  Easy and simple administration | No incentive:  - to provide desired effort  - to strive for quality and patient satisfaction  - for patient continuity  Incentive:  - to ask for informal payments  - for excessive referral to higher levels | Non-pecuniary incentives  Performance related bonuses  Monitoring  Quality control |
| Capitation | Strong incentives  - for cost saving  - for avoiding over-treatment and unnecessary interventions  - for making the whole episode of care efficient (favors prevention)  - for doctor continuity | Incentives for  - “cream skimming” – try to list relatively healthy persons  - under-treatment  - excessive patient referral to higher levels | Open enrolment  Encourage competition  Quality control and monitoring |
| Fee-for-service | Strong incentives  - to increase production and quality  - to give patients needed treatments  - not to exclude patients | Strong incentives  - to increase production cost escalation  - induce demand oversupply and overuse  High administrative costs both for providers and government | Increase control, monitoring  Introduce cost-sharing, co-payments  Set upper limits of reimbursement |

Source: Duy Pham (2022).

### Self-Check Questions

1. Please name at least two types of payment mechanisms for primary care physicians.

*salary, capitation, fee-for-service, and pay for performance*

1. Please name the three independent variables in the physician’s utility function in the Benchmark Model by McGuire and Pauly (1991).

*net income, leisure, and the demand inducement*

1. Please name one negative effect of salary payment mechanism.

*no incentive to provide desired effort, no incentive to limit services to reduce costs, no incentive to strive for quality and patient satisfaction, and no incentive to strive for patient continuity.*

## 4.2. Supplier-Induced Demand

The behavior of physicians significantly influences the efficiency in healthcare delivery. Physicians are decision-makers in most cases; thus, their behavior will affect the health outcomes of the patients. This section will describe the relationship between physicians and their patients as a principal–agent relationship. The characteristics of physicians acting as agents will be discussed, followed by focusing on supplier-induced demand (SID), which is one popular situation identified in the healthcare market. Lastly, real-world case studies of SID in Australia and Vietnam will be given as examples. Some criticisms of the concept of SID in these studies will also be given.

### Perfect and Imperfect Agency

In an area where there is a need for a professional who has more knowledge to consult customers about appropriate action, the relationship between agent and principal is developed. This relationship commonly appears, for example, in law, mechanics, and healthcare. In healthcare, the patients can be regarded as principlals and the physicians can be considered agents.

So, what is a perfect agent in healthcare and how can they be perfect?

#### What is a perfect agent?

A common understanding of a perfect agent is someone who makes decisions that are consistent with the principal if they have the same level of information (Folland et al., 2017). There are some attributes needed to consider physicians as a perfect agent from a patient’s perspective: (1) maximum patient health status, (2) maximum patient utility, or in a societal perspective, (3) maximum health status or utility of the population (McPake et al., 2020).

The first alternative, optimizing the patient’s health status, lets the physician abide by medical ethics but may result in paternalism, which limits freedom of choice for the patients. For instance, it recommends that, instead of informing patients about the dangers of smoking or drinking alcohol, physicians may bully patients into quitting them.

The second alternative, maximizing the patient’s utility, may imply that the physicians’ main responsibility is to deliver sufficient information, to the greatest extent feasible, and allow the patients to make the choices. This seems to be consistent with a number of current concepts; however, it creates the risk of patients making bad choices if the physicians do not provide appropriate information or the patients misunderstand it.

The third alternative is maximizing the health status or utility of the whole population. In this case, the principle is not limited to an individual patient, but opened up to a broader perspective of a whole society. This raises the awareness among physicians that delivering a service to one patient may involve the denial of other patients (McPake et al., 2020). This approach will happen less in a private clinic, where the main focus of physician is on their individual patient, but it normally happens if the physicians are employed in a public system where they act as the agent of the society and the individual patient. Especially in the situation of budget constraints, the physician’s decision-making may affect the utility of the whole population.

In general, the scope of perfect agency in healthcare is still vague. In any specific situation, there is a need to clarify the role of the agent (physician) so they are able to act appropriately.

#### How can a physician be a perfect agent?

It is critical to understand that physicians’ decisions are mostly based on the best of their scientific knowledge if they are not influenced by the profit motive (McPake et al., 2020). The patients would expect that a physician’s recommendation is the best option with the best evidence available. However, science is changing quickly, and physicians are human; in many cases, the best of their scientific knowledge may not be consistent with the best available evidence. This statement does not claim that their recommendations are bad for patients, but it may not be up-to-date. Additionally, in many cases of clinical research, there are different results in different studies, which may create conflict in physicians’ minds when making decisions.

Although the above statement claimed that medical science is sometimes imprecise, this should only contribute minimally, or not all. If a physician makes a bad decision, it is not necessarily for this reason. We have to investigate if there are other reasons that drove the physician to make that decision, especially the profit motive. These, other reasons create imperfect agents. One common agent imperfection in healthcare is “supplier-induced demand” (SID), which will be discussed in more detail in the following section.

### Supplier-Induced Demand

In the imperfect agency concept, SID refers to the decisions of physicians that are different to those of perfect agents. Again, the perspective of what a perfect agent should be has a strong influence on what is an inappropriate decision. For example, a physician may recommend that a patient have more frequent diagnostic screenings. If the incentive of the physician is to diagnose the disease early for a better treatment approach, this would help to increase the utility or health of the patient in the long term. This incentive should not be considered SID. However, if the incentive of the physician is to collect fee-for-service remuneration to earn more money by recommending an excessive and unnecessary amount of screening, this should be regarded as SID.

The above example is only a simple example of a physician’s decision that points out the difficulty of identifying SID in healthcare. Seyedin et al. (2021) classified the factors that may influence SID behavior among physicians into four groups: healthcare recipient, healthcare provider, insurer, and health system.

In order to appropriately identify SID and whose behavior it influences, some economic models have been proposed. One classical method is the “fee test” proposed by Sloan and Feldman (1978) to identify the SID by looking at the market when increasing the supply of physicians. This concept has been illustrated and explained by Folland et al. (2017) under the name of supply and demand models of SID. In a normal market, the increase in supply will lead to lower costs and/or higher quantity. In this case, it is expected that the increase in number of physicians will lead to a lower fee per visit and/or a higher number of visits (McPake et al., 2020). If the physicians have incentives to increase the demand, which means shifting the demand curve, the pattern of lower fee and higher quantity may be eliminated. The figure below illustrates the change in fee per visit and number of visits in both cases of normal market (only one demand curve ) and market with SID ( and ).

Fee test of inducement

Chart, radar chart

Description automatically generated

Source: Duy Pham (2022).

In a normal market, when the supply curve () moves to the right (), which represents the increase in the supply of physicians, the equilibrium between demand and supply curves meets at point and respectively. As can be seen, it results in a lower fee per visit ( and higher number of visits . However, if the physicians have an incentive to increase the demand, the demand curve will move to the right , which results in changes in equilibrium at points and . At point , the fee per visit is the same as at point , which means that the physicians have induced the demand to maintain their fee. Or even worse, point indicates that the fee per visit is higher than it should be at . In general, to test if the physician has the incentive to create SID, Sloan and Feldman (1978) suggested increasing the supply, and if the fee per visit does not reduce, there is a possibility of SID behavior. There are some concerns about this model. First, this cannot identify slight SID behavior, where the demand curve lies between and since, in this case, the fee per visit reduces. Second, the demand and supply curves in this model seem to be simplified and assumed to be in a perfect competitive market. In healthcare, it is more complicated due to market failures.

There are some other attempts at trying to identify SID, which are discussed further in the literature by McPake et al. (2020, pp. 52–53). The authors also mention that the identification method of SID in healthcare market is still vague, which has also been pointed out by many previous reviewers in the field.

### Real-World Identification of SID – Case Studies in Australia and Vietnam

As discussed, it is difficult to identify physicians’ SID behavior in the healthcare market. The following case studies will demonstrate how SID is identified and concluded in two contexts: Australia and Vietnam.

Peacock and Richardson (2007) develop specific equations in a demand and supply model to investigate the existence of SID among Australian’s physicians. The authors also re-examine previous criticisms of the cross-sectional methods for the identification of SID. Peacock and Richardson (2007) point out that the failure of many cross-sectional studies in diagnosing SID is because of the misspecification of the included demand and supply equations. As a result, they claim that if the demand and supply equations are sufficiently specified, they have the power to identify SID in the health market. Their result concluded that the impact of SID becomes stronger when the supply of doctors increases by showing the significant increase in the average price elasticity of demand (from 0.22 to 0.46).

In a study of Vietnam, the author compares the prescription quantity and quality of private and public physicians patients with similar profiles (Nguyen, 2011). The results show that the private physicians prescribed more medicines to the patients than the public physicians, and the amounts of injection drugs were substantially higher in the private physicians’ prescriptions. Although the types of injection drugs have not been specified in detail, it is claimed by the author that this drug is used to gain trust among the patients (Nguyen, 2011). Furthermore, this study points out that patients’ education acts as a preventative of SID. The concept of SID is identified in this study by pointing out the differences in the amount of medicine provided by public and private physicians. However, further investigation is needed to prove that the excessive prescription is a direct result of SID behavior. This might involve a review of the quality of prescription in the public sector compared to the private sector, an investigation into patient experiences and outcomes, etc.

### Self-Check Questions

1. Please name one objective used to decide whether a physician is perfect agent, proposed by McPake et al. (2020).

*maximum patient health status, maximum patient utility, and maximum health status or utility of the population*

1. Please fill in the blank.

*Supplier-induced demand* refers to the decisions of physicians that are different from those of perfect agents.

## 4.3 The Economics of Hospital Care

There is no internationally accepted definition of a hospital. It is normally defined as a combination between social and medical organization, which provide healthcare for those in need (both curative and preventive) and a training place for health workers and biosocial research (World Health Organization, 1963). In this section, we will focus more on the economic aspect of hospitals, by considering them as an economic unit. The payment mechanisms for the hospital will be discussed alongside the resources for hospital revenue. Additional resources will also be included, for example, private payments, charities, and collaborations. Finally, the expenditures in a hospital will be categorized, focusing on how the hospital revenue is spent.

### The Hospital as an Economic Unit

A hospital can be regarded as an economic unit that consists of a variety of inputs that undergo transformation (processing) to generate outputs. Simultaneously, it incorporates a feedback system, which serves as proof of evaluation of the hospitals’ performance and whether or not they have achieved their objectives. These mechanisms are important to consider when evaluating the economic efficiency and efficacy of a hospital. Upstream legal structures and rules in the healthcare system can also influence the hospital’s chances of economic survival. In the following, we will examine the purposes of hospitals and investigate what determines hospital behavior from an economic perspective.

#### Purposes of hospitals from an economic perspective

The hospital is a fundamental unit of the medical system and a social organization. A singular hospital is located in a specific geographical region; involves individuals; pursues specified objectives; and is guided by a variety of regulations and principles with the purposes of organizing, cooperating, and accomplishing their objectives (Chletsos & Saiti, 2019). To achieve its objectives, there is a need not only for material and human resources, but also organizational support, which is the provision of the resources so they can be used efficiently to achieve the best possible outcomes with the lowest costs. Indeed, the survival of a hospital’s economy is highly dependent on competent administration and cost containment. This is easier said than done, as the management of hospital services and controlling costs are not simple tasks. Success requires continual managerial oversight, the ability to adapt to the demand of a special healthcare market, the supply of high-quality services, and the reassurance of financial resources to maintain a cost-effective operation. These aspects should be the hospital’s primary focus (Chletsos & Saiti, 2019). However, the intensity of these aims varies according to some characteristics, for example, hospital type and its ownership structure, as well as broader characteristics of the healthcare system or insurance mechanism.

For example, there is a possibility that a hospital may merge with another one for economic reasons. However, the prerequisites for merging might be determined by the laws governing the health system and the applicable legal framework. The top hierarchical management level (i.e., the Ministry of Health) decide which hospitals can merge, and this is normally based on the form of hospital ownership. Consequently, a hospital that is not profitable is more likely to merge with a not-for-profit hospital than with one that is profitable. Indeed, before the merge, there is the possibility that a hospital with no profit will try to improve their economic status, which likely leads to putting the cost burden on patients with private payments, depending on the health system structure (Chletsos & Saiti, 2019).

The type of health system may also influence hospitals and their foci. Financial resources and contraints regulated by the management level via organizational support are important aspects of health systems that may significantly affect the way a hospital is run. Chletsos and Saiti (2019) point out that the culture of a healthcare system might affect both system membership and the type of ownership of a hospital (e.g., not-for-profit, for-profit, investor-owned, or public).

Considering a hospital as an economic unit, their purpose in most cases is to deliver the best possible outcome for the patients with the most efficient use of resources. However, the strength of this intention depends on different characteristics, from the top level of health system and health policy to the organizational and managerial level.

#### Hospital behavior and their determinants

A hospital’s objectives are achieved by the behavior of the hospital in response to their determinants. In the following, we will further discuss determinants by categorizing them into different levels: organizational, managerial in a hospital, health market, and governmental regulation.

First, hospital behavior may be influenced positively or negatively by organizational and managerial factors within the hospital. Organizational factors refer to type and system of ownership, size of the hospital, human resources, teaching status, and in- and outpatient care (Chletsos & Saiti, 2019). Regarding type and system of ownership, for instance, private hospitals are more open to developing groups and chains to strengthen the competitiveness of participating hospitals in the group. The collaboration can create a better variety of services to compete with the limited number of services in public hospitals. In many cases, public hospitals are linked with a university, providing education and conducting research on specialized services (Chletsos & Saiti, 2019).

The size of a hospital can also affect the diversity of services provided. The investment into acute or sub-acute care for patients with acute illnesses or long-term care for patients with chronic health conditions depends greatly on the size of the hospital (Al-Amin et al., 2018; Bai & Anderson, 2016). For example, the number of beds and analysis of health insurance coverage plays an important role in the long-term care strategy of the hospital, meaning that larger hospitals are better equipped to deliver long-term care for patients. However, Chletsos and Saiti (2019) emphasizes that the administration of the hospital has a critical role in its service delivery and financial performance, by claiming that if there is maladministration in the hospital, any kind of investment will not result in positive growth. This statement refers to the managerial factors influencing the hospital’s behavior, including the decisions of the hospital management on various aspects, namely the mix of services, the intensity of human resources, the utilization of debt, and the investment in technologies and infrastructures (Chletsos & Saiti, 2019).

Secondly, a hospital’s behavior is also related to market factors. Market factors are sometimes environmental factors based on where the hospital is located. The behavior of the hospital might be different in urban and rural areas. In a rural area, where there is limited number of hospitals available, the monopolistic power of the hospital will be increased, which might affect their behavior. Various other characteristics can be named, such as wage rate, number of physicians in the area, patient socioeconomic status, competition level, etc. (Chletsos & Saiti, 2019). All of these factors contribute to the specific characteristics of the market in which the hospital is located, thus influencing hospital behavior. In a competitive market, for example, hospitals have to think about how to survive by delivering cost-effective services or trying to find competitive advantages. This pattern is more common in high socioeconomic areas, such as urban area, where there are many additional sources of funding available aside from the government funds (Chletsos & Saiti, 2019). Funding depends on the legislation and regulation in a health system; however, in most health systems, there are existing mechanisms to compensate for extra work (e.g., fee-for-service and pay for performance). In general, being located in a wealthy area creates more incentive for the hospitals to improve their services’ quantity and quality, thanks to additional financial resources. Champagne et al. (1993) point out that many hospitals located in urban areas collaborate with universities to deliver trainings, which help increase their revenue. Additionally, Swanson-Kazley and Ozcan (2007) mention that the additional financial resources will help the hospitals invest in innovative medical technologies to improve their services.

Thirdly, regulations at health system level also influence the behavior of the hospital. Changes in regulations, also known as reform, may have positive or negative impacts on the hospitals’ financial performance (Chletsos & Saiti, 2019). There are three types of governmental regulation control mechanisms: (1) investment control, (2) utilization control, and (3) reimbursement control (Gruca & Nath, 1994, p. 346). The authors refer to investment controls as regulations that limit (or encourage) investment in hospital activities, for example, to invest in new medical technologies and services or opening a new private (or even public) hospital. Utilization controls refer to the regulations of quantity and quality of services recommended to the patients, or the length of hospitalization (Gruca & Nath, 1994). This type of control has more effect on public hospitals than private ones since they depend more on public funding (Chletsos & Saiti, 2019). Lastly, reimbursement controls are regulations that manage the hospital services fees and reimbursement (Gruca & Nath, 1994). The reimbursement mechanisms for hospitals will be discussed in detail in the upcoming section. Here, it is critical to keep in mind that, these controls not only affect public hospitals, according to Chletsos and Saiti (2019), but also private hospitals in cases of third-party payments.

### Hospital Revenue

There are various financial sources contributing to the hospital revenue, including governmental budget (from taxation), social insurance, and private payments (private insurance or direct out-of-pocket payment). In the following, we will examine reimbursement mechanisms and their advantages and disadvantages, followed by further discussion on the sources of hospital revenue.

#### Reimbursement mechanisms for hospitals

There is a parallel concept between the payment mechanism for primary care physicians, which was discussed in previous section, and hospital reimbursement mechanisms. A hospital, however, is a more complex organization, so their payment mechanisms also differ. We differentiate the reimbursement mechanisms by adapting two dimensions proposed by Jegers et al. (2002): retrospective versus prospective systems, and fixed versus variable systems. Prospective systems refer to payments that are determined without a link to the hospital’s activities, while a retrospective system refers to payments that are reimbursed after the costs are incurred from hospital activities. Fixed systems refer to the payments that do not vary based on the number of services delivered, while variable systems refer to the payments that vary based on the number of services delivered. By combining these two dimensions, we come up with three possible reimbursement mechanisms, as indicated in the table below. It is noted that the combination of retrospective and fixed systems is not feasible in practice (Olsen, 2017).

Table: Categorization of hospital payment mechanisms

|  |  |  |
| --- | --- | --- |
|  | Fixed | Variable |
| Retrospective | Not feasible | *Fee-for-service* |
| Prospective | *Global budget* | *Per patient/case* |

Source: Duy Pham (2022), based on Olsen (2017).

Global budget is the combination of a prospective and fixed system, where the hospitals receive a fixed budget from the government, normally based on the previous year’s activity and their plan for the upcoming year. This payment mechanism helps to reduce the constraints of budget for the third-party payer (i.e., governmental taxation) by transferring the financial risk to the provider (i.e., hospitals; Olsen 2017). Some problems will be created, for example, if the hospital admits an excessive number of patients or provides excessive services, as they will face financial deficit as a result. On the other hand, if they try to limit the number of patients or services by, for instance, increasing the waiting time, patient satisfaction will be diminished. At the end of the financial year, the blaming game usually happens between the providers and the payers: the hospitals blame the government for not providing sufficient money and the government blame the hospitals for not managing to use the money efficiently (Olsen 2017). In many cases, “soft” additional budgets are used as the solution, providing extra money for hospitals to cover the incurred costs. However, this action may reduce the incentive for cost efficiency among the hospitals (Olsen 2017).

Payments per patient or case is the second prospective variable mechanism. This type of payment mechanism is normally based on the homogeneity of resource use among clinical cases and characteristics (e.g., diagnosis, treatment procedure, and age; Olsen 2017). The data collected from different hospitals are used as a classification reference to come up with the appropriate amount of reimbursement. The most common classification used is a diagnosis-related group (DRG) where the payments vary based on the diagnosis and recommended treatments. This mechanism allows the hospitals to adjust their treatment recommendation in specific cases, which may help to improve the efficiency of resource use (Olsen, 2017). There are three facets of this DRG mechanism, pointed out by Olsen (2017):

1. It is a classification system based on diagnosis, with evidence from resource use and case-specific characteristics.
2. It provides cost information from the national average.
3. It is a reimbursement system that covers average treatment costs, in which the patients are classified in the DRG.

The main drawback of this mechanism is the incentive of the risk-based selection of patients. Hospitals tend to admit patients with mild symptoms in a specific diagnostic group and avoid more complicated cases to reduce the costs. Moreover, unnecessary treatments may be recommended in some cases to obtain a higher level of reimbursement from the DRG classification.

Fee-for-service is a retrospective and variable payment mechanism. This mechanism is also known as retrospective per diem, which used to be a common reimbursement mechanism in many settings (Olsen, 2017). In this mechanism, the bills are collected and reimbursed for each individual patient, which is more common when using private third-party payments (i.e., private insurance firms) rather than the governmental taxation budget. With the characteristics of a retrospective mechanism, this payment method may create an incentive to increase the number of diagnostic tests and the length of stay. The hospitals will not face any financial risk under this mechanism, which may cause inefficiency in the use of the healthcare resources (Olsen, 2017).

The discussion about the three common reimbursement mechanisms shows that there are advantages and disadvantages of each type. One practical solution is to have a mixed mechanism. Olsen (2017) recommends the possible solution of combining the fixed global budgets at the macro levels with the activity-based mechanisms (i.e., fee-for-service and case-based) at the micro levels. Accordingly, a total amount of the healthcare budget can be determined at the national level and allocated to the sub-national level (e.g., provinces and districts). With this budget, the sub-national level can create a suitable activity-based mechanism in addition to the basic budget reimbursement at hand. This mixed method is more attractive for the purchasers (the national government, in this case) than the providers (hospitals). Whatever is done at the national level is bound to influence the intermediate distributors at the sub-national level.

#### Sources of hospital revenue

So far, we have discussed the reimbursement mechanisms for hospitals, and the next question is where does the money come from? Previously, governmental funding via taxation was used as a common example since it is closely related to the reimbursement mechanisms. However, there are some additional sources of hospital revenue including, for example, insurance and private payments. Compulsory social insurance can be considered a public fund and is normally contributed by both employers and employees. Even though it is a separate fund from the taxation, the payment mechanisms would be pretty much similar to what was covered in the previous discussion. In many settings, the use of health services in private hospitals is covered partly by social insurance, and the remaining is paid privately (Olsen 2017). Private payments include private insurance by choice and out-of-pocket payment from the patients. Private insurance covers some additional health services that are not covered by the public insurance, and this insurance is normally paid by individuals or employers to prevent the risk of financial crisis due to illness. Depending on the health system, patients may also need to pay out-of-pocket money to cover the costs of health services directly from their own income.

A possible source of hospital revenue can also come from collaboration. For example, the hospital can provide training for physicians in collaboration with a university to earn extra revenue. Another source is charity; however, this kind of funding should only be used for the benefit of patients, not the hospitals themselves.

### Expenditures in a Hospital

Leading on from revenue, we will touch upon the expenditures, which is how the hospital spends their revenue on functional aspects to run the process of producing health outputs for their patients. We can divide the expenditures into fixed expenditures, which refers to the operating costs that is not affected by the number of health services provided or admitted patients (e.g., infrastructure and maintenance), and variable expenditures, which refer to the costs that are proportional to the services provided or number of patients admitted (e.g., number of staffs and pharmaceutical products). In the following, we will discuss three main categories: staff wages, pharmaceutical and clinical costs, and non-clinical costs.

#### Staff wages

The staff wages account for the largest percentage of total hospital expenditures, as healthcare is a labor-intensive service (Chletsos & Saiti, 2019). The number of staff depends on the size of the hospital or number of patients that can be accommodated. Although clinical staff, i.e., physicians, nurses, and midwives, are the core staff of the hospitals, there are other non-clinical staff, such as administrative, cleaning, and information technology staff. The World Health Organization (WHO) recommend a minimum of 4.45 clinical workers per 1000 patients in order to achieve universal health coverage in a setting (Scheffler et al., 2016).

#### Pharmaceutical and clinical costs

Pharmaceutical and clinical costs also account for a significant proportion of total hospital expenditure. This type of costs depends mostly on the kind of inpatient care a patient received, as well as how long they were hospitalized (Chletsos & Saiti, 2019). The costs of pharmaceutical products are usually regulated at upper level above the hospital; however, if there are similar medicines offered by different companies, the hospital can make a choice. The better the competition in the market, the more choices available from which the hospital can choose. In fact, depending on the flexibility of the health system regulations, hospital themselves can earn a profit by selling prescription medicines, especially in private hospitals.

#### Non-clinical costs

Non-clinical costs are incurred by the hospital’s activities. Hospitals have to pay for construction and maintenance of the infrastructure, equipment, or even advertisement. To build up a new hospital, specifically a private hospital, enormous amount of investment is needed. As a result, hospitals have to calculate these costs by adding overhead and maintenance costs on top of the services costs. This explains why most private hospitals charge higher prices than public ones for similar services. However, other aspects can compensate for this difference in cost with, for example, reduced waiting times and higher quality of services.

### Self-Check Questions

1. Please describe the purpose of a hospital when considering it as an economic unit.

*Considering a hospital as an economic unit, its purpose is to deliver the best possible outcome to the patients with the most efficient use of resources.*

1. Please name the two dimensions that are used to differentiate the reimbursement mechanisms by Jegers et al. (2002).

*retrospective versus prospective systems, and fixed versus variable systems*

1. Please list three categories of hospital expenditures.

*staff wages, pharmaceutical and clinical costs, and non-clinical costs*

Summary

There were two levels of healthcare delivery discussed in this unit: primary care physicians and secondary care hospitals. It is critical to note that physicians play an important role in any level of healthcare delivery, thus their behavior has significant impact on the efficiency of the health system. Healthcare is a special market, with some characteristics that are different from a perfect competitive market. Accordingly, the concept of profit-maximization may not be applied in the field of healthcare delivery. The benchmark model by McGuire and Pauly (1991) can better explain physician behavior based on utility-maximization concept. Based on this model, a physician’s utility function includes three independent variables: net income, leisure time, and inducement.

The trade-off between these variables in order to achieve the desired utility influences a physician’s behavior. In some cases, it leads to supplier-induced demand behavior. This phenomenon is difficult to identify in real practices and many models have been developed for the identification of SID. However, the concept of SID needs to be understood thoroughly before moving on to its analysis and identification, which seems not to be the case in the discussed case study about Vietnam.

Regarding the payment mechanisms, there are some similarities in payments in the primary and secondary levels of healthcare delivery (i.e., primary care physicians and hospitals). Each type of payment mechanism has its own strengths and limitations, and as a result, in practice the mixed mechanism is often chosen. So far, there is no proven gold standard for payment mechanism. Each context sets up the most suitable one for them based on their beliefs and experiences.

# 

# Unit 5 – Economic Evaluation and Priority Setting

**Study Goals**

On completion of this unit, you will be able to …

… understand the purpose of economic evaluation.

… describe the measurements of benefits.

… identify cost categories in healthcare interventions.

… apply economic evaluations to decision-making.

# 5. Economic Evaluation and Priority Setting

## Introduction

As in all economics, scarcity is at the heart of health economics. Health economic analyses support decision-making in questions of health policy and healthcare. Health outcomes are often the target of the analysis. In health economic evaluations, we consider two components: costs and benefits (or effects). The purpose of economic evaluations in healthcare interventions is to identify those interventions that bring relatively low costs in combination with substantial health outcomes for society. The results of economic evaluations can be used in priority setting to determine the interventions that are better to implement than others in the context of resource scarcity.

In this unit, we will start with a discussion on the two components of economic evaluations, benefits, and costs and how to measure them. Benefits can be divided into health-related and consumption benefits, which will be explored in more detail. Costs can be measured differently depending on the perspective, and this will also be discussed. Throughout the discussions, evidence from available resources in the field will be provided to show how costs and benefits are measured in different contexts around the world.

Types of economic evaluation will be further analyzed by comparing their characteristics and applications. This analysis will be followed by a deeper discussion on the criticisms of related measures in economic evaluation, including quality-adjusted life years (QALY) and the cost-effectiveness threshold. This will be helpful for health economists so they can look at the overall picture and identify the most suitable methods for economic evaluations in a specific context.

Lastly, policy implications of economic evaluations will be touched upon. We will examine the differences between research and practice by elaborating on the applications of economic evaluation in implementation science. This section will raise awareness of impactful research and informing the decision-making process, which needs to be taken into consideration when doing research. The unit will be concluded by examples of the applications of economic evaluation in public health in the U.S.

## 5.1 Benefits and the Measurement of Health Benefits

In economic evaluations, benefits can be measured in several ways. Most of the available literature in the field measures the health-related benefits of healthcare interventions, and this type of benefit appears in most of the guidelines in health technology assessment (ISPOR, 2022). However, there is also an increase in literature discussing the measurement of consumption benefits of healthcare, for example, **willingness to pay** (WTP) measures. This section will begin with a discussion on health-related benefits that are applied in economic evaluations, including clinical outcome, quality-of-life measures, and multi-attributable utility instruments (MAUIs). This is followed by the measurement of consumption benefits, with a focus on the monetary measures of WTP and its applications.

**Willingness to pay** This isthe maximum price that a consumer is willing to pay for one unit of a product (Varian, 1992).

### The Measurement of Health-Related Benefits

Health-related benefits can be observed in different data, both experimental and observational. In the following, we will focus on the measurements of health-related benefits in economic evaluation. These measures include clinical outcomes, quality-of-life (QoL) measures, and multi-attributable utility instruments (MAUIs), which will all be discussed.

#### Clinical outcomes

The clinical outcome is dependent on the objective of the health interventions. For serious cases, such as advanced stages of cancer or renal dialysis, the main objective of treatments is to extend life. As a result, the number of life-years gained is selected as a measurement of health benefits in some economic evaluations (Drummond et al., 2015). Tengs and Wallace (2000) point out that up to 90 percent of health gain comes from life extension of patients undergoing cancer therapies. However, Drummond et al. (2015) claim that the years gained by treatments should be adjusted based on quality of life. The concept of quality of life, which will be discussed further later in the unit, sparks the thought that being alive in certain conditions is worse than death. This line of thinking creates many debates in research area due to ethical issues, although it may be the preference of patients suffering such conditions. This is mentioned to show that life-year gain is an incomplete measure that may not be an ideal measurement of benefits in health economic evaluation. In fact, in health technology assessment (HTA) guidelines around Europe, quality-adjusted life years (QALYs) are the preferred measure (European Network for Health Technology Assessment, 2015).

Life-year gain may also be a redundant measure in comparisons between therapeutic areas, or even in distinct stages of one area. For example, the life-year gain of a patient with cancer may be valued differently to that of a patient with diabetes, and the life-year gain of a patient with advanced cancer may also be different to that of a patient with early-stage cancer. In healthcare decision-making, where there is budget constraint, it is necessary to have a measure that is comparable both within and between therapeutic areas, which may not be possible using life-year gain as an indicator.

An additional issue of clinical outcomes is that some of them are not the final end points. The final end point is the survival status of an individual. It might be feasible to follow a group of patients with severe diseases whose lives are predicted to end in a short period of time; however, for health interventions concerning chronic disease with a longer life expectancy, it might not be feasible to fully monitor them until they are dead. As a result, many studies encounter intermediate end points, for example, change in metabolism, blood pressure change, and time to progression (Drummond et al., 2015). These indicators can predict the final end points; however, significant errors may occur when estimating final end points based on these intermediate ones.

We also see the issue of using intermediate end points in the economic evaluation of preventive and diagnostic interventions. For example, by introducing a new screening intervention, the number of early diagnoses of a certain disease will increase. Many people may regard this as an effective intervention and thus conclude that it is cost-effective; however, this perception might be inappropriate. The increase in the number of early diagnoses is a positive result, but the benefits of the intervention should be investigated in a broader picture, accounting for the costs and effects of clinical treatments of these diagnosed cases in a long-term follow-up. Then, the comparison should be made with other alternatives, for example, a no screening scenario, to see if it is worth implementing such an intervention in practice. In summary, we need to be cautious when using clinical outcome as a measurement of benefits in economic evaluation.

#### Quality of life measures

The term “quality of life” originally came from the U.S., describing the effect of material resources (such as a house, car, and other goods) on people’s perception, and the term was expanded in other areas, including health, education, and social welfare (Carr et al., 1996). The concept of quality of life is becoming more common in health-related research as a supplement to clinical outcomes. The health-related quality of life (HRQoL) is measured using patient-reported outcome (PRO) measures, reflecting aspects other than the main clinical outcome. Valderas and Alonso (2008) adapt different models to create an integrated model of health outcome, showing different constructs affecting the outcome of HRQoL. These constructs include: “symptom status functional status; and general health perceptions” (Valderas & Alonso, 2008, p. 1129). All of these constructs are correlated to personal and environmental factors of an individual (Valderas & Alonso, 2008). There are two categories of HRQoL measures: generic and disease-specific (or condition-specific).

Generic HRQoL measures include items that are not specific to one disease, but rather address general dimensions that can be applied to any health condition. These measures can be further divided into two sub-groups: health profile and preference-based. Health profile measures provide an array of scores addressing different domains of HRQoL. One popular health profile measure is the *Medical Outcomes Study 36-Item Short Form (SF-36) Health Survey*, which includes eight domains of HRQoL (Coons et al., 2000). Health profile measures have the advantage of separating each domain of HRQoL to compare treatment effects more easily on a particular domain. Preference-based measures provide a single score, also known as a health index, ranging from 0 (death) to 1 (perfect health). This score reflects the cross-sectional subjective health status of an individual at the time of response based on that individual’s preference. Utility value can be derived from many of these preference-based measures. Popular generic preference-based measures of HRQoL are EuroQol (EQ-5D), the quality of wellbeing scale (QWS), and the health utilities index (HUI).

Disease-specific measures focus on a particular disease or group of diseases. One example is the EORTC QLQ-C30, which focuses on cancer. These measures are useful when investigating the efficacy of disease-specific treatment; however, they have a limitation for economic evaluation: it cannot be used as a comparison between different treatments in different diseases (Drummond et al., 2015). There are some efforts being made to develop an algorithm to create a summary score for comparison; for example, the QLU-C10D has been developed to derive a cancer-specific utility from the HRQoL measure of QLQ-C30 by the Quality of Life Group at the European Organization for Research and Treatment of Cancer (EORTC).

The aforementioned measures are HRQoL measures. In clinical research, quality of life is mostly referred to as HRQoL, which leads to the dominance of measures addressing it. However, we believe there might be components other than health that contribute to the overall quality of life of an individual, for example, social or economic impacts. Imagine one person has a child with cancer, which may affect their psychological wellbeing, thus reducing their quality of life. Another example is that given the same disease condition, one patient might suffer more than another if they cannot afford the treatment costs, resulting in more financial worries. These considerations are getting noticed in research, for example, the concept of financial burden and its impacts on patients and their households (Altice et al., 2017; Carrera et al., 2018; Newton et al., 2020).

#### Generic measures of health-related benefits

So far, we have discussed the clinical outcome of life-year gained and quality of life measures. To encompass the gain in health, we need to capture both the length and quality of life in a single measure. This is the purpose of generic measures of health-related benefits (or health gain). These measures can be used to compare the benefits between different health interventions and are dominantly used in economic evaluation. In the following, we will discuss the two most common generic measures of health gain used in economic evaluation: quality-adjusted life year (QALY) and disability-adjusted life year (DALY).

The first study to introduce the concept of QALY was in 1968, comparing kidney implant and dialysis in both quality of life gained (reduced morbidity) and survival gained (reduced mortality; Klarman et al., 1968). In this study, the measures of QALY were not developed yet, it was just mentioned for consideration. Since then, by recognizing the advantage of this concept in economic evaluation, many research efforts have been invested in developing the measures of QALY. The QALY measures are thus developed based on the preferences of health states, representing the HRQoL of those health states (Drummond et al., 2015). The figure below illustrates the concept of QALY gained when introducing a new health intervention.

QALY of a patient with and without treatment

Chart, histogram, waterfall chart

Description automatically generatedSource: Duy Pham (2022).

QALY weights are measured in a longitudinal follow-up to observe the changes over time and account for the gain by introducing a new treatment. The QALY loss and QALY without treatment areas in the above figure gives the sum of QALY of an individual without treatment. When introducing a new treatment, the total QALY is the colored area (all colors). It is quite common that at the early stage of the treatment, patients may suffer from side effects, that may lower their QALY weights, or even lead to QALY loss. However, after that, the QALY improves over time. This improvement also results in a longer life for the patient, illustrated by more years of follow-up until death in the horizontal axis of the figure. Thus, the gain in QALY over the lifetime should be calculated by subtracting the QALY loss from QALY gained. This figure can be used to compare QALY changes (gained or lost) between two or more health interventions after having estimated the QALY weights over the years of follow-up. The QALY weights presented in the figure above lie on an interval scale from 0 (=death) to 1 (=perfect health), which is a commonly used measurement. However, some preferences from individuals indicate that some conditions are even worse than death, which results in a negative QALY weight below 0. All in all, it is important to note that QALY is a preference-based concept.

The disability-adjusted life year (DALY) concept was developed to assess the Global Burden of Disease, which was first indicated in the study by Murray and Lopez (1997). Generally, this measure is conceptually similar to QALY, despite some critical distinctions from the original version in 1996:

* The life-expectancy used in the DALY is constant and based on the highest life expectancy in the world, which is that of Japanese women (Murray & Lopez, 1997). This differs from the case-specific QALY life-expectancy, which is normally based on a long follow-up of a specific disease. This difference helps to reduce the amount of follow-up time needed to identify the expected life expectancy; however, it may create an inappropriate measure in some diseases.
* Disability weights in DALY were pre-defined by using trade-off scores from a health professional panel that met in Geneva in August 1995 (Murray & Lopez, 1997). Preferences of patients have not been considered in this case. This received a substantial debate as described in the literature by Arnesen and Kapiriri (2004).
* Although DALY also used the same scale as QALY from 0 to 1 to indicate the states of death and perfect health respectively, it is more likely a categorical scale than interval one, since only seven values were in the range that can be collected for a disease.
* DALY initially used age weight, with a lower weight in young age than older age. This age weight might raise some equity concerns, according to (Murray & Lopez, 2013). Thus, the more updated Global Burden of Disease in 2010 had dropped this age weight because of ethical criticisms.

The primary purpose of the DALY measure is to estimate the burden of disease. Despite criticisms of this concept, some economic evaluations still use it to measure the health-related benefits (losses) of healthcare interventions, especially international agencies (e.g., the World Health Organization [WHO]).

#### Measuring preferences

The term “preference” is normally interchangeable with “value” and “utility”; however, based on Drummond et al.’s (2015) clarification, preference is an umbrella term that contains two other concepts. Therefore, we will use the term “preference” to generalize the measurement concept. According to Drummond et al. (2015), the three most popular techniques for measuring individual preferences are rating scale (and its variants), standard gamble (SG), and time-trade-off (TTO).

The rating scale is the simplest technique used to measure the preferences, and is normally a numbered scale ranging from 0 to 100. It has a variation named category rating or category scaling, which is a scale of a small number of categories, around ten. The visual analog scale (VAS) is another variant of the rating scale that is a line (normally ten centimeters) containing two ends without any mark in the middle of the line. An example of an instrument that uses VAS is EQ5D by EuroQoL. Although a rating scale constitutes a quick and sufficient technique to measure preferences, it has a risk of producing bias. One possible bias pointed out in the literature by Bleichrodt and Johannesson (1997) and Torrance et al. (2001) is end-of-scale bias, in which the respondents hesitate to use the end point of the scale to reflect their health outcome. Due to the risk of bias of the rating scale, Drummond et al. (2015) point out two approaches that can fix this. One approach is collecting the values for health states using the rating scale technique and mapping the score to other choice-based techniques, such as SG or TTO; however, this is not as good as direct use of the choice-based techniques. The second approach is to use a mixture of measuring techniques – the rating scale is a good starter – to help the respondents be familiar with the preferences before moving on to other techniques.

Standard gamble (SG) is a choice-based technique of measuring preferences. This technique can be used to measure the preferences for both chronic and temporary health states. In terms of chronic health state, SG is only feasible if the chronic state is preferred to death; it cannot be used if the state is considered worse than death. In the applicable case, two alternatives are given to the subject:

1. A treatment with two possible outcomes: a probability to return to perfect health for the remaining lifetime or a probability to die immediately.
2. Keeping the chronic state for the patient’s remaining lifetime

The technique is to vary the probability until there is indifference between two alternatives, in which case the preference score is equal to . Since ranges from 0 to 1, the assumption should be made that the optimal health state has a value of 1 and death has a value of 0. SG can also be used to measure the preferences for temporary health state (Drummond et al., 2015).

Time-trade-off is another choice-based technique that can be used to measure the preferences. Instead of gambling with the probability of death in SG, TTO is the tradeoff of time for better health. In the chronic health state preferred to death, two alternatives are available to the respondents:

1. Keeping the chronic state for the remaining lifetime and then dying
2. Being healthy for a shorter time period and then dying

will be varied until the respondents consider that it is indifferent between the two alternatives. The preference at that point is measured as . The range from 0 to 1 for death and perfect health respectively is also used in this technique. TTO can also be used to measure the preferences for temporary health state, which is explained in more detail in the literature by Drummond et al. (2015).

The discussion above summarized three common techniques for measuring preferences; however, to apply these techniques, the process is much more complex and time consuming than it looks. Alternatively, it is becoming more common to use pre-scored multi-attribute health status classification systems to measure health preferences (Drummond et al., 2015). These are also known as multi-attributable utility instruments (MAUIs) that derive a utility score for each health state based on the instruments of various attributes. Some popular MAUIs are the EQ-5D, SF-6D, and the health utilities index (HUI). The development and comparison of these instruments and their scoring techniques are outside the scope of this unit, but information can be obtained in the literature by Richardson et al. (2014, 2016).

### The Measurement of Consuming Benefit of Healthcare

Most of the available economic evaluation guidelines at national level recommend the use of QALY as a preferred outcome measure (ISPOR, 2022). However, there is a growing interest in alternative measures of healthcare benefits, for example, the willingness to pay (WTP) measurement (Drummond et al., 2015). In the Swedish guideline, it is mentioned that WTP can be used in some special circumstances as a preferred outcome measure (ISPOR, 2022). The rationales for using these techniques may come from different perceptions in research, as pointed out by Drummond et al. (2015, pp. 181–182). The use of WTP may come from supporters of welfarism. The main purpose of the WTP approach is to attach a monetary value to health-related benefits in order to make a comparison with other investments beyond the healthcare sector.

In the following, we will discuss the possible techniques used to assign the monetary value to the benefits of healthcare interventions, with a focus on preferences of WTP in the context of economic evaluation.

#### Monetary value of health outcome and willingness to pay (WTP) approach

When assigning a monetary value to health outcomes, this normally refers to the “cost-benefit analysis” (CBA) in health economic evaluation, which differs from those evaluations measuring health-related benefits by their natural units (Sculpher & Claxton, 2012). There are three general approaches to valuing health outcomes monetarily: (1) WTP approach based on stated preferences, (2) human capital approach, and (3) revealed preference approach. We will focus on the WTP approach in the following paragraphs (Drummond et al., 2015).

Through the concept of WTP, we can estimate how much people value their health benefits using a monetary term. As a result, we can compare these values to other commodities of interest, given the budget constraints. In health-related research, two main methods are used to measure WTP: contingent valuation method and choice experiments (Carson, 2012; Ryan et al., 2001; Steigenberger et al., 2022). Drummond et al. (2015) illustrate different perspectives on measuring WTP, ranging from restricted perspective considering only those benefits that do not have monetary value in market (i.e., value of changes in health), to global perspective considering not only the health benefits but also other costs, for example, future healthcare cost avoided or increased productive output as a result of improved health status. The latter can be valued based on the market from a societal perspective in economic evaluation. Steigenberger et al. (2022) point out that WTP varies among the respondents due to differences in socioeconomic factors, perceived benefit, threat and barrier, etc.

The WTP approach creates a substantial debate in health research. For the supporters, WTP approach is superior to the QALY approach in terms of considering additional non-health benefits and societal preferences instead of only focusing on health-related benefits (O’Brien & Viramontes, 1994; Olsen & Smith, 2001). Olsen and Smith (2001) also point out two other arguments favoring WTP: WTP is theoretically correct since it is rooted from welfare economics, and WTP has a rationale of improving allocative efficiency (Olsen & Smith, 2001). However, the authors also emphasize that these theoretical advantages of WTP have not been expressed appropriately in empirical studies using this approach.

Criticisms of this approach also exist. For example, Cookson (2003) points out that WTP responses might be under-sensitive to the magnitude of benefit. In other words, when the benefits vary significantly, WTP responses are relatively constant among the respondents. Moreover, WTP may inflate the monetary value of health outcome in the examining intervention compared to other interventions (Cookson, 2003). In this case, the WTP is excessive compared to its actual value. After all, the methodologies of valuing WTP need to be further developed and validated in order to deliver a reliable value for health economic evaluation (i.e., cost-benefit analysis; Drummond et al., 2015; Olsen & Smith, 2001).

### Self-Check Questions

1. What is the most recommended measure of health benefits in available guidelines?

*quality-adjusted life year (QALY)*

1. Please name three most popular methods for measuring preferences.

*rating scale, standard gamble, and time trade-off*

1. Please define willingness to pay (WTP).

*Willingness to pay is the maximum price that a consumer is willing to pay for one unit of a product.*

## 5.2 Costing Healthcare

Cost analysis is an indispensable part of any health economic evaluation. Although the cost of a health intervention might vary between settings, some general guidance can be given to observe costs for comparison between alternatives. This section will start with a discussion on the concept of cost in economics and explain why opportunity costs are preferred over market costs in economic evaluation. Then, different cost perspectives will be addressed, followed by cost categorizations in available guidance.

### Opportunity or Market Costs?

In economics, opportunity cost is the value of the best alternative that is forgone in order to produce an outcome. This concept is relevant in the healthcare market, which is an imperfect competitive market. In a perfect competitive market, market price is a good estimation of opportunity cost. However, if market failures exist, the market price does not fully reflect the economic opportunity cost; thus, prices need to be adjusted by accounting for not only the explicit market costs, but also other implicit costs incurred. This statement does not mean to ignore the market costs – in fact, many studies still use existing market prices whenever they are available, unless the additional consideration of other costs is justified. For example, if the intention is to determine how much a health intervention costs in order to identify the budget needs, market costs might be sufficient. On the other hand, if the intention is to analyze costs incurred from a societal perspective, other costs apart from market costs may be needed. Drummond et al. (2015) mention that if the costs are adjusted, they need to be convinced that (1) substantial bias will occur if the market prices are unadjusted, and (2) the adjustments are made in a clear and objective way.

### Costing Perspective

In economic evaluation, it is important to identify the perspective before calculating costs since the included costs can be different depending on the perspective. For example, traveling costs incurred by patients and their caregivers when seeking care might be notable from patient’s or societal perspectives, but this is not really relevant to the payers. Costing perspective needs to be defined prior to any analysis, and it depends greatly on the available data, resources, and time to conduct the study (Drummond et al., 2015). Some perspectives include those of the payer (or healthcare), societal, employer, patient, and family. The two most common perspectives in economic evaluation studies are payer (or healthcare) and societal perspectives. Most health technology assessment guidelines specify the perspective recommended for economic evaluations (ISPOR, 2022). In the following, we will categorize costs using different perspectives, from the narrower perspective of payer (or healthcare) to the broader perspective of society.

### Cost Categorizations

There are many ways to categorize costs, the most common of which in economic evaluation are direct and indirect costs that appear in almost all national health technology assessment guidelines (ISPOR, 2022). Previous efforts have been made to clarify these categories, including the influential first and recent second panel on cost-effectiveness in health and medicine (also known as Washington’s cost-effectiveness panel), in which the recommendations are applied in many economic evaluation studies (Sanders et al., 2016; Weinstein et al., 1996).

The first panel pointed out that in the literature, costs were traditionally divided into direct and indirect costs. Costs directly attributable to the examining intervention are considered direct costs (Weinstein et al., 1996). The components of indirect costs still vary among the literature; some refer to it as the productivity loss or gain resulting from illness or mortality, and some also include the intrinsic value of health (also known as intangible costs of the disease). However, the experts in this panel agreed to keep the term “direct costs” but rename “indirect costs” to “productivity costs” to reduce the confusion with the “intrinsic value” of health (Weinstein et al., 1996). In addition, they claimed that the intrinsic value of health should not be included in measuring costs since it is reflected in the measure of health-related quality of life already.

The second panel after a decade (Sanders et al., 2016) came up with other recommendations, which are starting to be noticed by economic evaluation studies in recent years. Accordingly, they categorized costs as the following: (1) costs within the “formal healthcare sector” and (2) costs outside the healthcare sector (or “informal healthcare sector”). We will discuss these categories in detail and refer to the second panel as “the panel” from now on.

#### Costs within the formal healthcare sector

When examining costs within the formal healthcare sector, we consider consumption of goods and services that are required as inputs (Sanders et al., 2016). Given an example of a surgical treatment, the formal healthcare sector costs should include the costs of the surgery itself, i.e., medical equipment, medicine, and physician time, and the costs of complication if any case occurs. The panel listed some examples of costs within the formal healthcare sector, namely drug costs, medical devices, physician times and services, outpatient and inpatient care, etc. (Sanders et al., 2016).

This cost categorization is in line with the healthcare (or payer) perspective, which is mostly required in health technology assessment guidelines around the world (ISPOR, 2022). This is also one advantage of the updated costing recommendation introduced by the second panel.

There are two approaches to valuing the formal healthcare sector costs: micro costing and gross costing. Micro costing is a specific approach that normally comes from primary data at the individual level of each patient based on the amount of resources used for their treatment. Gross costing is normally based on electronic claim (or reimbursement) data, which focuses on cost at the provider level rather than the individual patient level. More details on these approaches can be found in the literature by Heslin et al. (2018) and Sanders et al. (2016).

#### Costs outside the formal healthcare sector

It is important to note that there are some costs that arise from the intervention but fall outside the formal healthcare sector. These costs should be considered in the societal analysis, but not the health sector analysis (Sanders et al., 2016). The panel divided these costs into three sub-categories: (1) time costs, (2) productivity costs, and (3) other sector costs.

Time costs include the time of the patient that undergoes the health intervention, for example, time spent during treatment, travel time, and waiting time. Time of informal caregivers (e.g., family members) should also be included in this sub-category. The panel recommended to use the net (post-tax) wage rate plus fringe benefits to value the time costs (Sanders et al., 2016). This is under the human capital approach, which will be mentioned in the next paragraph. They also recommended performing a sensitivity analysis with alternative forms of time valuation.

Productivity costs reflect the “production value of time” in society (Sanders et al., 2016). The panel emphasized that the QALY measure does not capture the productivity of an individual. Two main approaches used to value the productivity costs are human capital and friction costs. The human capital approach counts hours not worked in the labor market by one individual because of their treatment as hours lost, while the friction-cost approach takes the employer’s perspective and only counts for the hours lost before finding a replacement to make up the hours (Koopmanschap et al., 1995; van den Hout, 2010). One needs to be careful not to duplicate the time count in the working hours lost and aforementioned time costs. The panel mentioned one drawback of the friction-cost approach: The substituting person is assumed to be totally unproductive before the substitution, which might not be correct (Sanders et al., 2016). As a result, the panel recommended using the human capital approach to measure productivity cost, as well as time costs.

Other sector costs might also need to be considered in some particular health programs, especially public health interventions where the costs (or savings) may fall outside the healthcare sector. Some examples of these costs can be further explored in the additional reading provided for this course (Sanders et al., 2016).

#### Other cost categories

The panel mentioned three other cost categories: future costs, friction costs, and transfer costs (Sanders et al., 2016). In this course book, we will focus on future costs.

A treatment may reduce the productivity of the patient during its course, but after the treatment, the patient can return to work and thus the productivity is increased again. This gain in productivity is one example of future costs (or earnings). In addition, the treatment may lengthen the life of the patient, thus, some consumption costs may incur during those added years, which is another example of future costs. The first panel in 1996 believed that the future costs has been accounted for in QALY measures and thus should not be included in the evaluation to avoid duplication (Weinstein et al., 1996). However, the second panel pointed out that these future costs should be considered separately (Sanders et al., 2016). By that means, the net resource use can be calculated by

The recommended cost categories that should be included in cost analysis are summarized in the figure below, adapted from the second panel’s recommendation (Sanders et al., 2016).

Recommended cost categories in two perspectives

A picture containing graphical user interface

Description automatically generated

Source: Duy Pham (2022), based on Sanders et al. (2016).

Keep in mind that the categorization above is based on one recommendation (Sanders et al., 2016), and there are many ways to categorize the costs included in economic evaluation. Since this is an updated work from the first panel in 1996, it has not been used as widely as the first one. The most recent versions of many health technology assessment guidelines still use terminology similar to that of the first panel (direct and indirect costs; ISPOR, 2022; Weinstein et al., 1996). However, we would like to show this comprehensive work in order to encourage the use of consistent concepts in future economic evaluation studies.

### Self-Check Questions

1. Please name the two most common perspectives used in economic evaluation studies.

*healthcare sector (or payer) perspective, and societal perspective*

1. Please name the two dimensions that are used to differentiate the reimbursement mechanisms as told by Jegers et al. (2002).

*retrospective versus prospective systems, and fixed versus variable systems*

1. Please specify the function that can be used to calculate net cost, recommended by Sandes et al. (2016).

*(Healthcare costs + Non-healthcare consumption costs) - Productivity*

## 5.3 Types of Economic Evaluation

The purpose of health economic evaluation is to select the most cost-effective option from available alternatives to inform the policymakers making decisions in healthcare. Health economics itself contain two parts: “health” and “economics.” Under the basic concept of economics, the firm tries to maximize their profits by finding the most efficient way to reduce costs of production and/or produce more outputs. Health economics also contains the “health” component, which needs to be considered along with efficiency in costs of production. This is why the term “cost-effectiveness” is used in health economic evaluation to reflect both the costs and the benefits of a healthcare intervention. The comparison of the costs of achieving a given benefit for a few alternatives is known as cost-minimization analysis (CMA). However, since this approach only focused on costs, it is not counted as a full economic evaluation and it is not recommended anymore in economic evaluation (Drummond et al., 2015). In fact, the CMA can be observed via cost analyses for each alternative and their comparisons. Recently, cost-effectiveness analysis (CEA) has been used as a general term for full health economic evaluation, in which the benefits are measured in health-related units, for example, life year gained, cases of disease prevented, or QALYs (Sanders et al., 2016). Some others separate the term cost-utility analysis (CUA) to refer to the CEA that uses preference-based measures, such as QALYs or DALYs, as the unit of benefits (Drummond et al., 2015). Economic evaluation that assigns monetary values to health benefits is referred to as cost-benefit analysis (CBA).

We will use the categorization by Drummond et al. (2015) to discuss three key types of full health economic evaluations: cost-effectiveness analysis (CEA), cost-utility analysis (CUA), and cost-benefit analysis (CBA).

### Cost-Effectiveness Analysis (CEA)

In a narrow perspective, Drummond et al. (2015) use the term CEA to describe those studies valuing the health benefits less than the clinical outcome unit, which can include number of cases prevented, disability days saved, and life years gained when introducing a new healthcare preventive or treatment program. This method is normally conducted alongside clinical trials by observing their clinical outcomes; however, the clinical outcomes to be observed need to be chosen carefully depending on the type and purpose of health interventions (Drummond et al., 2015).

CEA has been dominant in the past, but it is less popular in recent literature, possibly due to the fact that many method guidelines recommend the use of QALY as the measure of benefits (Weinstein et al., 1996; Sanders et al., 2016; ISPOR, 2022).

### Cost-Utility Analysis (CUA)

Another type of economic evaluation, in which the benefits are generic measures of health, is cost-utility analysis (CUA). Some researchers refer to this as a smaller branch of CEA. The two most common measure of benefits in CUA are QALY and DALY. Due to the recommendation of using QALY as the measure of benefits in many national guidelines (ISPOR, 2022), CUA has become the most popular economic evaluation type in published literature.

### Cost-Benefit Analysis (CBA)

Cost-benefit analysis differs from other types of economic evaluation, as it evaluates benefits in monetary terms. The advantage of this approach is that it can compare programs in a broader sense, as not only health outcomes are considered. This method has a long track record in economic analysis of other sectors, for example, transport and environment (Drummond et al., 2015). In healthcare, the health benefits need to be converted to monetary value in order to perform this analysis, which creates various concerns. One common concept developed for this purpose is willingness to pay (WTP). Although the number of articles addressing this concept has been increased recently, many of them are still considered to be limited in terms of methodology (Olsen & Smith, 2001).

### Self-Check Questions

1. Please name three types of full economic evaluations, as categorized by Drummond et al. (2015).

*cost-effectiveness analysis (CEA), cost-utility analysis (CUA), and cost-benefit analysis (CBA)*

## 5.4 QALYs and the Cost-Effectiveness Threshold

Quality-adjusted life years (QALYs) are becoming the most popular outcome measure in health economics research, as well as in health technology assessment guidelines around the world. As mentioned, QALY is used for cost-utility analysis.

The indicator commonly used to compare the cost-effectiveness of intervention alternatives is incremental cost-effectiveness ratio (ICER), which is the ratio between incremental cost and incremental benefit when moving from one alternative to another. The unit of this indicator is cost per QALY (or life year) gained or, in general, cost per unit of outcome gained from moving to a new intervention. Cost-effectiveness threshold is a concept used to compare the ICER among different interventions, which is used to inform policymakers.

However, these concepts have received some criticisms. In the following, we will further discuss the limitations of these concepts.

### Equity Issues of QALYs

Although the QALY has become a popular measure of health outcomes and is recommended in many national guidelines (ISPOR, 2022), it still receives some criticisms, especially regarding equity concerns in decision-making.

Cost-effectiveness analysis considers the ratio of costs and effects. QALYs, used to measure health effects, are used as the denominator to calculate the incremental cost-effectiveness ratio (ICER). However, ICER does not reflect the size and distribution of QALY gain and may therefore create misinterpretation in priority-setting. For example, given two scenarios, (1) incremental cost of €1 million and incremental QALY of 10 and (2) incremental cost of €10,000 and incremental QALY of 0.1, they both have an ICER of €100,000/QALY gained. However, the size of costs and QALY gained differs. If the budget is allowed in this case (> €1 million), the decision is usually to go for the one with higher number of QALY gain (first scenario). The distribution of QALY gain also matters. In the first scenario, it is better if the ten QALY gain is distributed equally among four people than ten QALY gain for one individual. However, Olsen (2017) points out that an enormous health gain for one (e.g., two QALYs for one individual), might be better than miniscule health gain for many people (e.g., one-month QALY gain for each of the 24 individuals).

Second, a QALY does not reflect certain characteristics that might lead to prioritizing one individual over another, such as the severity of the disease. The social value on the disease severity has been shown in existing literature, where the health gain is valued two to ten times higher among those with higher severity (Dolan et al., 2005; Ubel et al., 1998). This characteristic is taken into consideration in healthcare resource allocation in some countries, including Sweden, Norway, and Australia (Stafinski et al., 2011).

Third, QALY does not reflect age. The concept of a “fair innings” was introduced by Alan Williams (1997), stating the importance of age differences in healthcare prioritization. He notions that “death at 25 is viewed very differently from death at 85” (Williams, 1997, p. 119). Accordingly, the older individuals have experienced their lives with a longer period of being in a better health, thus they should be less entitled to more health, than young people who have not experienced their fair innings of health yet. For example, given two individuals, one aged 25 and one aged 85, experience the same illness with the same possibility of death, and due to the limited resources, only one can be treated. The fair innings argument suggests treating the younger individual and claims that this person should receive the chance of health because they have not lived as long as the older individual. The consideration of age has been touched upon in some health economic evaluation methodology, especially under lifetime or age-specific modeling approaches. Williams (1997) also points out another characteristic, social class (or socioeconomic status), that should be considered in healthcare prioritization.

**Cost-effectiveness threshold**

This is the willingness of decision-makers to pay for a unit of health outcome (i.e., QALY gain and life-year gain; Drummond et al., 2015; Olsen, 2017).

### Cost-Effectiveness Threshold

**Cost-effectiveness threshold** is used to compare the level of cost-effectiveness of the health interventions and make decisions about which interventions should be prioritized within a limited health budget (Bertram et al., 2016). Some national guidelines have explicitly stated the threshold value (or range), such as the UK National Institute for Health and Care Excellence (NICE; Claxton et al., 2015). In many cases, if the threshold is not explicitly published in a health policy document, one might look into the previous decisions from relevant authorities to find the maximum ICER that has been accepted for a similar intervention and can therefore estimate the willingness to pay (WTP) for a unit of health outcome. Since health budgets vary based on context, the WTP for a unit of health outcome (i.e., cost-effectiveness threshold) also varies.

From the cost-effectiveness threshold, net monetary benefit (NMB) can be calculated using the formula where is the cost-effectiveness threshold, is the incremental effect (health benefit), and is the incremental cost. This NMB can be used as one indicator to develop the list from most to least cost-effective interventions (league table).

However, determining the cost-effectiveness threshold value is not a simple task, and it creates various debates in the field of health economic evaluations. In the following, we will discuss the popular recommendation of the WHO’s Choosing Interventions that are Cost-Effective project (WHO-CHOICE project) for estimating cost-effectiveness threshold based on gross domestic product (GDP; Hutubessy et al., 2003) and discuss the criticisms of this approach as pointed out in existing evidence.

#### WHO-CHOICE threshold recommendation

In 2001, the cost-effectiveness threshold based on per-capita GDP was introduced by the WHO’s Commission on Macroeconomics and Health. Later, in the WHO’s Choosing Interventions that are Cost-Effective project (WHO-CHOICE), the threshold is the cost per DALY averted when it is equal to three times the per capita income. This means those who have less than this threshold would be considered cost-effective (Hutubessy et al., 2003). Although this recommendation has been used as the most common threshold in literature (Bertram et al., 2016), some criticisms exist.

In fact, some countries in which the initiative of measuring cost-effectiveness threshold are followed recommend other multiplicative numbers of GDP. For example, the UK’s NICE uses the most relevant “central” cost-effectiveness threshold at £12,936, which is much lower than one GDP in the UK (Claxton et al., 2015). Thailand also set a threshold at only 0.8 of the national per-capita GDP per QALY gained (Thavorncharoensap et al., 2013). In addition, the use of DALY in measuring health benefits are limited, so those countries apply the threshold per QALY gained instead.

Another criticism of this recommended threshold by WHO-CHOICE is that it cannot reflect the feasibility of implementation regarding the affordability and budget impact of the investigated intervention (Bertram et al., 2016). One example is that although in Peru, the addition of trastuzumab would be cost-effective when accounting for the WHO-CHOICE threshold recommendation, it exceeds the entire budget for breast cancer in the country (Bertram et al., 2016).

Further reading on the limitation of WHO-CHOICE cost-effectiveness threshold recommendation can be found in the work of Marseille et al. (2015).

### Self-Check Questions

1. What is the cost-effectiveness threshold recommendation by WHO-CHOICE?

*cost per DALY averted when it is equal to three times the per capita income*

1. Please fill the following sentence with a right word.

*The concept of fair innings represents the importance of age differences in healthcare prioritization.*

## 5.5 Policy Implications

The key purpose of economic evaluation is to generate evidence-based information on the costs and effects of a health intervention to assist and improve the decision-making process and resource allocation in healthcare. Economic evaluation appears mostly before implementing changes in healthcare to inform decision-makers; however, sometimes it also appears during and after implementation. Respectively, it will contribute to the role of informing, monitoring, and evaluating the intervention. In clinical practice, the data from RCTs are the most reliable input of health outcome in economic evaluation, however, they normally need a long time to follow up. Some health interventions need to be implemented in a more urgent manner. For example, as a response to infectious disease (e.g., COVID-19), waiting for the full process of clinical trials before making a decision may not be appropriate.

In this section, we will start with a discussion on the importance of economic evaluations in implementation science, focusing on how to incorporate it into implementation strategies and decision-making processes. This is followed with some practical evidence on the application of economic evaluation in public health interventions in the U.S.

### Economic Evaluation in Implementation Science

The economic aspect of implementation strategies in healthcare may differ between related stakeholders, especially in places where the health economics concept is underdeveloped. For example, some physicians may discard the information on costs by reverting to the ethical obligations and believe that clinical outcome should be the sole measure for making decisions in healthcare. This perception is certainly an essential tenet in medical care; however, the amount of resources (budgets) available should also be considered. In this regard, we believe that policymakers should be informed by the work of a multidisciplinary team to obtain the evaluation from various perspectives before making any decision.

Substantial efforts have been made in recent decades to integrate economic evaluation in healthcare decisions, resulting in national guidelines in some countries for health technology assessment (ISPOR, 2022). However, many other countries have not established such guidelines yet, which may create difficulties in identifying the appropriate methods of economic evaluations for decision-making. Nevertheless, only once the methods for economic evaluation are agreed upon, can the comparison and prioritization among healthcare interventions be implemented consistently. Next, we will look at when and how economic evaluation methodologies provide substantial benefits for implementation strategies.

#### Limited health outcome data available

The economic evaluation is more impactful if there are limited data on a health outcome and the decision needs to be made in a relatively short period of time. RCT is the gold standard in clinical research as the most reliable evidence for assessing health outcome; however, it is time and cost consuming. Most health technology assessment national guidelines require a primary RCT study or systematic review of RCTs as the source for clinical effectiveness (ISPOR, 2022). In case of limited health outcome data available with an urgent need for decision, alternative data sources, such as a large-scale observational database, can be applied. However, the use of these data should be considered carefully due to their potential bias (Franklin et al., 2017; Rovithis, 2013). Over the years, the field of econometrics has been developed substantially; thus, many methods are proposed to deal with the selection bias of observational studies. Examples include matching, regression analysis, instrumental variables, and propensity scores. It is important to note that to address potential bias using the econometric method, the bias itself should be fully recognized, which is not the case in economic evaluation literature (Rovithis, 2013).

#### Mixed-method approaches

Economic evaluations are often based on quantitative data to estimate the cost-effectiveness of a health intervention, which can sometimes derive potential bias. In implementation science, additional qualitative approaches to these quantitative components help address the limitations (O’Leary et al., 2022). Using qualitative approaches, stakeholders of the health intervention should be interviewed, surveyed, or discussed in focus groups, etc., to collect their perspectives and additional inputs on the overall economic evaluations. Patients may perceive additional costs. A systemic approach to evaluate the costs and benefits of a health intervention in a specific context is recommended (O’Leary et al., 2022).

#### The use of decision-analytic model

Adding to the aforementioned mixed-method approach, additional inputs that come from qualitative studies should be incorporated in the decision-analytic model and its sensitivity analysis (O’Leary et al., 2022). Thus, a complete stimulation model can provide a context-specific output to inform the decision-making and implementation strategies. In addition, some additional analysis can be performed to discover equity issues by comparing subgroups in the population, referred to as distributional cost-effectiveness analysis (O’Leary et al., 2022).

Sensitivity analysis is used to handle the uncertainty in health economic models. It serves two main purposes: to estimate how important a particular parameter is in the analysis and to evaluate how the results change when adjusting the parameters and some assumptions. Consequently, the model aims to reflect context-specific parameters and outputs that should be used to inform the decision-making process.

### Application of Economic Evaluation in Public Health – The U.S.

The limited resource phenomenon is also present in public health interventions. Epidemiological data are normally used as essential evidence of public health issues. The role of epidemiology is to provide information about diseases by identifying the risk factors, incidences, and distribution with the objective of controlling them. Epidemiological data may only help detect the effects of the interventions and rank which interventions can deliver the best possible outcome for population health. However, in case of limited resources, the decision-makers need to consider the costs of the interventions. Economic evaluation is good to incorporate with epidemiology in this case to evaluate which public health interventions are efficient in terms of both costs and effects. As a result, economic evaluation is beneficial when prioritizing public health interventions in decision-making (Rabarison et al., 2015).

Beside epidemiology, economic evaluations of public health interventions are becoming important in informing decision-making in public health, as shown by evidence from the U.S. In 2012, the Committee for the Study of the Future of Public Health, Institute of Medicine in the U.S. made a call for action for the application of economic evaluation in public health, including the following (Rabarison et al., 2015, p. 3):

* Public health agencies at all levels should develop a model chart to better track the program outputs and outcomes to prioritize the funds.
* A robust research institute needs to be established to assess the effectiveness of public health interventions.
* Key components of public health delivery should be captured and measured systematically by obtaining data on relevant aspects, including implementation costs of the program.
* The methodologies for cost and effectiveness analyses should be developed and validated to compare alternative strategies.

From these recommendations, some efforts have been made in the U.S. to improve the application of economic evaluation in public health. For example, two research networks have been established: including the Center of Disease Control – Prevention Research Center Program and the Public Health-Practice Based Research Network. In these networks, the efforts of involving the consideration of costs, quality, and equity of public health interventions are explicitly considered (Rabarison et al., 2015).

### Self-Check Questions

1. What is the main purpose of sensitivity analysis?

*handling uncertainties*

1. Please fill in the blanks.

*By means of qualitative approaches, stakeholders of the health intervention should be interviewed, surveyed, or discuss in focus groups, etc., to collect their perspectives and additional inputs on overall economic evaluations.*

Summary

Economic evaluations are an important tool for decision-making in the healthcare sector. The more reliable they are, the stronger evidence can be generated. Therefore, we need to understand the concepts and measures to develop an appropriate economic evaluation. There are two critical components that need to be focused on: costs and benefits (or effects), meaning that the purpose of economic evaluations is to observe those interventions bringing relatively low costs and a substantial health outcome.

The measurement of costs and benefits vary based on context; however, some standard recommendations have been made in health technology assessment guidelines in some countries (ISPOR, 2022). For costs, considering both healthcare and societal perspectives is recommended in most guidelines and for benefits, QALY is the most popular recommended unit of measurement.

Various types of economic evaluations have been used, ranging from cost analysis to full economic evaluations in the form of cost-effectiveness, cost-utility, and cost-benefit analyses. The units of measurement used in these methods differ, especially on the benefit (outcome) side. Along with the discussion on the power of these economic evaluations, some debates on QALY measure and the cost-effectiveness threshold have been generated in research. The number of cost-benefit analysis studies is growing with the supporters of WTP, whereas cost-utility analysis is still dominant in the field.

The distance between research and policy implication needs to be considered, as researchers may need to find a suitable approach to inform policymakers effectively. Some context-specific models might be suitable and should come with sensitivity analysis to handle the uncertainties.

# Unit 6 – Health Econometrics

**Study Goals**

On completion of this unit, you will be able to …

… understand the concept of applied health econometrics.

… identify popular data sources used in health econometrics.

… understand the importance of data sources and study designs.

… describe techniques for causal analysis.

# 6. Health Econometrics

## Introduction

The effectiveness of an intervention can be evaluated by the changes it brings to the outcome of interest. To determine whether the changes are attributed to that intervention or other factors, statistical and/or mathematical analysis needs to be conducted. As a result, relationships between the cause and outcome of interest can be discovered. Depending on the field investigated, the methodology used is named differently, for example, psychometrics, econometrics, etc. Econometrics is statistical and mathematical methods used to investigate relationships in economics. Health econometrics is a branch of econometrics that focuses on the relationships related to health economics and interventions, for example, health insurance and health outcome.

In this unit, we will introduce the field of applied health econometrics and explore some popular econometric methodologies used to prove causal inference in health interventions. First, the role of data science in healthcare and health economics will be discussed. Popular data for applied health economics are also mentioned. Descriptions of different types of econometric models will follow with examples of their uses for various data types.

In the second section, methods for causal analysis will be the focus of discussion. Suitable study designs for causal inference will be further explored, including randomized control trials (RCTs) and quasi-experiments. For quasi-experiments, specific techniques in study designs used to build up the causal relationship will be discussed in more detail.

## 6.1 Introduction to Applied Health Econometrics

Generally, econometrics is a method of investigating economic problems that links three subjects (economics, statistics, and mathematics) to create the power of evidence (Fisher, 1941; Tintner, 1953). Applied health econometrics can be described as a branch of econometrics that uses statistical and mathematical methodologies to evaluate economic issues in healthcare. To develop an econometric model for evaluating health interventions, the availability and quality of data are essential. In this section, we will discover the role of data in science and common types of data used in applied health econometrics and their inference. Then, econometrics modeling, including linear and non-linear regression, will be discussed in more detail. Finally, the range of application of econometrics in healthcare will be addressed.

### The Role of Data Science in Healthcare and Health Economics

Data science can be simply understood as the extract of knowledge from data (structured or unstructured) and the application of the extracted knowledge explaining phenomena in different areas of interest. Hayashi (1998, pp. 40–45) define data science as a multidisciplinary field unifying “statistics, data analysis, informatics, and their related methods” to “understand and analyze actual phenomena” with data. The field of data science has boomed in recent years, especially the development of machine learning and big data.

In healthcare, data also play a critical role in creating evidence for practices. We can split the tasks of data science in healthcare into three categories, based on Hernán et al. (2019):

1. Data have a descriptive task, which provides quantitative summary of certain features. For example, by collecting the number of patients with a particular disease over time, we can calculate the prevalence and/or incidence of the disease in a population. An example in health economics is observing health insurance coverage in the population by collecting data on health insurance types among individuals.
2. Data can be used to predict patterns. For example, epidemiologists can use data to predict the risk factors of a particular disease. If some risk factors relate to socioeconomic status (i.e., income, employment, and education); insurance types; etc., it falls under the interest of a health economist.
3. Data can reflect the causal inference, which means that it can predict whether a factor (determinant) causes a particular output.

However, some requirements are essential for the establishment of causal inference. Causal inference is very important in healthcare and is used to identify the cause of a problem (e.g., a disease) in order to prevent it. An example is that in order to answer a question – does drug A work against COVID-19? – we need to prove the causal relationship between drug A and its treatment of COVID-19.

### Common Data Sources in Applied Health Econometrics

To build up an econometric model to explain the economic relationships of interest, we need to observe the data available for analysis (McPake et al., 2020). In the field of applied health econometrics, the majority of work relies on survey data from observational studies of a group of participants. In addition, even though the primary purpose of administrative data is not for research, these data are also applicable in many circumstances (McPake et al., 2020). In the following, we will discuss two most popular data sources used in applied health econometric literature: administrative and survey.

#### Administrative data

Administrative data are usually a population-based dataset with variables that are not primarily used for research. However, it can provide some relevant and useful data for applied health econometrics. One advantage of this data source is that it includes the characteristics of the full population of interest over time. Another advantage is that multiple administrative data can usually be linked by a common identification, which can create an even bigger dataset for research purpose (McPake et al., 2020). However, researchers need to be careful about ethical or data protection issues when linking these data sources. One disadvantage of this data source is that it may not be specific enough for the research purpose, which means it may not contain the variables of interest required for analysis (McPake et al., 2020). Some examples of administrative data include insurance claim data, reimbursement data, birth and registry data, and hospital discharge data.

#### Survey data

Survey data can be collected by asking participants to answer a set of closed (quantitative or observational studies) or open-ended questions (qualitative studies). In the field of health econometrics, quantitative data are more often of interest. Various designs of observational studies, including cohort, case-control, cross-sectional, ecological, case study, and mixed designs, are used to collect data for analysis. Key differences among these designs are the group settings and times of follow-up (at frequent intervals). By conducting observational studies, the variables of interest could be measured via the survey, thus providing a useful dataset for analysis. However, the common disadvantage of all types of observational studies is non-response (or missing) data, which are the items that are not filled or somehow missing from some participants. Missing data may or may not contain useful meaning in analysis. Handling missing data is an interesting topic, that is discussed in the literature by Kang (2013).

### Econometrics Modeling

After understanding comprehensively what data are available and their structure, the next step is to specify a suitable econometric model to explore the relationship of interest. First, some terminology should be clearly understood. For example, we may want to explore the relationship between insurance coverage and health outcome by hypothesizing that higher insurance coverage may result in better health outcomes for an individual. The insurance coverage is an independent or explanatory variable, and the health outcome is named as a dependent variable or outcome (Jones, 2007). The relationship can be illustrated as follows.

Causal Relationship in Econometric Analysis

Graphical user interface

Description automatically generated with low confidence

Source: Duy Pham (2022).

We should use the terminology in pairs: If we name X as an independent variable, Y should be named as a dependent variable, and if we name X as explanatory variable, Y should be named as an outcome variable. For consistency in this course book, we will use the pair independent and dependent variable.

In the following, we will describe the key characteristics of two types of regression models: linear and non-linear, as well as when to use them. The statistical explanation and coding used in these models are outside the scope of this book and can be explored further in the literature by Jones (2007, 2017).

#### Linear regression

Linear regression is a simple model reflecting a linear relationship between one or more independent variables and a dependent variable. The function of one independent and one dependent variable is as follows::

In this function, is the intercept, which is the predicted value of Y when X = 0; is the regression coefficient that represents the degree of impact of independent variable X on the dependent available Y; and is the error term, capturing all variation in Y that is not explained by X. Ordinary least squares (OLS) is the most common method use to estimate the parameters of a linear regression model (i.e., ) based on the practical data value sets of X and Y. Indeed, this method helps generate the minimum error term () in the model.

Linear regression relies on some assumptions. These assumptions and methods for checking assumptions can be discovered in the literature by Casson and Farmer (2014) and McPake et al. (2020). In the field of applied health economics, sometimes these assumptions are violated, which leads to the use of other suitable models (McPake et al., 2020; Jones, 2017). For example, healthcare cost and expense data are usually right skewed, which may not satisfy the assumption of normal distribution of the residuals; thus, the linear regression cannot be applied (Jones, 2017). One solution suggested is transforming the data by taking logarithms, then running a log-linear model instead (Jones, 2017). However, the transformation needs to be carefully done since it may be confusing when doing interpretation. Furthermore, in some cases, log-transformation problems may happen due to the special characteristics of data, such as data with a significant number of zero observations (Jones, 2017).

#### Non-linear regression

In health economics, survey and administrative data are the most common data sources, which means that many health outcomes are not measured as continuous variables, but rather binary; multi-nominal; integer counts; and duration (e.g., time to death) variables. Non-linear regression models are more suitable for these types of variable; thus, non-linear regressions are more common in health economics (Jones, 2017). In the following, we will briefly describe the popular non-linear model based on each type of dependent variable.

First, we will discuss which models are most suitable for an outcome variable that has a discrete number of choices available. If the dependent variable is a binary variable, which means the answer can be only one of two options (e.g., yes or no questions), the most common regression model is logit (assuming a logistic distribution of error) and probit (assuming a normal distribution of error; McPake et al., 2020). For a multi-nominal outcome, in which the choice can be made between a number of options that are not in order (e.g., questions like: “to which hospitals did you go for the surgery?” where the answer would be a hospital name), the multi-nominal or conditional logit model is normally applied. Other flexible models, such as nested logit and multi-nominal probit, can also been used in some cases (McPake et al., 2020). In case of an ordinal outcome, where a choice can be made between naturally ordered options (e.g., “How do you rate your health today?”, with answer options like “Very good,” “Good,” “Fair,” “Bad,” and “Very bad”), the ordered probit model is normally applied.

Second, if the outcome variable is measured as a count of events (e.g., “How many times have you visited the doctors in last three months?” where the answer should be an integer number equal to or larger than 0), the classical model used is a Poisson regression (McPake et al., 2020). Other regression models for count data can be found in the work of Cameron and Trividi (2013).

Third, many health outcomes in applied health economics are measured using the time elapsed before an event happens (e.g., time to death). For this particular type of outcome, the most common model used is the Cox regression model with the application of hazard or time-to-event function (McPake et al., 2020).

In general, depending on the availability and types of data, suitable regression model(s) can be chosen for analysis. After making a choice, assumptions of each model type need to be considered carefully during the analysis. The detailed application of each mentioned model is outside the scope of this course book and is explained in the literature by Jones (2017).

### Self-Check Questions

1. Please define the concept of econometrics.

*Econometrics is a method of investigating economic problems that links the three subjects – economics, statistics and mathematics – together to create the power of evidence.*

1. Please name the two most popular data sources for applied health econometrics.

*survey and administrative data*

1. Please name the two types of econometrics modeling.

*linear and non-linear regression*

## 6.2 Methods for Causal Analysis

In this section, we will focus on the research designs and methodologies used to verify the causal relationship in applied health economics. Causal analysis can prove the impact of a health intervention on a health outcome of interest, which creates strong evidence for the effectiveness of the intervention, and thus brings reliable information for decision-making. However, to prove a causal analysis, a variety of conditions and assumptions have to be fulfilled. In the first part of this section, we will explore the study designs that can be applied for causal analysis, including randomized control trial (RCT) and quasi- or natural experiments. RCT, is a gold standard for proving a causal relationship; however, consumes many resources and sometimes faces ethical issues. Therefore, quasi- or natural experiments, can also be used for causal analysis with a careful consideration of econometric techniques and assumptions.

### Research Designs for Causal Analysis

There are many research designs to answer research questions. While some questions may not need complicated research designs to be answered, others do. According to Evans (2003), the level of evidence not only depends on the study designs but also the focus of the research question, which is related to the effectiveness, appropriateness, or feasibility of healthcare intervention. Evans (2003) develops a hierarchy of evidence in these two dimensions and claims that RCT is not a gold-standard design in every case. In fact, RCT is a gold-standard only when evaluating the effectiveness of interventions. If the research questions are beyond the effectiveness, a different approach might be needed. In this section, we will discuss in which situations the RCT is necessary to determine a causal relationship, and in what situations other approaches are also considered appropriate.

#### Randomized control trials

RCTs are common in the field of clinical science and in economic evaluations of medical treatments (e.g., drugs, medical procedures, and new medical technologies), normally known as micro-economic evaluations. RCT is an experimental trial where participants are allocated to an experimental or control group randomly and receive different treatments (Kendall, 2003). These two groups will be followed to see the differences in outcome. The concept seems to be simple; however, a well-designed RCT is not easily planned. Kendall (2003) points out the features of a well-designed RCT:

* The study sample needs to be selected appropriately and be consistent with the hypothesis so the results are appropriately generalizable. The sample size needs to be sufficient to detect important differences between groups.
* A good randomization technique is needed to minimize the confounding variables and selection bias.
* Except the tested treatment (intervention), experimental and control groups are treated identically.
* The investigator is blinded to which group the individual is assigned and to which group the treatment is allocated.
* Subjects are analyzed within their allocated group, either with intended or unintended (blinded) intervention.
* Analysis should purely focus on answering the research question; either the result is significant, or it is not.

Double-blind, which means that neither the subjects nor the investigators know which treatment the subjects are receiving until the trial is over, is also applied in many RCTs to prevent bias when the investigators analyze the results. However, this strategy is not always feasible, for example, in surgical intervention.

RCTs are less common in behavioral and social science due to the scope of the interventions. For example, some public health interventions, such as tobacco control or tackling obesity, may apply some concepts of behavioral and social science to design their implementation strategies (Public Health England, 2018). These interventions have too broad a range of activities from different levels of stakeholders that might be too complicated to design RCTs; thus, RCTs are likely infeasible in this case. However, there is at least one enormous RCT in relation to macro-health economic: the RAND Health Insurance Experiment in the U.S. between 1971 and 1982 (Brook et al., 2006).

#### Quasi- or natural experiments

In RCTs, the investigators have the intention of fully controlling the variables to see the difference in outcome. A well-designed RCT needs to treat the groups identically, and the only different should be the intervention of interest to see the differences. This design is sometimes unethical. As an example, we cannot stop breastfeeding a controlled group of children to see the effectiveness of breastfeeding since it may affect their lives later. Moreover, although RCTs are useful for causal analysis in a relatively small environment of laboratory experiments, it becomes difficult for large scale interventions because it can be time consuming and expensive.

Instead, quasi-experimental designs are more common in applied health economics, as they allow the researchers to answer questions in a more feasible way. The key difference between RCT and quasi-experiments is the randomization: quasi-experiments do not allocate subjects randomly to different groups for comparison. In fact, the researchers have less control over the intervention in quasi-experiments than RCT. In natural experiments, which is one type of quasi-experiments, although some random assignments are possible, it is not considered true randomization, and is therefore not a true experiment. Natural experiments are often conducted when introducing a policy or law so researchers can investigate its effects (McPake et al., 2020). One example of a natural experiments is the Oregon Health Study, which is a landmark study exploring the relationship between public health insurance expansion in the U.S. and well-being, healthcare utilization, financial strain, and health outcomes of low-income adults (Baicker et al., 2013; Finkelstein et al., 2012; Taubman et al., 2014). In this natural experiment, the researchers study the effects of the public health insurance by assigning the eligible low-income people into the program based on a random lottery, since they could not afford to test all eligible ones. This design also helps to reduce the ethical issue of selection since people are chosen via a random lottery.

With the growth of real-world data and evidence, data can be linked together to create an enormous retrospective data source, which is an advantage for conducting quasi-experiments to build up better evidence for decision-making. Indeed, with the development of econometric methodologies, these data sources can also be used in causal analysis to some extent.

### Techniques for Causal Inference of Quasi-Experimental Designs

In this section, we will focus on the four quasi-experimental designs that can be applied for causal inference, following the recommendation of Kim and Steiner (2016):

* regression discontinuity (RD)
* instrumental variable (IV)
* matching and propensity score (PS)
* comparative interrupted time series (CITS)

#### Regression discontinuity design

Regression discontinuity (RD) design is when the assignment of an intervention group is determined by a fixed threshold of a particular factor, which itself may be correlated with the dependent variable (the outcome). Any discontinuity in the outcome at the cut-off value of the included factor can be interpreted as a causal effect of a treatment. One common factor used as a cut-off point in applied health economic studies is age. One example is a Japanese natural experiment study on the cost-sharing mechanism in healthcare based on age (Shigeoka, 2012). In Japan, after reaching the age of 70, the cost-sharing in healthcare drops from 30 percent to ten percent. This study uses individual-level survey data to discover the causal inference of this cost-sharing mechanism on health outcome and healthcare utilization. The results show that although the mechanism helps to reduce healthcare expenditure among the elderly above 70 years of age, there is no significant effect on self-reported health outcome (Shigeoka, 2012).

By means of RD design, a major validity threat is the assignment of the factor and its threshold. Other potential drawbacks are noncompliance, generalizability, and statistical power (Kim & Steiner, 2016).

#### Instrumental variable design

Instrument variable (IV) is an exogenous variable that satisfy two assumptions (Kim & Steiner, 2016):

1. It is associated with the independent variable (treatment).
2. It is theoretically unrelated to the dependent variable (outcome), except going through the independent variable.

To elaborate, the instrumental variable only links to the dependent variable when considering the independent variable as a mediator. Given that, the strategy for causal inference is to see the variation in dependent variable, explained by the IV. However, in practice, the identification of IV is a challenge, since in theory, assumption one mentioned above can be tested but assumption two empirically cannot (Newhouse & McClellan, 1998). It can be claimed that there may be another exogenous variable linked with both IV and the dependent variable without being related to the independent variable. Some other practical issues of IV design are pointed out in the work of Kim and Steiner (2016).

#### Matching and propensity score design

Next, the matching and propensity score (PS) design is particularly helpful when investigators have a lack of control over intervention but a good knowledge of the potential confounders that may determine the effect of the intervention on health outcome. By knowing a list of potential confounders, causal inference can be theoretically achieved with matching or PS design. The matching activities will help cluster the most similar groups for comparison, at least in terms of identified confounders, for causal inference analysis. Sometimes, matching by hand is difficult, although the number of confounders is normally small when using this technique; therefore, the use of PS is helpful (Rosenbaum & Rubin, 1983). The PS is defined as “the conditional probability of receiving the treatment given the set of observed covariates” (Kim & Steiner, 2016, p. 9). One of the most common challenges of matching design is selecting the potential confounding variables, while other possible practical issues can be found elsewhere (Kim & Steiner, 2016).

#### Comparative interrupted time series design

So far, we have discussed the designs that meet some of the requirements, from assignment control to the knowledge of exogenous and endogenous variables that are related to the investigated relationship between independent and dependent variables of interest. If none of these requirements exist, an alternative design is comparative interrupted time series (CITS) design. This design contains its own two obligatory requirements (Kim & Steiner, 2016):

1. The outcome has been measured in multiple time points (time series) for all subjects.
2. The intervention group has been interrupted by an intervention.

Some concerns about this design include the appropriate length of time between series and the need of autoregressive models for accounting for serial dependency structures (Kim & Steiner, 2016).

### Self-Check Questions

1. Please name the two types of study designs that can be used in causal analysis.

*randomized control trial and quasi-experiments*

1. Please name one technique that can be used to prove causal relationship in quasi-experimental studies.

*regression discontinuity, instrumental variable, matching and propensity score, and comparative interrupted time series*

Summary

Applied health econometrics have been developing substantially in recent decades to investigate the effectiveness of healthcare interventions. We have discussed the popular types of data available in the field, including survey and administrative data. These data sources in combination with clinical data create a concrete basement for applied health econometrics and economic evaluations. While RCT study designs are essential for micro-economic evaluations in proving the causal inference of a medical technology (e.g., treatment and medicines) on health outcome, quasi-experiments are more commonly used for causal inference at the macro-economic level, such as health insurance or health policy.

Econometrics models need to be appropriate and suitable for the data, resources available as well as the research questions. Different types of techniques for causal inference for quasi-experimental studies were covered in this unit. However, further understanding outside the scope of this book is needed to apply them appropriately. It is important to note that, in many cases, RCTs are not feasible because they consume extensive resources (time and cost) and may entail some ethical issues. Thus, the use of other study designs should be implemented with careful consideration. With the development of real-world data, researchers have more feasible choices for study designs in the field of health economics.