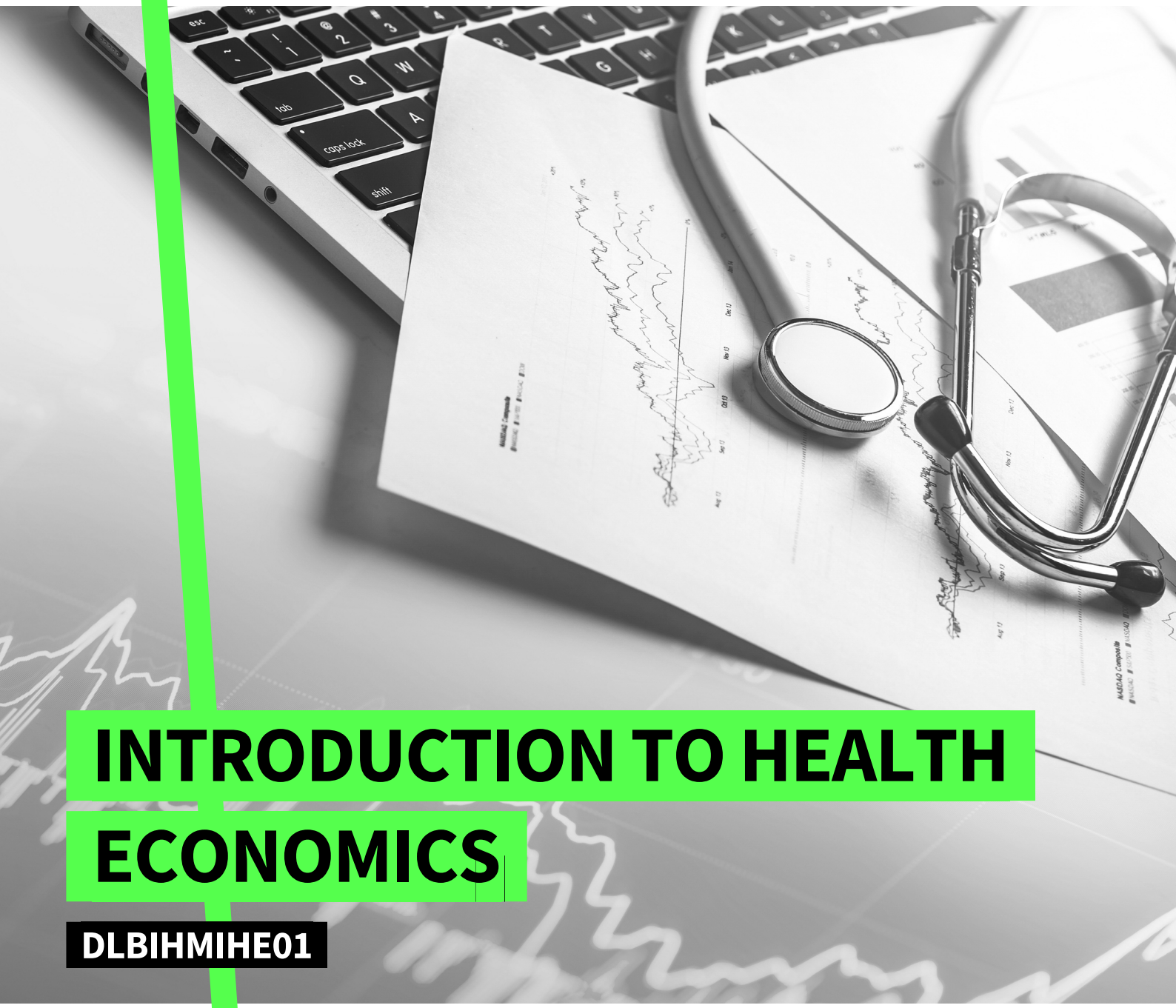
Course Book



INTRODUCTION TO HEALTH ECONOMICS

DLBIHMIHE01

iu
INTERNATIONAL
UNIVERSITY OF
APPLIED SCIENCES

INTRODUCTION

LEARNING OBJECTIVES

While health itself cannot be traded on a market, products and services that produce health are traded on a number of different markets. Competition between providers and different provider payment methods (e.g., fee-for-service, capitation, and pay for performance) create financial incentives that influence provider behavior. Pre-payment systems are common (e.g., tax and public and private health insurance), meaning that healthcare is rarely fully paid for by the consumer at the point of use.

Severe market imperfections, such as information asymmetries between patients, providers, and payers, mean that traditional economic models are of little to no use when analyzing health financing and healthcare markets. Several mechanisms – mostly in the form of financial incentives – have been devised to help correct these imperfections from a single actor perspective. Furthermore, these failures warrant government intervention to correct them from a public health perspective.

This **Introduction to Health Economics** will enable you to describe and discuss the analysis of economics, the production of health, and the financing and production of healthcare in terms of efficiency.

UNIT 1

HEALTH, ECONOMICS, AND HEALTH ECONOMICS

STUDY GOALS

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

- use economic concepts to analyze health production in relation to efficiency and market conditions.
- identify the main drivers of costs in healthcare.
- understand the dynamics of market forces and failures in the healthcare sector.
- recognize the role of the government in the production of health.
- analyze health as an economic good.
- critically reflect on the approach taken by economics regarding the determinants of health.

1. HEALTH, ECONOMICS, AND HEALTH ECONOMICS

Case Study

One evening, Olga decides to go shopping for some ingredients to host a dinner for a couple of friends. Being a budget-conscious student, she carefully looks through all the different brands and sizes of each of the products with the goal of choosing the best ingredients while saving as much money as possible. While shopping, she recalls that her friends enjoy drinking a particular brand of wine and picks up three bottles. She also remembers that she is running low on detergent and dish soap and decides to buy them. However, she realizes close to the register that if she takes all the extra items she will not be able to afford some of the necessary ingredients for her recipe. Reluctantly, she leaves two wine bottles behind and buys everything else she needs without going over her budget.

The next day, Olga checks her receipts and reflects on how this type of trade-off decision applies to almost everything in life. She ponders what would happen to the world if everyone could always buy everything that they wanted and decides to read about how the market works. As a public health student, she wonders how much these choices and dynamics apply to the healthcare sector.

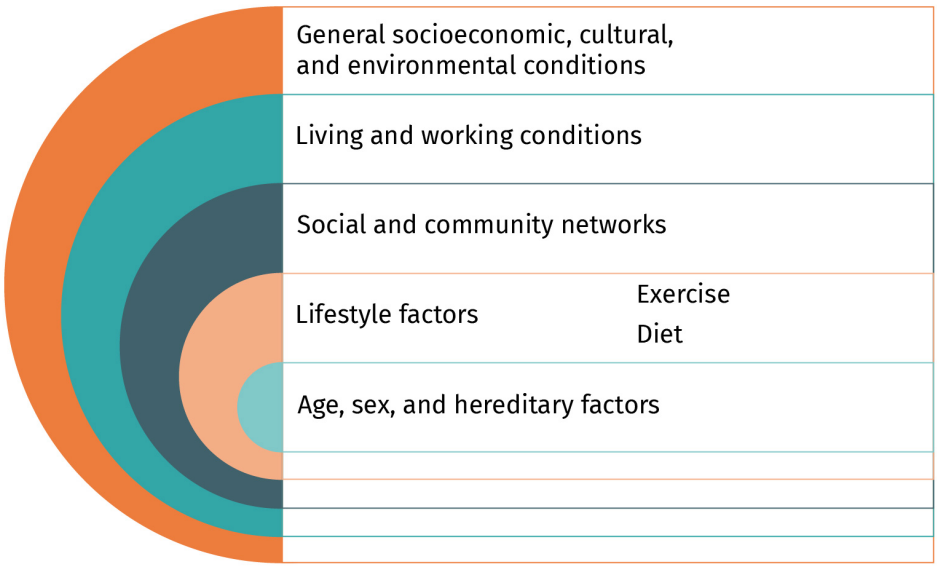
1.1 The Demand for Health and Healthcare

Health was defined by the World Health Organization (WHO) constitution of 1948as “a state of complete physical, mental and social wellbeing and not merely the absence of disease or infirmity” (p. 1). Humans intrinsically value maintaining physical and mental abilities while avoiding or alleviating diseases since it allows individuals to adequately cope with the demands of daily life. This implies that fluctuations of health can manifest in a variety of ways and at various stages of life, depending on when and by whom they are experienced. For example, a woman of fertile age trying to have children experiences a health context that is vastly different to that of a man in his late seventies. While the former is interested in securing safe and affordable care to ensure the best chance of having healthy children, the latter is more likely to be invested in managing potential chronic diseases and maintaining independence.

Health is determined by the life continuum, as well as many other contextual factors, as visualized in the figure below. Dahlgren and Whitehead (1991) describe the relationship between an individual, their environment, and disease. While the innermost layer is composed of factors we cannot modify (such as our genes), the rest of the layers are made of

influences on health that can be modified, such as exercise and diet choices; presence and composition of family and community networks; and broader socioeconomic, cultural, and environmental factors (Dahlgren & Whitehead, 1991, as cited in Jinks et al., 2010).

Figure 1: Main Determinants of Health I



Source: Sergio Flores (2022), based on Dahlgren & Whitehead (1991).

Sometimes the terms health and healthcare are used interchangeably. Nevertheless, it is important to draw a distinction: Health is the state of complete wellbeing and is the ultimate goal, whereas healthcare is the set of tools, services, and actions that improve or preserve health. Healthcare is the pathway to health. There is no inherent value in healthcare by itself; however, the two concepts become codependent when we talk about health need and demand, which can only exist as a function of available healthcare.

Health Need and Demand

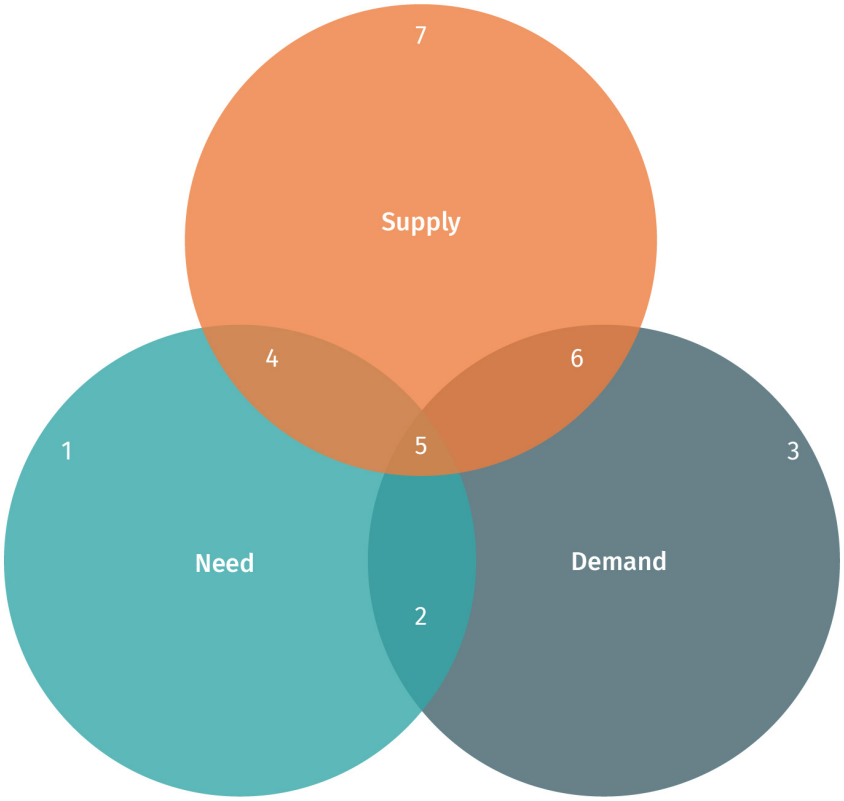
Health need is succinctly defined as the capacity to benefit from healthcare or wider environmental changes (Wright et al., 1998). These health needs can be either met or unmet, and that is mostly determined by the supply and demand of healthcare. Healthcare supply can be understood as the availability of resources needed to cover health needs, such as qualified personnel, facilities, or medications.

The demand for health refers to what patients ask for. The interactions between the existence of health needs and the supply and demand of healthcare lead to several scenarios that are always present in the healthcare sector and are vital to understanding the peculiar dynamics of the healthcare market. These interactions are illustrated using the **Venn diagram** below.

Venn diagram
This is a diagram style consisting of overlapping circles to illustrate the

relationships between different variables.

Figure 2: Health Need, Supply, and Demand Venn Diagram



Source: Sergio Flores (2022), based on Santana et al. (2021).

Using the Venn diagram above as a guide, (Allin et al., 2010) distinguished the following five types of unmet need.

Unperceived (by the patient)

This is the type of need that the patient does not know they have and is therefore unable to report. A professional could potentially recognize this sort of need because of their training, but since the patient is unaware of it, they do not attempt to address it. Consequently, the health need goes unmet because the patient does not realize it is present. A case of asymptomatic early-stage cancer is an example of such a scenario. Space one in the diagram includes unmet and unperceived needs.

Chosen (informed)

This is the type of unmet need that is the result of a patient's personal, informed choice. For example, a patient could be sick and fully cognitively aware but still choose to avoid any kind of treatment. This could be the case of patients reaching the end of their life and choosing not to continue with uncomfortable treatment that might only extend their life by a short time. Space one in the diagram also includes unmet needs by choice.

Unchosen

This is the type of unmet need that happens as a result of factors out of the patient's control and is more related to healthcare supply. Some of these factors could include a shortage of available healthcare providers, difficult access to providers due to long commuting distances and/or high travel costs, and long waiting lists. Space two in the diagram includes these supply-constrained needs.

Clinician validated

This type of unmet need happens when the patient cannot obtain the healthcare they demand and see a need for (that the professional healthcare community would also agree on). Therefore, the individual's need is (at least partially) unmet. An example of this is if a clinician commits some kind of malpractice or negligence to a patient aware of their health need. This type of unmet need is found in space two of the diagram.

Subjective unmet expectations

This type of unmet need results from the patient feeling that their health need was not met by the healthcare professionals. This unmet health need is subjectively perceived as such by the patient and can thus be considered a demand. This type of unmet need falls into space three of the diagram.

As in the case of health needs, there are also three different types of demands exemplified in the previous figure and explained by Santana et al. (2021):

1. Need-based demand refers to demand for healthcare that is backed up by a healthcare need, corresponding to spaces two and five.
2. Unnecessary demand is represented in the Venn diagram by spaces three and six. This category indicates demand that is not based on need. If care is provided even though no need is present, it is represented by space six. If care is not provided when demand not based on need is present, then it is represented by space three. Space three represents demand that is visible in some form but is not based on need and does not result in more healthcare utilization. A clinical consultation appointment prompted by a desire for social interaction rather than a medical requirement is one example of this. Space six involves demand for healthcare services that are not based on need, such as unnecessary follow-up dental or outpatient appointments.
3. Avoidable demand can occur for a variety of reasons:

a) A need initially goes undetected, resulting in demand at a later stage of the illness, for example, when a person is diagnosed with late-stage cancer and needs surgery or chemotherapy that could have been avoided if the cancer had been caught earlier.
b) Some healthcare demand may be preventable if it is caused by behavioral risk factors. These include a sedentary lifestyle, smoking, and substance abuse, which may trigger conditions like coronary heart disease, type-2 diabetes, or lung cancer.
c) Displaced demand is also potentially avoidable. 3a and 3b cases are represented in space five in the Venn diagram. Proactive preventative care (space four), as well as other types of early intervention for unmet needs (space one), might result in the effective transfer of cases out of space five or could at least lower the share of resources needed to treat them.

1.2 Health Production: Efficient Use of Resources

Health economics is the study of all resources, activities, and institutions within the healthcare sector that are involved in producing goods or providing services. Within any economy, resources are the inputs used to produce outputs. These are often categorized as factors of production by mainstream economic theory:

- labor, which includes all human resources
- capital, which includes all goods that are used in the production of other goods, such as machines, factories, and equipment
- land, which usually refers to natural resources, like water or wood

When these resources (inputs) are combined to produce something, we call this process production. Good health is what we ultimately want to achieve as an outcome of health production; however, it can be hard to measure and define, with many composite measures proposed, such as life expectancy, disability-adjusted life years (DALYs), and quality-adjusted life years (QALYs). Therefore, intermediate outputs are often used to measure production and supply in healthcare (i.e., births attended, fractures treated, and home visits provided).

Utility
This is the total satisfaction received from consuming a good or service.

Because inputs (resources) are always limited, decisions about where to allocate them must be made. These decisions must prioritize allocations to create outputs that provide the greatest **utility**. For each output that is successfully created, there are several unrealized potential outputs, which are the trade-offs (Guinness & Wiseman, 2011). The production function of healthcare is a tool that allows us to visualize how different outputs can be achieved while using a certain combination of inputs. Consequently, it also lets us see which amount and type of inputs produce the most outputs (Guinness & Wiseman, 2011).

Imagine you work in a hospital and have to decide how many surgeons to hire (variable input) to produce a set number of surgeries (output). Surgeons are not the only inputs needed; supporting staff, operating rooms, and equipment shall be treated as fixed variables (but only for the sake of this example). Imagine the case as described in the table below.

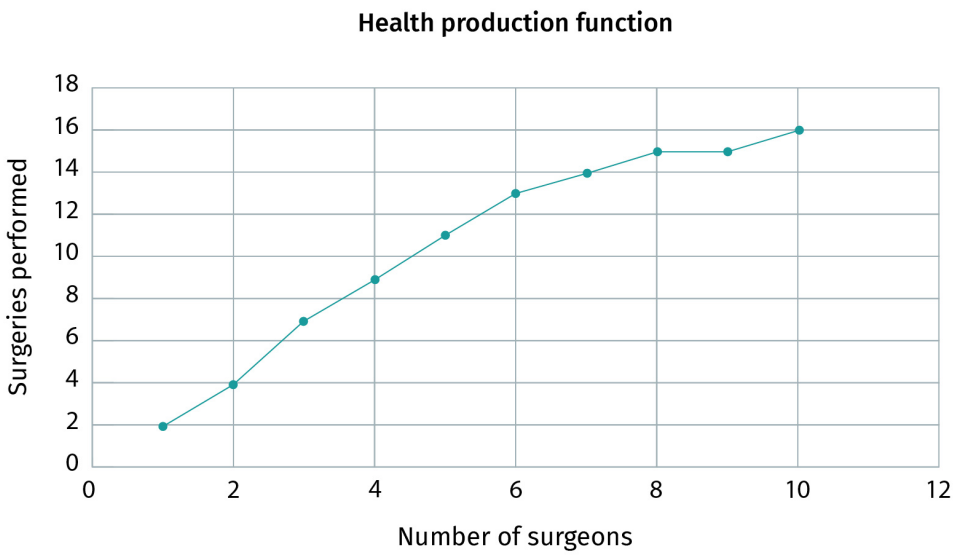
Table 1: Production Function for Healthcare Table: Hospital Example

Number of surgeons	Total surgeries performed	Marginal output	Total surgeon cost	Total fixed costs	Total cost	Average cost
1	2	2	50	200	250	125
2	4	2	100	200	300	75
3	7	3	150	200	350	50
4	9	2	200	200	400	44.5
5	11	2	250	200	450	41
6	13	2	300	200	500	38.5
7	14	1	350	200	550	39.3
8	15	0	400	200	600	40
9	15	0	450	200	650	43.3
10	16	0	500	200	700	46.7

Source: Sergio Flores (2022).

When plotting inputs versus outputs on a graph, we obtain the production function, which is shown in the figure below.

Figure 3: Health Production Function for Healthcare Plot: Hospital Example



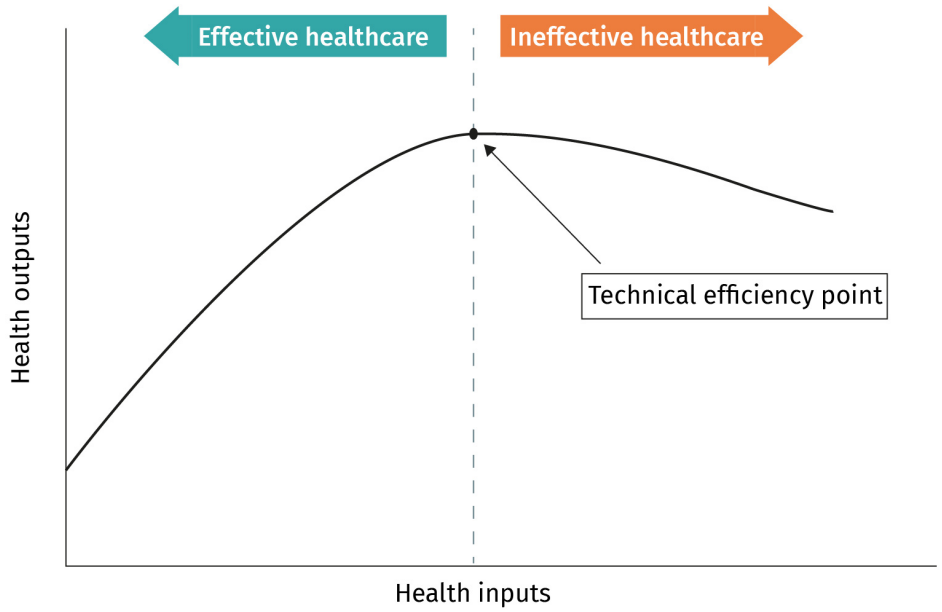
Source: Sergio Flores (2022).

Marginal analysis
This is the analysis of whether the costs of engaging in more of a particular unit of action can produce enough benefits to compensate for the costs.

When facing these kinds of decisions, we make use of **marginal analysis**, which is an evaluation of the extent to which each additional unit of consumption or production of something yields further benefit or incurs greater loss (Guinness & Wiseman, 2011). As observed in this example, our first surgeon can perform two surgeries. As we add more surgeons, we can produce more surgeries. The first additional surgeon hired produces two more additional (marginal) surgeries, and hiring a third would produce three more surgeries. The total cost for the hospital administrator rises correspondingly, but the average cost per surgery dramatically drops from 125 to 50 units per surgery with three surgeons.

If the production function followed a linear trajectory, we could assume that the more surgeons we hired, the more outputs we could obtain at a lower cost. However, in practice, this is not the case. Real-world scenarios have several constraints that would not allow that; in our example, the amount of supporting staff available for surgeries, the number of operating rooms present at the installations, the availability of pharmaceuticals, and the demand for surgeries are all factors that could reduce the number of outputs compared to inputs. The figures above show how the marginal benefit of each surgeon decreases, while the marginal cost stays the same or even starts to increase as more surgeons are hired. So, how many surgeons should be hired? As health economists, we want an optimal mix of inputs and outputs, ergo, the most efficient one. The situation in which at least one more input is required for a producer to create more output is referred to as technical efficiency (Guinness & Wiseman, 2011). The figure below illustrates this.

Figure 4: Technical Efficiency Plot



Source: Sergio Flores (2022).

The initial marginal outputs of this health production function are high, plotting the curve into an upward trajectory. As we add more inputs into our production function, the curve seems to flatten out, reaches its highest point, and ends up following a downward slope. This highest point is the point of technical efficiency where a maximum capacity to benefit is obtained. There are scenarios in which too many input units can actually cause detrimental health outputs. For example, the prescription of too many medications or a liberal use of surgical treatment can result in side effects for patients and create a health deficit. As the figure above shows, the use of inputs before reaching the technical efficiency point is considered an efficient use of resources, while any use of inputs beyond this point is deemed ineffective.

1.3 The Costs of Healthcare

Healthcare resources are all personnel, materials, infrastructure, earmarked accounts, and anything else that can be used to provide healthcare services. These can all be inputs in the production of health outputs (and, by extension, outcomes). We can divide healthcare costs into three main categories: human resources, physical capital, and consumables.

Human Resources

The most significant inputs into the health system are human resources, which include the various types of clinical (physicians, nurses, pharmacists, and dentists) and non-clinical workers (management and support staff) who make each health intervention possible.

Human resources is frequently the largest single expense within healthcare. In many nations, labor costs can account for two-thirds or more of total recurrent expenditures. The healthcare sector is labor intensive and requires qualified and experienced staff to function well.

Physical Capital

This is infrastructure and technology/equipment in the healthcare system. The material basis on which care is delivered is provided by physical resources. We can further divide this into three broad subcategories:

1. Buildings/structures with auxiliary facilities (i.e., energy and water systems)
2. Medical equipment (i.e., diagnostic laboratory equipment and radiological machines)
3. Logistics (i.e., supply systems, transport, warehouses, and their logistic facilities)

Consumables

These are items that are used for a short length of time and must be replaced on a frequent basis. This category includes pharmaceuticals and disposable (one-time) equipment and other supplies.

Healthcare Cost Variations Among Countries

The costs of all these inputs varies significantly across countries for many reasons, including demographic characteristics, health system characteristics, workforce and structural capacity, health utilization, and pharmaceuticals.

Demographic characteristics

The population structure of a country, based on parameters such as age; gender; and proportion of overweight, unemployed, drinking, or smoking population, can significantly alter the epidemiological characteristics of the population and, consequently, the type of resources that each country has to purchase to maximize health welfare. The demographic characteristics of a country also largely determine the epidemiological needs of the population.

Health system characteristics

The type of service provision, health financing, and provider payment mechanisms greatly influence the overall health costs. For example, countries like Sweden, Canada, and the United Kingdom with single-payer systems in place (where the government is the single actor that buys healthcare on behalf of everyone) are able to negotiate or establish lower, more uniform costs due to the volume of patients the government represents. In contrast, the US has a fragmented multi-payer system (with multiple public and private payers) that has less bargaining power on behalf of health consumers, making healthcare more expensive. In the same way, payment mechanisms also affect healthcare costs. For example, if

healthcare providers are paid for each service they provide (fee-for-service payment mechanism), there is an incentive to overprescribe healthcare, making it inefficient and more expensive overall.

Workforce and structural capacity

The amount and concentration of the health workforce in relation to the country's population is often the biggest source of cost for many countries. Furthermore, the variation of the types of skills health workers possess plays a significant role. Some countries have significantly higher numbers of professionals working in intensive units or specialized hospital care than others that emphasize the importance of nurses and general practitioners (GPs). In others, much of the workforce profile is geared towards primary care or even relies heavily on community health workers and nurses. Additionally, the wages healthcare workers obtain are often a direct function of the income per capita a country reports.

Health utilization

The amount and extent to which health services and health resources are consumed by the population is a major driver for health costs. Annual hospital discharges, use of imaging and laboratory services, average length of stay, etc. serve as proxies to measure this factor.

Variation of technological use in medical practice

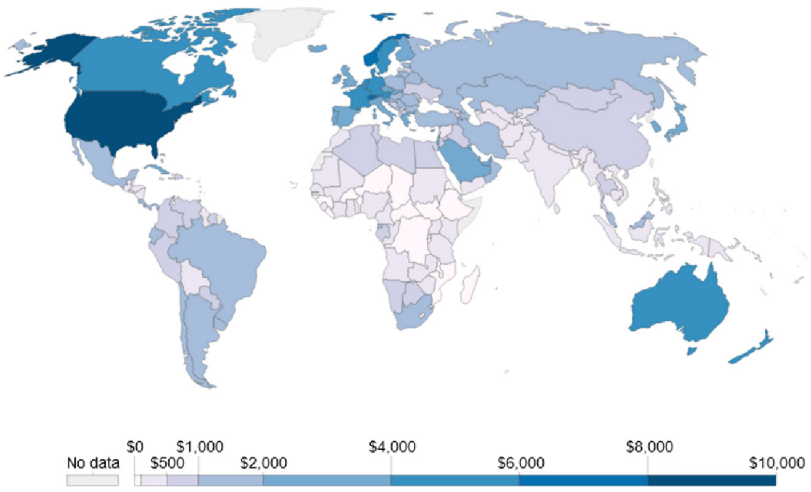
Technological progress is widely considered an important driver of health costs. The complexity of the interventions and the type and amount of equipment required to perform them vary greatly among countries.

Pharmaceuticals

The price, availability, and coverage of pharmaceuticals varies across countries and is mainly driven by the type and source of the drugs purchased.

In general, the variations of costs related to healthcare use are the result of either variations in the prices of goods and services, the volume of care provided, or both (Health and Europe Centre, n.d.). As seen in the figure below, even countries that share similar overall economic conditions manifest variations in their healthcare costs, with the US being the clear outlier. In this particular case, the US spends almost twice as much as ten other high-income countries despite performing worse on many population health indicators (Papanicolas et al., 2018). We can observe that the characteristics of the US healthcare system (multi-payer system) result in a very complex structure that requires high administrative costs. Additionally, healthcare workers in the US have considerably higher salaries on average than in many other countries (Tijdens et al., 2013) and the prices of both drugs and tests are considerably higher there than anywhere else (Mulcahy et al., 2021).

Figure 5: Country Variation in Healthcare Costs



Source: Our World in Data (2020). CC BY 3.0.

1.4 Health and the Market

Market
A market is a place where buyers and sellers can meet to facilitate the exchange or transaction of goods and services.

Economists study the decision-making process humans engage in when facing scarcity. Markets are one of the ideal scenarios where this happens. A **market** is a situation in which suppliers of goods and services meet with consumers who want – or demand – those goods and services and agree on a price to purchase them. The benefit that an individual gains from consuming these goods or services is called a utility, and the more utility a consumer expects to receive from their purchase, the higher the price they are willing to pay. When we pool the utility that all individuals in a society experience, we obtain the welfare (Guinness & Wiseman, 2011).

Supply
This is the willingness and ability of producers to create goods and services to take them to market.

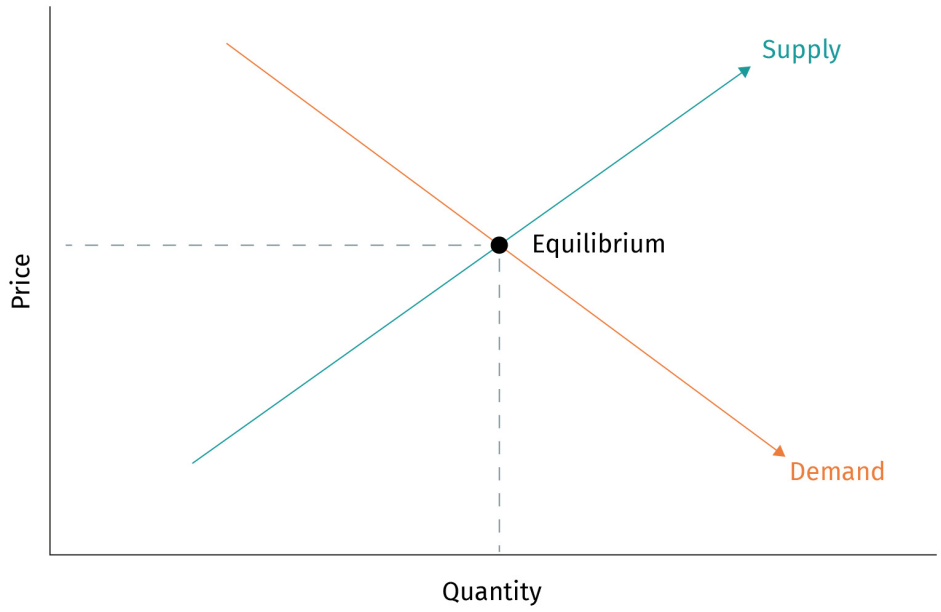
Demand
This is the consumer's desire to purchase goods and services and willingness to pay a price for a specific good or service.

In a market, **supply** and **demand** from producers and consumers push and pull each other to determine the prices of products and services that will be exchanged. Supply refers to the goods and services that producers are willing and able to sell in the market, and demand includes the goods and services that consumers are willing and able to purchase in the same market. The dynamic between the interests of buyers and sellers in this market is known as market force of supply and demand. Demand from the buyers stems from their interest in obtaining as much utility as possible (utility maximization) at the lowest possible price. Supply from sellers originates from their interest in maximizing profit by producing at the lowest possible cost while selling at the highest possible price (Guinness & Wiseman, 2011).

A way to visualize this is a tug of war between supply and demand. When supply wants to increase their output price too much (pulls the rope), demand will simply buy less of it (pulls the rope back) and supply will have to reduce prices. If demand asks supply to pro-

duce a significantly larger amount of product, then supply will tug back by increasing prices. This process leads to a state of equilibrium in which the amount demanded equals the quantity supplied, as seen in the figure below.

Figure 6: Market Equilibrium: Supply and Demand



Source: Sergio Flores (2022).

If a state of market equilibrium is reached, then we should observe a state of efficiency from both the producer's and the consumer's point of view. In this equilibrium, goods and services are produced at the lowest possible unit cost for the benefit of the producer, and consumers perceive that they are obtaining the most possible utility out of their money within their budget. From both a consumer's and a supplier's perspective, money is not wasted in either production or consumption. **Pareto optimality** is how economists characterize this situation. Given their resources, everyone is at their best feasible welfare level (Mwachofi & Al-Assaf, 2011). Most economists would argue that the ideal scenarios of free markets like the ones explained above are the best placed to promote efficiency. Necessary requirements for perfect free markets to produce efficient allocations include

Pareto optimality
This is an economic state where resources cannot be reallocated to make one individual better off without making at least one individual worse off, implying efficiency but not equality or fairness.

- the number and size of "actors" (i.e., producers and consumers) in the market. If you have few suppliers, prices can be set among them without taking consumer demand into consideration (Eastin & Arbogast, n.d.).
- the ease with which "actors" can enter and exit the market. For example, if a consumer has restricted access to the market through barriers, such as long waiting times, long distances, or extremely high prices, they do not have ease to enter the market and demand may not apply the right effect on the price (Eastin & Arbogast, n.d.).

- the degree to which producer outputs are differentiated from other producers.
- the price and product information available to both buyers and sellers. If a buyer has knowledge of the product, then they can set a price they are willing to pay for a good or service. In the same sense, if an insurance provider does not know about the health risks of a population, they may set prices that are too low to make insurance sustainable for everyone.

Other aspects to consider include the following (Mwachofi & Al-Assaf, 2011):

- how easy it is to advertise the goods and services in the market
- how steady and predictable demand from sellers is
- the reliability of the quality and amount of goods from suppliers
- how easy is it for consumers to test a good or service before purchase
- access to uniform information between buyers and sellers
- standardized prices for the exact same product regardless of the buyer
- that all market suppliers have a profit motive

All of these requirements are very hard to meet in the real world, but they provide a navigation guide to how equilibriums are influenced by market forces.

The Special Case of the Healthcare Market

The healthcare market differs from an ideal market in several ways. The special characteristics of this market are explained below.

Information Asymmetry

Information asymmetry exists if one party has information that the other lacks. On the one hand, suppliers (doctors, physical therapists, nurses, etc.) have vastly more knowledge of diseases and their treatments than patients. Therefore, even if patients look for a second opinion, they ultimately have to trust a healthcare provider about how much healthcare to consume. This creates an unusual circumstance in which the supplier of goods and services is also setting the demand, which can lead to a market failure.

On the other hand, and in a similar fashion, healthcare providers make decisions on how much healthcare to supply based on information provided by the consumer. A healthcare professional relies on a patient's previous history and account of their current illness to chart a treatment plan. In the case of healthcare insurance organizations, they require information from the consumer to determine the most efficient production costs, which should ultimately lead to a good price in the market (Mwachofi & Al-Assaf, 2011). The two following failures derive from information asymmetry.

Adverse selection

Adverse selection is the exploitation of information asymmetry. For example, people who are less healthy might identify that their costs when covered by health insurance are much lower than without it since their use of healthcare is significantly higher than those who are healthy. Therefore, they have an incentive to sign up for insurance without disclosing

information that would make healthcare providers adjust their costs. Insurance providers may then boost premiums after unexpectedly seeing a rise in their costs in order to avert losses. More expensive premiums can result in healthy people deciding that the cost is no longer providing them enough utility in return and therefore leaving the scheme, leaving only the less healthy and possibly leading to the market's collapse (Mwachofi & Al-Assaf, 2011).

Moral hazard

A moral hazard is a situation in which someone will take risks because they will not be affected by the cost that they could incur. Individuals who are aware that their healthcare costs are being subsidized by other people are likely to consume more healthcare (even if it is not warranted) or take risks that could be detrimental to their health that would not be taken if they had to pay the full cost of healthcare themselves. As a result, healthcare resources are not efficiently allocated based on health need but on artificially created need (Mwachofi & Al-Assaf, 2011).

Externalities

Externalities are the effects of consumption or production that have an impact on people who do not participate in the transaction; they can be either positive or negative. Positive externalities materialize when one person's actions in the market have a positive effect on an individual that was not directly part of the transaction (i.e., herd immunity caused by vaccination of multiple individuals), while negative externalities have a negative effect on another person (i.e., secondhand smoke from tobacco users). Because spill-over effects are not evident to either the producer or consumer, they are frequently overlooked in decision-making. As a result, the consumption or output level chosen is inefficient or ineffective (Mwachofi & Al-Assaf, 2011).

Consumer Rationality and Ability to Make the Best Judgments About Their Welfare

Even if conditions allow them to do so, consumers looking for healthcare are not always in a position to make the best decisions regarding their health. To start with, they may not have enough information about their illness or how to treat it, as opposed to how they might approach purchases in a different market. Furthermore, intense stress as a result of illness makes it hard for an individual consumer to make an informed decision. Moreover, consumers are unable to precisely foresee the outcomes of healthcare consumption (Mwachofi & Al-Assaf, 2011).

Interdependent Demand and Supply Determination

Increased demand for healthcare (due to an influx of people or an epidemic, for example) can lead to increased prices. In this scenario, as a result of the price rise, a physician may now be willing to provide fewer hours of service. This exemplifies how healthcare supply and demand are not set independently, resulting in market failures (Mwachofi & Al-Assaf, 2011).

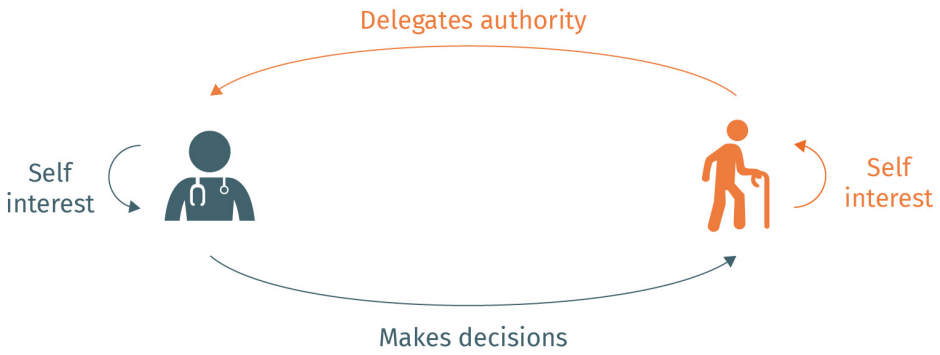
As you can see from the previous examples, healthcare markets do not have the conditions of a perfect market and thus cannot behave as a perfect economic market would. Many of the necessary requirements for a perfect market simply cannot be met, and if left alone, will derive into multiple market failures and inefficiencies. To counter this and keep markets as close to Pareto optimality as possible, markets require intervention in the form of regulation from actors other than suppliers and consumers.

1.5 Supplier-Induced Demand and Agency

One of the ways to overcome the information asymmetry between patients and physicians can be explained with the agency relationship.

Agency Relationship

Figure 7: Principal-Agent Relationship in Healthcare



Source: Sergio Flores (2022).

The principal (patient) appoints an agent (healthcare provider) to counsel them in making decisions through recommendations and information they lack. The principal then combines the advice given by the agent with personal preferences to make decisions as if perfectly informed. However, it is more likely that a decision is made when an agent combines information with a principal's preferences; that is, doctors make decisions for patients (Nguyen, 2011). This can lead to several special considerations from the agent: Should the agent maximize patient or societal utility? What if the patient is unable to communicate or be part of the decision-making process? Furthermore, as the agent is usually also a supplier, this leads to a situation where one actor (the agent) is both the demander and supplier in the market.

Supplier-Induced Demand

Because of the aforementioned scenario, supplier-induced demand is a very common scenario in the healthcare sector. In a normal market, a patient would use their own judgment to demand a certain amount of healthcare based on open and available information and knowledge, whereas in this scenario, an excess demand is created based purely on the supplier's own knowledge, which is never entirely privy to the patient.

Induced demand is described as a change in healthcare demand caused by a provider's discretionary influence over patients, especially coming from medical doctors. Even when patients cover the full costs, induced demand obstructs the efficient deployment of state resources. This situation may alter the supply-and-demand balance in the healthcare sector. Induced demand presents two economic challenges: First, it increases healthcare expenses while also putting a burden on government resources. Second, because a larger amount of a country's resources is spent on healthcare with low benefits, it has an important influence on efficiency (Seyedin et al., 2021). Additionally, from a health perspective, induced demand from healthcare providers does not always reflect health gains for the patient. For example, a 1985 report by the WHO suggested that no health benefits were identified when cesarean section rates exceeded 10–15 percent, and this was confirmed by more recent studies (Althabe et al., 2006; WHO, 1985, 2021b). We also know that, as with any kind of surgical intervention, cesarean sections carry inherent unavoidable risks. Even so, when private obstetricians in the Indian region of Madhya Pradesh were paid to deliver babies in 2016 with higher amounts paid for cesarean sections, the rates of cesarean delivery increased from 26.6 to 40.7 percent (Bogg et al., 2016).

Naturally, since physicians do not always put the needs of others ahead of themselves, some measures have been designed to avoid the double agent problem (in which the roles of principal and agent are both filled by the physician), such as a professional self-regulation/ethical code, standardized clinical guidelines and protocols, and incentives to affect provider behavior through efficient monitoring and policing.

1.6 Market Failure and the Role of the State

Healthcare markets are prone to market imperfections if unregulated and subjected to the forces of the free market, thus throttling its efficiency. Since healthcare is a type of good that will always be consumed because it is necessary for our survival no matter the price, an unregulated market would inexorably drift towards higher costs and fewer health gains. Because it lacks the conditions a perfect market needs, it would not be able to self-regulate. Furthermore, no one would be incentivized to invest in health initiatives that benefit payer and non-payer patients, such as vaccines or certain types of medical research.

Imagine the following scenario: You live in a town called Freeville where, for the last ten years, the previous administration did not allow government regulation in the healthcare sector. In Freeville, a very popular chain of fast food called "Bad Burger" has introduced

extremely cheap food that is very dense in sugars and fat. In fact, they have introduced a new soda beverage of 64 ounces (half a gallon) in their most popular kid's menu. Most people see this option as a great deal and embrace it quite quickly, and Bad Burger becomes the biggest business in town with ads everywhere inviting kids to eat. Eventually, it also becomes the provider of lunches in schools for kids, as it is cheaper and saves people money. Soon after, the biggest park in Freeville, which had a children's playground, running and biking tracks, and a large area of forest, is bought by Bad Burger and transformed into a soda factory for the chain. As a result, childhood obesity rates have shot up. As a resident of Freeville, you wonder why the market isn't fixing this issue itself and realize that the conditions for a perfect free market are not always present in the healthcare sector.

These imperfections are arguably best addressed through the market intervention of an external force. Governments are usually the best placed actors to intervene, considering that they are the only actors with enough power to set parameters for all actors in the healthcare market. A government theoretically represents the best interest of all stakeholders and is able to consider a broad and longer-term perspective. The free market, however, will not factor in health and the need for health services as a fundamental human right and an indispensable foundational condition to any other human activity. The main ways a government intervenes are as follows:

- informing consumers, providers, and suppliers that they must act in a certain way. For example, the use of cigarette labels warns consumers of health hazards incurred by consuming the product.
- regulating how a private activity may be undertaken. For example, the government may pass laws limiting how much pollution a factory may produce.
- financing healthcare with pooled funding, for example, from tax revenue or employee/employer contributions
- providing or delivering health services using publicly-owned facilities and civil service staff, for example, by building, maintaining, and staffing public health centers
- taxing and subsidizing goods, for example, taxing alcohol (making it more expensive and harder to obtain) but subsidizing vaccines (making them free and easier to access)

The following two types of goods that the private free market is not designed to incentivize are provided/enabled by the government and play vital roles in maintaining the health of a population:

1. Public goods. These are goods that produce large amounts of **positive externalities** and are provided to the population at large. These include actions such as initiatives to control populations of insects that could transmit diseases, vaccination campaigns, and investment in research.
2. Merit goods. These are goods that are considered beneficial to the individual or society regardless of their ability or willingness to pay. These include subsidized education, use of helmets and seatbelts, and fire departments.

Positive externalities
This is a trait of a good or service that, when consumed, has a positive effect on (or benefits) a third party that is not the consumer or producer.

SUMMARY

In this unit, we defined health as a state of complete physical, mental, and social wellbeing and not merely the absence of disease or infirmity; health need as the capacity to benefit from healthcare or from wider and environmental changes; and health demand as what patients ask of the healthcare sector. We then identified human resources; physical capital (which includes infrastructure, equipment, and logistic networks); and consumables as the biggest categories of inputs in healthcare. The volume and the price of these inputs determine how healthcare costs vary per country around the world.

The healthcare market diverts significantly from what a theoretical perfect market should be for several reasons, mainly the information asymmetry existing between suppliers and consumers. This propitiates the existence of phenomena such as moral hazard and adverse selection; the frequent presence of both positive and negative externalities; the absence of consumer rationality in many scenarios where health need is present; and a particular interdependent supply and demand determination, with a disproportionate influence coming from the supplier. In many instances, these lead to the phenomena of supplier-induced demand. Because of this, the healthcare market suffers from critical imperfections that require government intervention to function appropriately.

UNIT 2

FORMS OF DELIVERY OF MEDICAL CARE

STUDY GOALS

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

- understand the role and influence of healthcare providers in the patient's care pathway.
- identify different provision schemes aimed at modifying healthcare provider behaviors.
- recognize the evolving trends in healthcare delivery.

2. FORMS OF DELIVERY OF MEDICAL CARE

Case Study

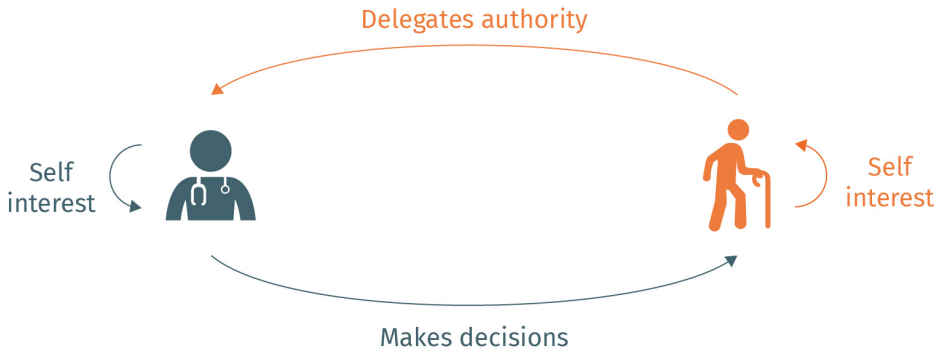
Olga has been feeling pretty good after her dinner with friends. However, a few weeks later, she starts experiencing regular abdominal pain and cramping. She decides to visit a doctor but is uncertain which specialist she should book an appointment with, when to do so, or what kind of treatment she should expect to receive. She ultimately decides to go to her family doctor, who refers her to a gynecologist. The gynecologist carefully explains the different treatment options available, but Olga finds it hard to decide for herself and ultimately decides on following the treatment the physician recommends, even if it is more expensive than she would like.

While leaving the office, she wonders how hard it is for her to look for, select, and purchase appropriate healthcare treatment compared to most other things in her life. For example, whenever she needs a maintenance check on her car, she has a general idea of what the car might require and can choose from different mechanic shops. She is also free to choose where to buy the needed car parts, or even if she is willing to replace them. Enthralled by this, she decides to enroll in a health economics course to understand the underlying mechanisms.

2.1 The Principal-Agent Relationship as the Key Problem

In a perfect market, consumers have a high degree of independence and power over the direction of supply and demand. They are the ones that decide the amount and price of goods they are willing to consume, partly because both prices and product/service information are available to both buyers and sellers beforehand and on a comparable scale. Even if that was not possible, the market allows consumers to test the product or service prior to consumption.

Figure 8: Principal-Agent Relationship in Healthcare



Source: Sergio Flores (2022).

In the healthcare market, one of the most important characteristics distinguishing it from a perfect market is the asymmetry in information between a consumer (patient) and a supplier (healthcare professional). Patients are often not familiar or updated enough with all the nuances of pathologies and the wide array of treatments available to them. Therefore, they are in a weaker position to make an informed, rational decision about their health.

Because of the scenario just described, supplier-induced demand is a very common scenario in the healthcare sector. This translates to a demand greater than what would be necessary if patients and physicians had the same level of knowledge. When a provider prioritizes other interests at the expense of the patient's interests, a principal-agent problem is created (Nguyen, 2011). In most cases, this creates a higher demand than if the patient alone made the decision.

To exemplify this kind of relationship, try to imagine the same dynamic occurring in a market that is not healthcare. Imagine you are an owner of a mechanic shop that fixes cars and sells spare parts. A person with a car that is experiencing mechanical problems comes to you. It is in your best interest to retain the customers and try to sell them your services. In fact, if we saw this scenario solely from your interests as a provider of car repair services, the best possible scenario is to sell this costumer as many services as possible to maximize your utility, even if the car does not require immediate repairs. If the customer knows enough about the mechanical upkeep of cars, they might directly tell you the services they need and will then decide if your prices are fair. If the customer's and your interests do not align, the customer will either explicitly tell you what they need or go to another shop. If, however, the customer does not know anything about car repair (i.e., information asymmetry), then they will probably rely on your advice. For all intents and purposes, you are playing both the role of the supplier and consumer, which might lead to a conflict of interest.

These dynamics between potentially conflicting interests were observed in India in 2016. When the government tried to increase the hospital delivery rates in the regions of Gujarat and Madhya Pradesh, two different approaches were implemented: In Madhya Pradesh,

private obstetricians were paid to deliver babies, with higher rates paid for cesarean sections than vaginal deliveries. In this region, cesarean rates increased from 26.6 to 40.7 percent. For Gujarat, obstetricians were paid for each block of 100 deliveries, regardless of the type of delivery. In this case, cesarean sections decreased from 8.1 to 4.3 percent (Bogg et al., 2016).

2.2 The Physician as a Supplier of Medical Services

When a patient decides to demand healthcare following a perceived health need, a complex process of mutual decision-making and organization of care is started. Physicians play a key role in this process. For example, they can establish a relationship with the patient to the point that they can influence the patient's demand for health, playing a dual role as supplier and demander of healthcare. Additionally, physicians act as important gatekeepers for specialized medical care and the supply of drugs. They also play a major role as researchers, policymakers, and even commercial actors within the healthcare sector. Their presence throughout the care process is constant and highly influential.

A literature review and metanalysis (Stewart, 1995) provided evidence that just the presence of good communication between physicians and patients has a positive effect on a patient's health outcomes. Even when physicians are not actively taking part in the patient care, their effect on the patient's behavior can be felt. For example, the amount of healthcare providers that choose to establish themselves in a certain area (thus increasing the density of health professionals in an area) tends to increase the healthcare utilization of consumers. How? A larger amount of health professionals reduces waiting times for patients and leads to higher competition, which can bring prices down. It could also mean an increase in healthcare centers, which translates into shorter travel distances. All these "barriers" to the consumer are reduced, which facilitates (and tends to increase) consumption.

Even when patients are facing life threatening scenarios, such as prostate cancer, evidence suggests that physician recommendations are the most likely determinant of the treatment choice over patients' preferences or worries about side effects, such as diminished sex life (Scherr et al., 2016). Similar studies involving other diseases, such as pneumonia, gangrene, cancer, and asthma, show similar results (Adams et al., 2001; Sekimoto et al., 2004). Sekimoto et al. (2004) found that only 12 percent of patients would prefer an active role in decision-making, and Adams et al. (2001) found that, on average, patients with severe asthma do not wish to be predominantly responsible for decision-making about asthma treatment. Considering the huge influence that physicians have as health suppliers, healthcare provision schemes aim to modify healthcare provider behavior to align with the patient's best interests.

Figure 9: Physicians as Gatekeepers to the Patient Care Pathway

Source: Sergio Flores (2022).

When a patient has a health need and demands healthcare that is accessible to them, they go through several stages of care, including first contact with a health professional, testing and diagnosis, subsequent consultations with more specialized healthcare personnel, treatment of the disease, and (sometimes) palliative care. A patient would find it difficult to navigate all this alone. It is physicians who determine the best path for the patient by performing diagnostic examinations, referring them to other professionals, and providing necessary treatment. These decisions are, in turn, backed up by evidence-based guidelines stemming from research and health legislation, which are spheres in which physicians and other healthcare professionals also partake. It is clear that physicians can have a very strong influence on the patient care pathway.

Because of this unbalanced relationship between patient and physician, several mechanisms attempting to compensate for this have been proposed. Mainstream economic theory establishes that the use of financial incentives can lead to different types of healthcare provider behaviors, each with their own strengths and limitations. Some of them are listed below.

Capitation

The provider agrees to supply a predetermined list of health services to a predetermined group of people for a set fee per person and period. When the real cost of these services exceeds that specified level, the provider takes a financial risk. In contrast, when the cost is less than the predetermined reimbursement, the provider keeps a portion of the money. The most common concern about capitation is that the provider organization receiving the payment may overly restrict service use, potentially removing certain essential services alongside unnecessary ones. As a result, patients may receive lower-quality care.

To illustrate this with an example, if you are a healthcare provider under a capitation payment mechanism, you sign an agreement stating (in a fictional setting) that you will be covering 1,000 patients in a community. The monetary terms are as follows.

Table 2: Capitation Monetary Terms Example

Subpopulation	Number of patients	Capitation per month (US dollars)	Total
Children	350	$30	$10,500
Women	250	$25	$6,250
Men	200	$20	$4,000
Seniors	200	$30	$6,000
Total	1,000		$26,750

Source: Sergio Flores (2022).

For each subpopulation (children, women, men, and senior patients) you offer to provide a specific set of services dependent on their needs. Ultimately, in this example, you as a provider will get 26,750 USD per month to cover the entire population. If you spend less than this amount, you can keep the difference. However, if you exceed it, you as a provider bear the financial risk and must cover the difference.

Fee-for-Service

Healthcare providers are paid for each service they deliver to a patient. Assuming more medicals services benefit patients, fee-for-service payments would be presumed to improve quality of care and patient outcomes. More care, however, does not necessarily imply higher quality or better outcomes (Fisher, 2003; Smeets et al., 2009; Tsugawa et al., 2017) Therefore, this method could incentivize overprovision of health services or prioritize fee-paying patients.

To illustrate, let us assume you are the healthcare provider for a clinic subscribed to a fee-for-service payment scheme. If a patient came with a fracture and you managed to treat it with a cast, you would get paid for each service provided: the emergency visit, radiological

studies, costs of materials used to immobilize the limb, and medications prescribed to the patient. If a surgical approach was required, then a higher reimbursement fee would be given since the costs for you are higher.

Salary

The healthcare provider is paid a set sum for a set length of time. There is no financial motive to provide unnecessary services or underprovide. Because there is no financial incentive for providers to give high-quality treatment when they are paid on a salary basis, payers often rely heavily on the implementation of regulations and processes that are supposed to improve quality.

To illustrate, let us assume you are a healthcare provider under a salary payment mechanism. In this case, you get paid a fixed amount every month, no matter the volume or complexity of services you provide.

2.3 Managed Care and Alternative Forms of Provision of Care

Imagine you are in charge of the health centers of a rural region in Honduras. For many years, people with low income have struggled to improve their health. Clinics are always understaffed and the population is underserved. Only wealthy people have access to the private doctors, while the rest are struggling. To improve things, you decide to employ the help of private healthcare providers to improve your population's health. You read that one way forward is through managed care.

Principles of Managed Care

In an attempt to keep costs low for patients (consumers), managed care plans were invented. Managed care plans combine finance and delivery of healthcare services that aim to keep costs low and reduce the unilateral influence of healthcare providers, designed to manage cost, health utilization, and quality via the use of guidelines and protocols; strict administrative controls; and focus on certain diseases, such as chronic pathologies. However, the extent to which this integration occurs varies greatly, ranging from health maintenance organizations (HMOs) in which the health plan subscribes a limited network of physicians to preferred provider organization plans, which generally negotiate a discounted fee-for-service price schedule and provide access to a broad provider network (Glied & Smith, 2011). Various tools are used to influence care and/or costs across the managed care spectrum. These include utilization review/management systems, changes in how physicians are compensated, and provider network restrictions (Glied & Smith, 2011), which are explained in more detail below.

Provider network restrictions

Managed care organizations find the most cost-effective providers by comparing the prices charged by various hospitals and practitioners for the same procedures. When such organizations are large enough, having hundreds of thousands of subscribers, they can opt to bring specific providers into their network, thereby procuring discounts for the provision of health services to their members in healthcare marketplaces with multiple providers. To attract more patients, providers are willing to offer steep discounts to these organizations. By offering these reductions, providers can also keep or gain market share (Glied & Smith, 2011).

Changes in physician compensation

Managed care organizations are constantly trying to keep costs down for their members; to do so, they are constantly persuading both patients and physicians to choose cheaper healthcare options via economic incentives. For example, managed care organization may require preauthorization from patients before they access hospital emergency rooms or specialized outpatient care, or they may make it harder for patients to seek treatment at more expensive facilities. Costs are also controlled by capping physician salaries, setting them at the outset, and allowing annual adjustments (Glied & Smith, 2011). Furthermore, physician compensation comes from a mix of salary, productivity bonuses, and other factors to keep it attractive to healthcare professionals (Ryan et al., 2015; Darves, 2011).

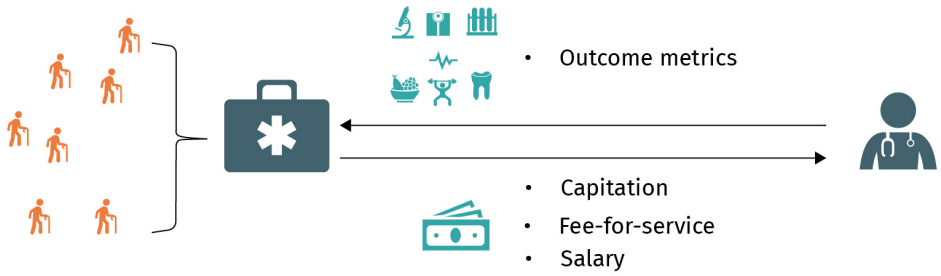
Utilization reviews and management systems

Many managed care organizations have created complex information systems that track provider prices and the quality of healthcare obtained by their members in order to conduct **utilization reviews** (Glied & Smith, 2011).

Utilization reviews
These are processes in which the patient's care plan is carefully reviewed on a case-to-case basis, usually against a set of guidelines to make sure the patient is receiving appropriate care at minimal costs.

Managed care plans work by spreading financial risk and through strict healthcare management, that is, detailed records of patient's health consumption; very careful administrative control of expenses and membership; and rationing measures in place, such as the use of gatekeeping and a limited wait time. However, their incentive system (fee-for-service, capitation, or salaries) still allows through some of the vices that come with a principal–agent relationship, such as supply-sided increased demand in the case of fee-for-service and a supply-sided decreased demand or favoring patients with the least health needs (cream skimming) in the case of capitation or salaries. To avoid the limitations of these traditional payment mechanisms, pay for performance (P4P) schemes as an alternative form of provision have been introduced. These are meant to directly align the interest of the agent with the interest of the principal (Guinness & Wiseman, 2011).

Figure 10: Pay for Performance Mechanisms



Source: Sergio Flores (2022).

Pay for Performance

Imagine that, after learning about managed care organizations, you decide to implement these health provision mechanisms in a rural setting. However, you start experiencing some of the inherent disadvantages of managed care organizations after some time. Your healthcare providers are not being productive enough, trying to only treat the patients with the highest chance of being healthy or over diagnosing people to earn more profit. At the same time, epidemiological evidence seems to suggest that there is an increase of diabetes in the region, along with new reports of macrosomia (babies born much larger than normal) that have not been identified during pregnancy due to deficient prenatal care. So, you now decide to approach payment in a different way. Instead of paying private healthcare providers based on a salary, capitation, or fee-for-service mechanism, you decide to pay them only if they reach certain goals.

Pay for performance was implemented in this setting as a means for payers to focus on quality while also lowering expenses. A pay for performance program pays healthcare providers an extra amount for achieving or surpassing predetermined quality, performance, or health outcome goals, such as lowering body mass indices (BMIs) in overweight patients. Improvements in the achievement of these goals over time, such as decreases in the rate of anti-hypertensives prescribed, may be rewarded with extra compensation. Providers who do not meet these goals may face financial penalties under pay for performance arrangements (James, 2012).

The quality indicators employed in pay for performance generally fall into one of four categories: process measurements, output or outcome metrics, patient experience metrics, and structural measures (James, 2012).

Process measurements

Process measurements evaluate the frequency and completion of intermediate actions taken to assure health outcomes of patients are achieved, for example, how often aspirin was dispensed to patients vulnerable to myocardial infarctions per guidelines or how often smoke cessation initiatives were held for patients with compromised lung function.

Outcome metrics

Whereas process measurements evaluate intermediate actions, outcome metrics focus on the final results of healthcare provision. Some examples of these metrics are whether a patient's diabetes is under control based on laboratory testing, incidence of major cardiac events, survival measures, and remissions rates for oncological patients.

The use of these metrics in pay for performance is not straightforward since outcomes also depend on several other contextual factors that are beyond the healthcare provider's control. For example, if the provider is following best practices in diabetes management and providing the patient with the latest generation of blood glucose controlling drugs, but the patient is not disciplined with its consumption or keeps a diet rich in carbohydrates, the outcome metrics can be skewed. Partly as a result of this, cost savings are increasingly being included as outcome indicators.

Patient experience metrics

Patient experience metrics look at how patients feel about the care they've received and how satisfied they are with it. In the inpatient context, patients' perceptions of the quality of communication with healthcare professionals, as well as whether health facilities were clean and in an appropriate state, are examples.

Infrastructure measures

Infrastructure measures evaluate treatment facilities, personnel, and equipment. Many pay for performance systems, for example, provide incentives for clinicians to use health information technology (James, 2012).

To revisit the example at the beginning of this section in which you are trying to get the providers in the rural community to work under a pay for performance scheme, you now have to decide how to evaluate your providers to pay them. Based on the previous list of quality indicators categories, you decide to create three instruments:

1. A fidelity checklist that measures process outcomes, such as the number of home visits that the healthcare professionals delivered during the month or how many sex education talks were given at schools
2. A review of medical files assessing blood HbA1 levels glucose (measure of blood sugar levels) for the last three months in patients diagnosed with insulin resistance (diabetes) or number of children delivered at a healthcare facility during the last trimester
3. Surveys delivered to patients to assess their perceptions of how the healthcare provider has explained their illnesses and proposed treatments, how clean and ordered the facilities were, and if they feel their health problems have been addressed in a satisfactory way

Some of the programs currently in place that base their financial incentives on the pay for performance as an alternative form of provision are disease management programs (DMPs) and wellness programs. DMPs are systematic treatment regimens aimed at assisting patients in better managing their chronic illnesses and maintaining and improving

their quality of life. They are also conducted with the long-term goal of enhancing medical care (Institute for Quality and Efficiency in Healthcare, 2016). Wellness programs include activities such as (often online) risk assessment programs for nutrition, weight management, stress management, and smoking cessation.

Some common challenges when designing P4P programs are as follows:

- It is difficult to set reliable and consistent quality indicators that can be used to evaluate healthcare provider performance. Some patient populations present specific challenges, such as differences in intrinsic motivation to improve health, socioeconomic factors, and very different baseline measures.
- It requires a high level of administrative complexity. Careful records must be maintained and organized, verifying it for accuracy and linking health indicators with treatments and costs. Accountants and other record keeping professionals must be involved, as this method relies heavily on carefully curated documentation.
- Cases must be evaluated on an individual, case-to-case basis, which precludes the possibility of easily incorporating standardized guidelines or protocols to evaluate performance.

SUMMARY

In this unit, we have explored the special relationship existing between a patient and a healthcare provider due to the steep information asymmetry between them. This relationship, described under the framework of the principal-agent relationship, is so extraordinary that it often modifies normal demand and supply dynamics. Healthcare providers can very often affect the whole patient care pathway in particular means, setting up the conditions for the existence of supplier-induced demand.

In order to counteract this phenomenon, many supplier behavior-inducing mechanisms have been introduced, mostly based on differing payment methods. At first, these methods focused mostly on spreading financial risk while aiming to obtain the most healthcare provision possible at a lower cost by connecting healthcare providers and patients through mechanisms such as capitation, fee-for-service, and salaries. However, over time, these evolved into strategies aimed at aligning more closely the interests of both agent (healthcare provider) and principal (patient) by reimbursing the achievement of process, quality, and health output indicators.

UNIT 3

THE HOSPITAL AS AN ECONOMIC AGENT

STUDY GOALS

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

- understand the production function of hospitals.
- visualize a standard inpatient care path from beginning to end.
- incorporate economic principles into the analysis of hospitals.
- identify main factors and elements influencing hospital costs.

3. THE HOSPITAL AS AN ECONOMIC AGENT

Case Study

Mario is really excited. He has recently been hired as assistant to the hospital director and his mission is to help the hospital achieve a new level of productivity and efficiency. He holds a master's degree in business administration and is sure that, by applying mainstream financial and managerial theory, he can turn things around.

However, he soon realizes that hospitals have some unique characteristics as productive units. He learns that, unlike in other industries, he cannot simply streamline personnel, stocks of equipment, and medications to the bare minimum in pursuit of maximum effectiveness; he must be ready for surge capacity scenarios. He realizes receiving reimbursements for services rendered is more complex than he thought and observes that hiring more doctors or expanding the number of beds in the hospital does not produce correlated returns.

He decides to take a course in health economics to better understand the dynamics at play. He quickly learns about the framework of managed care and how reimbursements are shaped, realizing that, for a few decades now, hospitals have shifted their focus from revenue creators to cost containment centers.

3.1 The Hospital as a Productive Unit

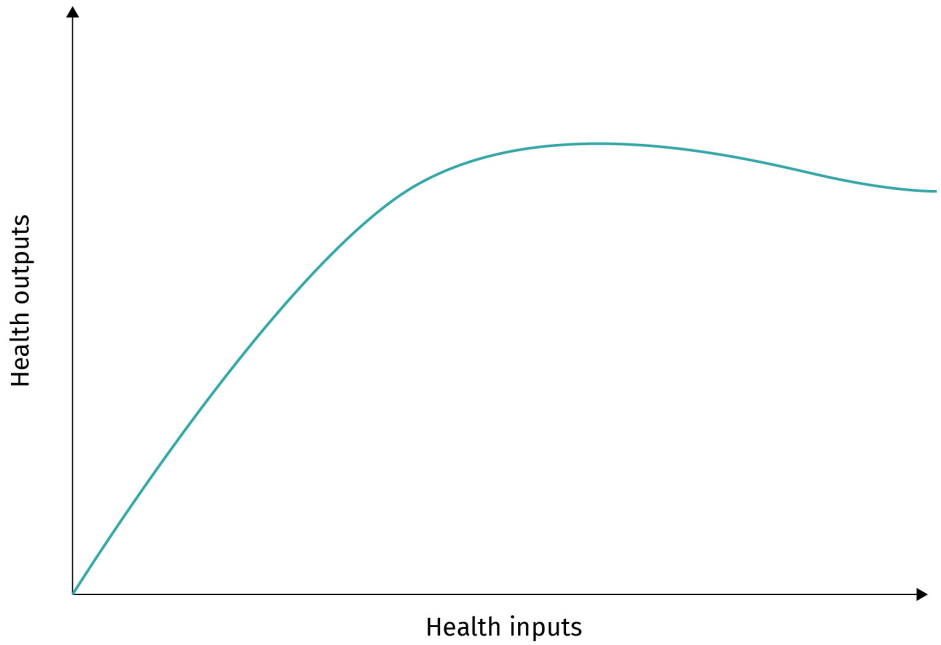
Hospitals are important components of the healthcare system and one of the most important units operating in healthcare markets. They are typically able to provide more complex and specialized healthcare than any other kind of facility by bringing together a wide range of health resources in a single location, organized in such a way that services can be provided 24 hours per day. Additionally, in many instances, they serve as research centers aimed at better understanding disease and treatment. Finally, they can be hubs of medical training and health technology innovation. Their medical, research, academic, and even social role is such that they are often on the receiving end of the biggest allocations in the healthcare budget. An analysis of 148 countries in the world from 1995 to 2017 determined that 34.5 percent of the healthcare budget globally goes into hospitals, which is more than any other healthcare provider (Schneider et al., 2021).

Although a standardized definition is hard to find in the literature, the American Association of Hospitals defines hospitals as "a healthcare facility … with organized medical and professional staff, inpatient beds available 24 hours a day and … providing inpatient healthcare services for surgical and non-surgical conditions and usually provides some outpatient services, especially emergency care" (Abdelhak & Hanken, 2014, p. 726). The type of staff and equipment hospitals possess to cover all the services they provide make them one of the most complex healthcare facilities in existence.

From a strictly economic point of view, a hospital transforms different inputs – such as qualified personnel, equipment, and consumables like medication – into outputs, such as number of patients discharged, surgeries performed, or inpatient childbirths. From this same economic point of view, its main objective is either sustainability or profit maximization, depending on, for example, whether they are owned by a commercial or public entity. Essentially, it follows the same **production function** of any other business unit, but with the main distinction that hospitals always need to be ready to provide precise healthcare services that specifically suit each patient at any given moment of the progression of their disease. That readiness can be planned to a certain extent, as in elective surgeries or outpatient consultation appointments, or completely unplanned, as in emergencies (which require reserve capacity) or extraordinary events such as natural disasters (which require surge capacity). This creates a cloud of uncertainty regarding resource use, which also translates into issues of scale and scope that few other business units outside the sector ever experience.

Production function
This is the relationship between the number of inputs and the amount of product obtained.

Figure 11: Hospital Production Function



Source: Sergio Flores (2022).

To best understand the production function of a hospital, we must first attempt to identify the inputs, outputs, and processes that transform inputs into outputs. Within the context of a production function, inputs are the units of resources (factors of production in mainstream economic theory) needed to produce outputs, which are the goods or services we intend to create out of the inputs. Since this production function attempts to establish a relationship between inputs and outputs, these need to be quantifiable or measurable. Inputs can be generally classified into one of three main categories:

1. Labor, which corresponds to healthcare personnel and supporting staff
2. Equipment and consumables, which correspond to devices, machines, instruments, and pharmaceuticals
3. Infrastructure, which includes physical installations

In a hospital setting, each of these resources serve specific and specialized roles that are not normally present in other kind of healthcare facilities. For example, healthcare personnel are usually trained at a specialized level, depending on the kind of pathologies they need to attend to, such as pediatricians or anesthesiologists. Hospital equipment is similarly more technologically advanced to support specialized diagnosis and treatment services, such as positron emission tomography or genetic laboratories. Physical installations within a hospital must be able to support both personnel and equipment requirements and can be broadly divided into the kind of care they are designed to provide. For example, hospitals may have neonatal, pediatric, and maternal units; intensive care units; operating theaters; or emergency units, among others.

Outputs are more difficult to define in the hospital context. The main purpose of hospitals is to provide healthcare, but that may be achieved through different pathways and lead to different outputs. The reader might be aware that the ultimate goal of healthcare organizations is to improve health, so the ideal unit of output measurement would be a unit of health improvement. Although researchers over time have managed methods to quantify health in different ways, it is not practical to apply them to routine administrative data and we must therefore select intermediate outputs, such as physical measures of activity.

Hospitals provide inpatient treatment, outpatient visits, and emergency care (Glied & Smith, 2011). When analyzing inpatient treatment, perhaps the most common output measure is **hospital discharge**. When considering **outpatient care**, the number of consultations is generally the most commonly used measure; when considering emergency visits, and due to the ever open and uncertain nature of its service, it is preferably measured just by logging the amount of patient emergencies treated.

Hospital discharge
This is the transfer of care of a patient from a hospital to other providers of healthcare or home.

Outpatient care
This is the care of patients administered without overnight stays in the hospitals.

In any case, when considering the hospital as a productive unit, we understand that inputs are converted through several healthcare delivery processes into outputs. The inputs and processes used within a hospital are tailored to each patient. Even patients with the same diagnosis can follow different care pathways depending on their inherent characteristics, such as age or sex. We can exemplify this through a **patient flow**.

Patient flow
This is the movement of patients through a healthcare facility.

Imagine a woman in her thirty-eighth week of pregnancy starting to feel Braxton Hicks contractions. She decides to go to the hospital's maternity unit. There, a group of healthcare professionals assess her condition (using physical examination and equipment, such as ultrasound and laboratory tests) and decide to admit her into the hospital. This prompts a host of administrative and medical actions: A unique identifier and a bed is assigned to the patient; she goes into observation; and, if needed, she will start receiving medication to help her during labor. During labor, she will be constantly monitored to measure her progress and assure her wellbeing by more personnel using more equipment and medication. Once it's time for childbirth, the patient will be transferred to another space within the hospital especially suited for the event. More specialized personnel will be equipped and ready to handle the health needs of the child, effectively starting a new

patient flow within the hospital. At this particular point in the patient flow, a different patient could instead qualify for a cesarean section, requiring a different set of resources in an operating room. Nevertheless, once childbirth is complete, a separate process starts for the child and the mother enters a new observation and monitoring phase. During the next few hours, the mother recovers and receives maternal education from another set of professionals, while the child receives a first set of vaccines and is examined for any kind of abnormality. If their health status evolves satisfactorily, both patients are discharged from the hospital and expected to continue their health monitoring with a different healthcare provider.

Using the previous example, we can attempt to identify how a hospital transformed inputs into outputs: The hospital used healthcare personnel (midwife, gynecologists, pediatricians, nurses, administrative personnel, and potentially anesthesiologists); medications (analgesics, oxytocin, and prostaglandins); equipment (ultrasonography, labor cot, and cardiotocograph); and facilities within the hospital (expulsion room, operating rooms, gynecological admission unit, and neonatal unit) to produce a dual discharge of both a healthy mother and child. The healthcare delivery process that makes this possible includes steps such as assessment, diagnosis, treatment, and monitoring.

3.2 Hospital Cost Functions

From an economic point of view, a hospital's main objective is either sustainability or profit maximization. However, after the introduction of case-based payments like diagnosis-related groups (DRG), managed care organization, and capitation-based payments (where payment is pre-established based on the number and type of patient attended regardless of how much is actually spent), hospitals are now more focused on cost containment than revenue generation. This means that a hospital needs to keep costs as low as possible.

Costs can be classified in many ways. We can start by introducing average and marginal costs. In the simplest terms, average costs are the total costs the production of outputs incurred divided by the total amount of outputs. Marginal costs are the costs of producing an additional unit of output.

Function equations are mathematical expressions that define the relationship between independent variables and a dependent variable. In the specific case of a cost function, it shows the influence of a cost driver on a cost. Therefore, we can use cost function tools to predict what a cost will be based on a cost driver. The reader might be able to infer that, in this relationship, if a cost driver increases, then the cost would also increase. However, a hospital cost function also allows us to forecast how much (and not only if) the cost goes up based on how much a cost driver increases. It also allows us to identify how much the total cost can be attributed to other subtypes of cost.

Hospital costs can be broadly divided into two types: fixed and variable costs. Fixed costs remain the same no matter the production output (i.e., rent for the space). Variable costs change according to the amount of output produced. In the hospital context, the rent for a

building could be a fixed cost, whereas the expenses for analgesic medication are variable costs. Let us exemplify what we have learned so far. Imagine that the following simplification is a real scenario.

Table 3: Types of Costs

Clinic	Rent per month (A)	Maintenance costs per month (B)	Total fixed costs (A+B)	Number of consultations (C)	(Variable) cost per consultation (D)	Total variable costs (CxD)	Average cost per consultation (A+B+(C/D))/C
1	1,000,000	500,000	1,500,000	100	5,000	500,000	20,000
2	1,200,000	500,000	1,700,000	140	5,000	700,000	12,143
3	1,600,000	500,000	2,100,000	180	5,000	900,000	16,667
4	2,000,000	500,000	2,000,000	220	5,000	1,100,000	14,091

Source: Sergio Flores (2022).

In this example, we can identify four key components of our cost function:

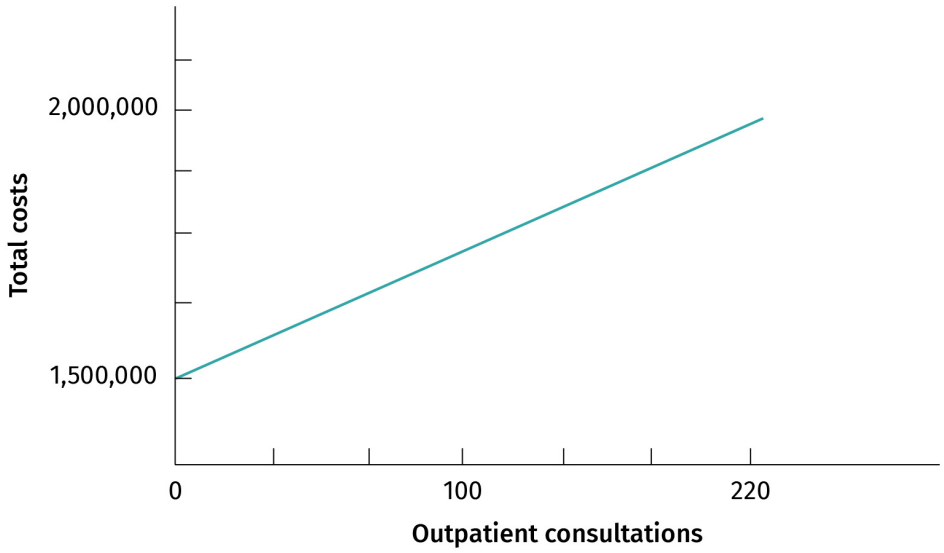
1. Fixed costs (A and B), such as rent and maintenance costs for outpatient clinics, are costs that stay the same no matter the amount of outputs we aim to produce. In this example, the fixed cost for clinic 1 is 1,000,000 for rent plus 500,000 for maintenance. Some common hospital fixed costs could be building maintenance, utilities, and salaries.
2. Variable costs (D) in this case are the costs each outpatient consultation carries. The more output produced, the higher the total variable costs. Common variable costs in hospitals are equipment, medication and supplies, and payments for personnel. For this example, we assume variable costs per consultation of 5,000.
3. Total costs are the sum of both variable and fixed costs.
4. Average cost per consultation is the total cost divided by the number of consultations.

In this example, the variable costs per consultation are the same. Therefore, the difference in average costs per consultation comes from the different fixed costs and case volume of each clinic. If we wanted to calculate our total costs, the relationship between these items could then look like this:

$$\begin{array}{rcl} Y & = & MX + B \\ Y & = & Total \text{ costs} \\ B & = & Fixed \text{ costs} \\ M & = & Cost \ driver \ (number \ of \ units) \\ X & = & Variable \text{ cost } per \ unit \end{array}$$

This is an example of a linear cost function. Linear means that every change of the input changes the output in the same way. However, in reality, cost functions are seldom linear. If we created a graph from this function, it would look something like the following figure.

Figure 12: Example of Linear Cost Function



Source: Sergio Flores (2022).

Two economic concepts that are key to understanding hospital cost functions will now be introduced: economies of scale and economies of scope.

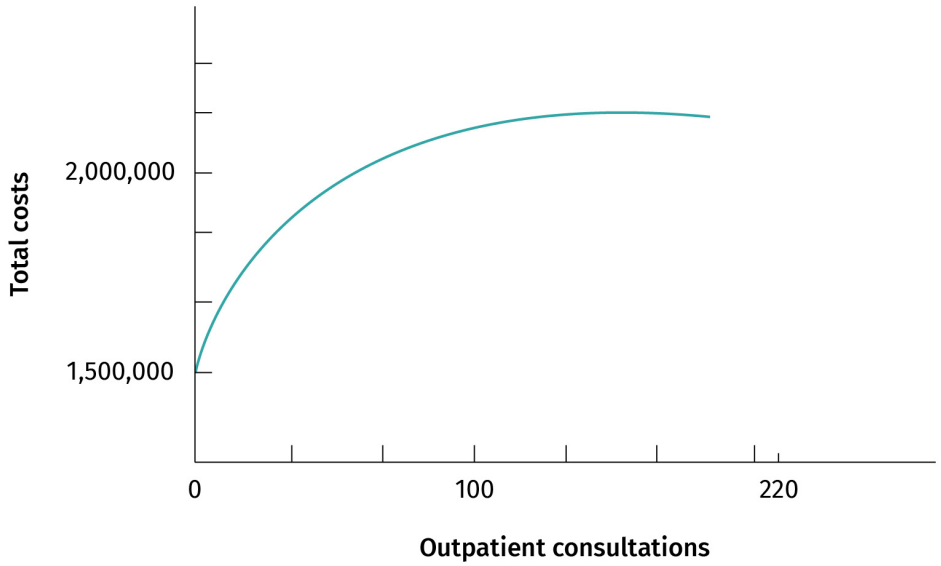
Economies of Scale

Table 4: Economies of Scale

Amount of output we want	Amount of input needed	Scale
1	1	Increasing returns to scale
2	2	Increasing returns to scale
4	3	Increasing returns to scale
8	4	Increasing returns to scale
16	6	Increasing returns to scale

Source: Sergio Flores (2022).

Figure 13: Visualization of Economies of Scale



Source: Sergio Flores (2022).

When we talk about economies of scale, we describe the situation in which a hospital seeks to increase an output level by a value larger than one (e.g., double, triple, or quadruple the amount of output), and the number of inputs necessary to achieve it increases less than proportionally. Therefore, adding inputs provides us with increasing returns to scale. In the previous table, we can observe how the increase of inputs needed to generate more output is proportionally less than the increase in outputs. In other words, costs decrease as the number of outputs produced increases. The figure above helps us visualize how our cost function changes when economies of scale are introduced, modifying the linear relationship of our inputs and outputs into a nonlinear one.

The opposite situation is called a diseconomy of scale, which refers to situations in which a hospital seeks to increase an output level by a value larger than one (e.g., double, triple, or quadruple the amount of output) and the amount of inputs necessary to achieve it increases more than proportionally. Here, adding inputs provides us with decreasing returns to scale. The table below also exemplifies this concept. In other words, costs increase as the number of outputs produced increases. The figure above helps us visualize how our cost function changes when diseconomies of scale are introduced.

Table 5: Diseconomies of Scale

Amount of output we want	Amount of input needed	Scale
1	1	Decreasing returns to scale
2	3	Decreasing returns to scale

4	6	Decreasing returns to scale
8	10	Decreasing returns to scale
16	20	Decreasing returns to scale

Source: Sergio Flores (2022).

Economies of Scope

Another important economic concept to consider in our hospital cost function analysis is economies of scope. Whereas scale gives us an understanding of whether increasing the volume of an input also increases the amount of output, scope tells us if increasing the variety of a firm provides an increase in the amount of output. It lets us know if producing a certain type of good or service will reduce the cost of producing another related good or service. In other words, it tells us whether producing a range of goods and services together is cheaper than producing them individually. One key reason for this to happen is if the different outputs share (at least to a certain extent) the same kind of inputs. In the case of a hospital, this could be the case if it is found that producing both surgical and emergency services is cheaper than producing them in different healthcare facilities.

Table 6: Economies of Scope

Service	Cost per consultation when produced at facility 1	Cost per consultation when produced at facility 2	Cost per consultation when produced together at facility 3
Pediatric services	4,500	4,800	2,000
Prenatal services	2,000	2,200	1,000
Gynecology outpatient consultations	3,000	2,800	1,500

Source: Sergio Flores (2022).

Both the economies of scope and scale determine the size and ranges of services that a hospital can provide. An analysis of the economies of scale tells us the optimal size of a hospital, at least from an economic perspective. Economies of scope informs us about the ideal combination of types of services.

To illustrate economies of scope, we can look at hospital laboratories. Inpatient blood drawing usually occurs in the morning and samples get sent to the laboratory and processed in the morning. Afternoons, evenings, and nights are less busy, but laboratories must still be staffed 24/7 and ready to run urgent analyses. There is an opportunity for economies of scale when one hospital laboratory consolidates the samples from different hospitals to better utilize its equipment. The senders, in turn, only keep a small emergency laboratory. There is also an opportunity for economies of scope when the hospital starts servicing outpatient physician practices where samples tend to be collected from late morning to early afternoon. These samples can be shipped to the hospital laboratory

and processed in the calmer afternoon hours. This new service segment is a source of added revenue where the hospital mainly pays for variable costs of reagents etc., using existing staff and equipment. If economies of scale and scope did not have limits within hospital functions, there would only be a few very large hospitals.

Relationship Between Cost and Outputs

Outputs
A health output is a summary measure of a good or service provided by healthcare providers or short-term results of a health intervention.

Outcome
Health outcomes are longer-term results of health interventions and the impact on health that health outputs have.

After categorizing and understanding costs, we face the challenge of choosing the best **outputs** to use. Given that the final **outcome** hospitals look for is the improvement of health (which is hard to measure in an objective and standardized way), many alternative approaches have been brought forward. The simplest of them rely on proxies for health improvement by using intermediate outputs, such as number of patients admitted, total hospital discharges, or total number of bed days (the number of days that beds in the hospital were used for curative care). However, these measures also have limitations: By just using the number of patients admitted, we ignore other important elements that are important for cost considerations, such as length of stay or severity of the disease. However, using number of bed days does not include outpatient visit, becomes very sensitive to length of stay fluctuations, and leaves out information about the complexity and severity of the disease.

An output measure that can address variation in resource intensiveness and complexity is the case mix. This is a type of patient classification system that aims to group cohorts of statistically similar patients that will require similar treatments or care pathways. Examples of classification systems that use case mix are diagnostic-related groups (DRGs) in the US or healthcare resource groups in the United Kingdom (UK). These classification systems are used as a basis for reimbursement. They are engineered so that groups of patients falling under each of the case mixes are expected to consume comparable amount of inputs in hospitals. A couple of examples of these case mix groups could be

- heart failure,
- hip/knee replacement,
- cesarean section, or
- neonate with significant problems.

Because these DRGs are defined in advance, hospitals know how much money they can expect in the form of reimbursements and will try to keep the costs for those patients within the pre-agreed terms, which potentially improve efficiency (like in a capitation system). However, in practice, these case mixes are complex to manage and could also encourage hospitals to make patients' cases seem more severe (Mihailovic et al., 2016).

3.3 Hospital Cost Inflation

Healthcare expenditure around the world has been rising consistently over the last hundred years, with one out of every three dollars spent on healthcare allocated to hospital costs (World Health Organization [WHO], 2021a). We can broadly categorize hospital drivers of cost into two groups: contextual and intrinsic factors.

Contextual Cost Drivers

Contextual factors are variables that are very difficult or impossible to change for the hospital administration in the short term. These include the size of the hospital, location, whether it is a research or teaching center, type of ownership, or regulation from higher authorities. Theoretically, larger hospitals should experience economies of scale, which should translate into lower average costs. Hospitals localized in urban areas with higher-density populations and shorter average distances from patients are expected to have lower costs due to higher occupancy rates and lower variability in healthcare demand. Teaching hospitals, due to their dual nature, also experience increases in costs (Stock & McDermott, 2011; Li et al., 2002).

Intrinsic Cost Drivers

Intrinsic factors are variables that are under the direct control of hospital management. One of the variables that researchers have identified over the last years as a main driver of hospital costs is technology. The introduction of new treatments over the years has improved healthcare services but made them more expensive. These technologies can come in the form of new and better medication and pharmaceuticals, new surgical procedures, and better radiological tools. In other industries, technological innovation is often designed to increase productivity and outcomes for the industry player. Technology in healthcare may improve productivity but is often designed to improve quality over quantity of outcomes, and potential cost savings may appear out of the hospital setting. This could be the case when an expensive piece of equipment cures a condition which, therefore, does not create downstream costs for outpatient therapy.

Another rising cost driver over the last years is administrative expenses. New healthcare delivery models demand exhaustive utilization reviews, strict accounting controls, and careful record-keeping of patient profiles. Although the US is a clear outlier, many countries in the world are seeing administrative costs rise steadily. Administration is most expensive in countries where day-to-day operations surpluses are the primary source of hospital capital funding, such as the US and, increasingly, the Netherlands and England. The complexity of the payment system and the form of capital funding appear to be driving hospital administrative costs (Himmelstein, 2014). In a similar fashion, liability and insurance costs for all types of healthcare providers are on the rise. These drive costs up via two differing mechanisms:

1. The use of defensive medicine: the departure from standard practice to avoid prosecution by negligence, which pushes healthcare providers to utilize more healthcare per patient without producing more health gains
2. Transferring the costs of higher malpractice insurance premiums to the payer: charging patients more for their insurance plans

SUMMARY

In this unit, we have explored the role of hospitals as productive units within the healthcare system. Hospitals are unique facilities and, as such, are bound by specific constraints and factors pertaining to their health production. Hospitals use different types of inputs within their own production process that are relatively easy to categorize into labor, equipment, and infrastructure. However, identifying appropriate, measurable hospital outputs is a challenge for healthcare administrators. This has evolved from simply measuring hospital activity measures and discharges to deriving complex statistical measures, such as diagnostic-related groups.

To explore the relationship between inputs and outputs in the hospital setting, the hospital cost function is a mathematical tool that allows us to identify if and how much total costs would increase in relation to different cost drivers. If the relationship between inputs and outputs were proportional, we could assume a linear function. However, the concepts of economies of scale and scope show us how changing the amount or types of outputs we aim for also changes the number of inputs required in a non-colinear way, thus producing variable returns in most real-world scenarios.

Nevertheless, hospital costs have steadily risen over the last decades, which has been mostly driven by a combination of contextual factors, such as size of the hospital, location, whether it is a research or teaching center, type of ownership, or regulation from higher authorities, as well as intrinsic factors including new technology, more complex administrative expenses, and the rise of defensive medicine.

UNIT 4

HEALTH INSURANCE

STUDY GOALS

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

- understand the objective of health insurances.
- identify the different kinds of health insurances available in the healthcare markets.
- explore the potential unintended byproducts that health insurance may prompt.
- recognize some of the mechanisms insurance companies apply to correct moral hazard and adverse selection.

4. HEALTH INSURANCE

Case Study

Olga decides to spend a year in another country to gain some professional experience. One day, she is cooking some dinner when she accidentally cuts her finger with the knife and panics. Olga realizes that, as a foreign resident, she has not yet registered for public or private health insurance. This is the first time in her life she has faced this situation. Fortunately, the injury is small and a plaster is enough.

The next day, she goes to the municipal office to inscribe herself to the national health insurance, and she is asked about her employment status and tax number. She unfortunately does not have the necessary documentation and is told she can't enroll yet. The requirements she was asked for pique her curiosity: What do employment information and tax contributions have to do with applying for public health insurance?

She calls home and enrolls in a private healthcare insurance scheme until she qualifies for the public option. As a young student, she is told that she is welcome to join but must first fill out several forms concerning her medical history and daily habits. She agrees and sends the forms by post the next day. A week later, she is sent her insurance policy contract and terms of use. She reads them all and sighs. She wonders why applying for health insurance is such a complicated process and decides to read more about the topic.

4.1 The Demand of Insurance

Health is highly valued by humans, as it is understood to be a necessary prerequisite to cope with all demands of daily life. In his seminal work, Grossmann (1972) argued that health could be viewed as an initial stock that depreciates with age but can be increased with investment. Therefore, health is not something that can be purchased directly; we can only influence determinants of health to preserve it or purchase health services that can improve it.

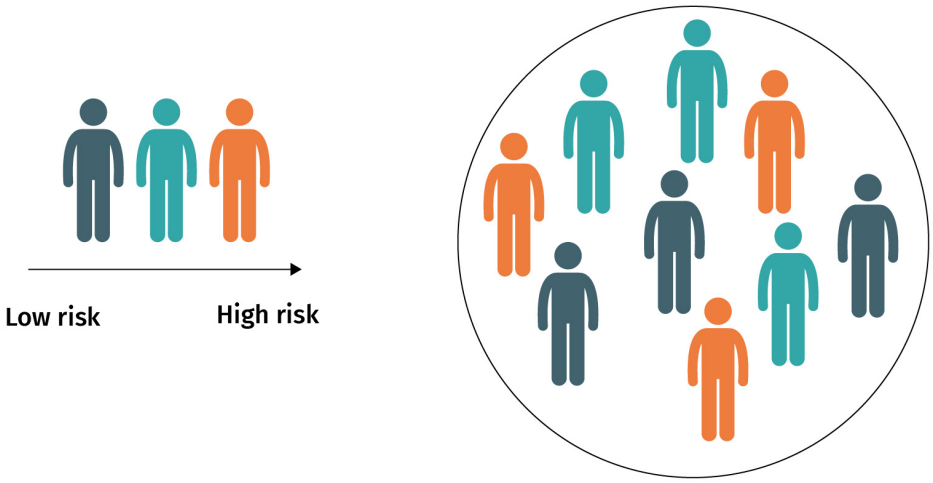
Since health can be affected by a myriad of factors, each person's stock of health is permanently at risk of being depleted. Some of these factors can be influenced (like personal behaviors or habits), while others cannot (such as genetic predisposition, accidents, or the emergence of highly transmissible diseases). As a consequence of this, even within the same population at any given time, individuals have differing risk health profiles. In a similar sense, each individual also has a differing risk profile over time. Therefore, people cannot reliably predict when they will fall ill and need healthcare services. Since healthcare cannot be substituted and is necessary for survival, it is an inelastic good; namely, its consumption remains the same without much influence from changes in price. A single individual facing a health mishap by chance could find themselves with huge healthcare costs and risk **catastrophic health expenditures** (United Nations, n.d.).

Catastrophic health expenditures
This describes a situation in which household

To counter this, risk pooling has emerged as a financial and systematic solution. The World Health Organization (WHO) defines risk pooling as the "accumulation and management of revenues in such a way as to ensure that the risk of having to pay for healthcare is borne by all members of the pool and not by each contributor individually" (Ahangar et al., 2018, p. 1). Risk pooling therefore increases the possibility that individuals in need of healthcare can reach it in an affordable and timely manner. It also allows for crucial resource transference from the healthy to the sick, from younger to older individuals, and from high- to low-income individuals. In other words, risk pooling offsets the higher healthcare costs of the less healthy with the lower healthcare costs of healthy people. The larger and more balanced a pool of people, the more spread out the financial risk.

expenditure on health is a certain large percentage of total household income, the two most common thresholds proposed being 10 and 25 percent (United Nations, n.d.).

Figure 14: Risk Pooling



Source: Sergio Flores (2022).

For all these reasons, risk pooling has been brought forward as a central element of health financing systems to ultimately achieve **universal health coverage** (WHO, n.d., para. 1). Risk pooling is also the intuition behind health insurance schemes, which come in different formats.

Universal health coverage
This is a global initiative aiming to "ensure all people access to the health services they need, when and where they need them, without risking financial hardship" (World Health Organization [WHO], n.d., para. 1).

4.2 The Supply of Insurance

The demand for health insurance from rational risk-averse actors (i.e., individuals, organizations, and governments) has prompted the existence of several health insurance schemes. Historically, governments and charities have financed health services for groups of people that need help in India, China, Arabia, and medieval Europe (Guinness & Wiseman, 2011). There is evidence of private health insurance in Europe from around the eighteenth century, and social insurance was introduced in Germany in 1883 by Otto van Bismarck. Employment-based insurance systems also developed in the rest of Europe, Latin America, and Asia. Later, the United Kingdom introduced social reforms that extended

coverage through government provision with the aim of covering the whole population. In the context of the global push to achieve universal health coverage, many countries around the world followed this example. Even in countries like the US where private insurance plays a major role, examples of social insurance exist, including Medicare, Medicaid, and the Veteran Affairs (Guinness & Wiseman, 2011). Nowadays, we can broadly categorize health insurance in three groups based on how they are financed: tax-based, social, and private health insurance.

Tax-Based Health Insurance

Tax-based health insurance, also called national health insurance, is based on the example set by the United Kingdom in 1948 by William Beveridge (Light, 2003). This type of healthcare is funded, for the most part, through general taxation and therefore administered by the government. With these funds, the government either pays healthcare providers directly or through intermediaries. This type of insurance is usually also more reliant on funds coming from the segments of the population with the highest income through either progressive income or corporate taxation. In some countries, these tax-based systems cover the population unprotected by social health insurance schemes.

Social Health Insurance

This type of health insurance comprises obligatory, income-related contributions paid through payroll but most of the time operated by a public agency. This contribution is usually shared between employees and employers. If everyone in the population of a country is required to join a social health insurance plan, it becomes very similar to tax-based insurance. Within this type of health insurance, some countries opt to pool all the social health insurance financing together, whereas other countries have multiple funds that can receive this financing.

Private Health Insurance

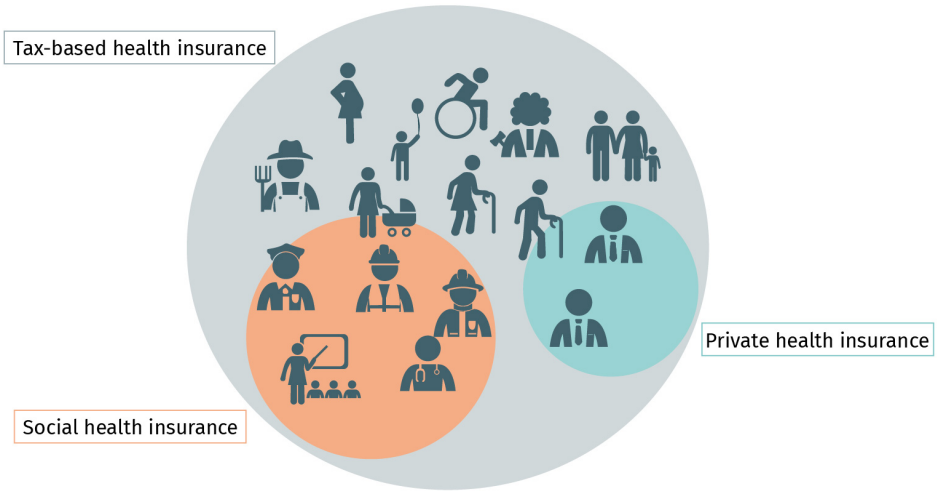
The main distinction between private and tax-based or social health insurance is that private health insurance is voluntary and financed through risk-based payments called premiums. Individuals pay these premiums to private insurance companies in advance, and these payments may be part of individual or group packages. The types of private health insurance are as follows (Guinness & Wiseman, 2011):

- principal: This occurs when the private health insurance acts as the individual's main coverage, for the most part because public health insurance is not available (e.g., Switzerland and the US).
- substitute: This occurs when an individual has the option to replace a public health insurance scheme to which they have access and does so with a private health insurance.
- complementary: This occurs when private health insurance pays for **copayments** that the public health insurance does not, therefore acting as a complement to public health insurances.

Copayments
These are fees that an insured person must pay for each healthcare service received.

The figure below illustrates the scope of each of the insurance categories discussed. All three of the insurances are not mutually exclusive and can coexist within each country. Tax-based insurance is designed to cover the whole population, as it is funded by mandatory general taxation, and citizens not subscribed to either a social health insurance or private insurance can be covered. Social health insurance is more common among employees contributing to the funds through mandatory payroll deductions. Therefore, it generally only covers employed people able to contribute through their payroll, as well as their direct family members. Finally, private health insurance pools are considerably smaller because they are financed through voluntary contributions.

Figure 15: Health Insurance Categories



Source: Sergio Flores (2022).

4.3 The Case for Moral Hazard

A common issue driving healthcare prices up is the presence of moral hazard. Since an insured person has their healthcare costs covered, they may behave differently to if they had to cover healthcare costs themselves.

Ex ante moral hazard is the behavior of individuals that increases the chances of falling ill or being injured. For example, imagine Olga has been cycling her whole life with a helmet, drives her car with care, and tries to avoid unhealthy habits. However, after purchasing private health insurance, assuming that she takes no risk in terms of the payment for any injury, she stops taking all these precautions and starts engaging in higher risk activities, like diving from cliffs and horse riding. These activities increase the chances of her healthcare consumption going up compared to when she was not covered by insurance.

Ex post moral hazard is when the individual consumes healthcare in excess of what they normally would if uninsured. In this case, let's imagine that Olga – being a young healthy woman – normally went for consultation once a year for checkups or the occasional flu. Now that she is covered by health insurance, she is more likely to request professional healthcare advice or consultations based on the smallest symptoms or suspicion of a disease. She might believe that a checkup every month is now justified even though there is no reason or health gain obtained.

In both scenarios, the most likely outcome is that Olga will consume more healthcare than she would before being insured, at least partly motivated by the fact that she does not have to pay for it herself. If this happens with a significant amount of the individuals that have come together to pool risks, then risks are not spread out but rather increased for all. Insured individuals therefore do not pay for the total price of the healthcare they consume. The only way to cover for these collective increased risks would be to increase the contribution each individual makes, resulting in an overall increase of costs transferred to the patients in the form of higher contributions or premiums.

Researchers have confirmed the existence of moral hazard through several notorious randomized evaluations and quasi-experimental observations (Einav & Finkelstein, 2018). Health insurance managers have identified this as an issue that could put the schemes at risk and have therefore put forward measures to dissuade moral hazard, such as the following (Guinness & Wiseman, 2011):

- copayments: a fee that an insured person must pay for each healthcare service received
- deductible: the amount of money that an insured person must pay before the insurance company begins to cover expenses
- co-insurance: a percentage of the healthcare costs that the insured person must pay from the total costs after subtracting the deductible (with an upper limit called the out-of-pocket maximum)

4.4 Asymmetric Information and Adverse Selection

For a market to reach an equilibrium point between supply and demand, and therefore the best price for both consumers and producers, information access has to be symmetrical between both parties. When information asymmetry exists, one of the actors with the most information may abuse their position and make choices that maximize their utility unilaterally.

In health insurance markets, one of the most common scenarios of asymmetric information is through adverse selection. In this scenario, patients have more information about their own health status than the insurer. For example, a patient with severe preexisting conditions (such as stage B heart failure) may be tempted to join a health insurance at a standard premium price since it might still be cheaper than their actual treatment would be without health insurance. As in the case of moral hazard, the individual is not paying

the total price of healthcare consumed. When other individuals with reduced health do the same thing without disclosing it, the health insurer has to raise prices to match healthcare consumption. A rise in costs may then dissuade healthy individuals from enrolling in the same health insurance scheme, as they might calculate that the cost of the premium outweighs the benefits of being insured. This can lead to a catastrophic negative reinforcing loop (sometimes called a death spiral) in which less healthy individuals join, raising prices and prompting healthy individuals to leave, collapsing the health insurance plan. As with moral hazard, insurance companies have implemented the following measures to attempt to avoid the pitfalls of adverse selection to the extent that is possible. They are as follows:

- accurate identification of risk factors
- robust information verifying systems to confirm information coming from new members about risk profiles (such as age and preexisting conditions)
- aggregate limits of liability, which means that insurance companies set a maximum amount of coverage (e.g., for a case or period of time).

SUMMARY

In this unit, we have introduced the reader to the basic concepts of health insurance. Demand for health insurance is always present for different individuals (or even for the same individual at different time periods). Discarding health risks exposes individuals to catastrophic health expenditures. To overcome individual health risks and expenditures, risk pools have been devised. Here, several individuals with differing risk profiles are brought together, compensating for one another's health status and allowing transference of money between the pool.

Different kinds of health insurances have been devised over time, culminating in our current three main models: tax-based insurance, social health insurance, and private insurance. The distinguishing factor between them is the voluntary or mandatory nature of the contributions.

Finally, we have reviewed information asymmetry; this is a peril inherent to health markets that endangers risk pooling schemes. To counter risks, different tools and mechanisms from health insurance providers have been put forth.

UNIT 5

ECONOMIC EVALUATION

STUDY GOALS

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

- understand the theoretical basis for economic evaluations within healthcare.
- identify both costs and benefits measured and computed in economic evaluations.
- perform basic economic evaluation calculations using comparative analysis.
- recognize the practical applications of economic evaluations within the healthcare sector.

5. ECONOMIC EVALUATION

Case Study

Haitao is a technical assistant for a province health authority and has been tasked with determining whether a change in the official state-subsidized drugs list should be approved. The situation is as follows: A pharmaceutical company is offering new antiparasitic medication for children that has fewer side effects and fewer interactions with other drugs, thus making it safer than the current antiparasitic medication. It requires a single dose, whereas the current standard is a three-dose treatment. Furthermore, it seems to be almost twice as effective at killing parasites compared to the current treatment. However, the new drug costs three times as much as the one in use, and antiparasitics are not the only drugs that need funding from a limited budget. Therefore, Haitao needs to find a method to establish whether or not the benefits of the medication are worth the extra cost.

While they are reaching out to health economists, Haitao hears from their boss that things are more complicated. The government's main mission is to improve the academic performance of children in schools; therefore, it is considering allocating money away from the health province authority and giving it to the educational authority to start a meal program in schools. Haitao must now not only be able to argue why investing more money in a newer drug would be worthwhile but also why money in antiparasitic medication is a better investment than school meals.

Haitao starts to wonder how to make a choice between two things that could benefit the population, when they hear back from a health economist. The health economist listens to the dilemma and says "we'll need to perform a health economic evaluation of all choices involved." Haitao's situation is one that occurs many times, regardless of how resource rich or constrained a country might be. Decisions regarding the allocation of resources to maximize the welfare of the population require a systematic approach. We will explore the toolkit available to health economists in this unit.

5.1 Theoretical Bases of Economic Evaluation

As a branch of both public health and economics, health economics aims to prevent disease, promote health, and prolong life among society from a public health perspective, as well as understand the most efficient use of health resources from an economics perspective. In other words, health economics is a discipline that is primarily concerned with the interconnection between maximizing the population health outcomes and the best use of available resources.

One of the central theorems in the field of economics is how humans deal with scarcity, that is, a situation in which resources are limited but the demand for goods or services is unlimited. Scarcity invariably leads humans to make choices regarding where and how to allocate resources to obtain the greatest amount of benefit. However, the crux of this decision-making is that for every resource allocation choice made, a tacit decision is also made to not allocate these resources to other choices. The resulting **trade-off** of the resource allocation choice we make is called opportunity cost. Scarcity and opportunity costs are central concepts in the field of economics. However, within the healthcare sector, decisions regarding resource allocation have special significance for the following reasons (Drummond et al., 2015):

Trade-off
This is a situation in which a decrease of an element must be made to increase the gain of another.

- Health economics is more concerned with choices that maximize **utility** for the population than saving costs as in other fields of economics.
- Health problems impact societies at different levels. They may begin as the number of cases and deaths reported or the intrinsic pain, suffering, and disability individuals may experience. They might then ripple into broader issues, such as the amount of money spent to fix the health problem or the amount of income lost by sick individuals. Other individuals are also affected: Lost income may affect the individual's family and their employer must incur costs to replace the person.

Utility
This is the total satisfaction received from consuming a good or service.

Impacts of decisions made within healthcare often extend beyond the healthcare sector and reach many (if not all) parts of a society. That is to be expected, as good health is a prerequisite for all activities in life. However, in a similar fashion, resources spent in healthcare not only represent opportunity costs for the healthcare sector but for society as a whole. Every unit of a resource spent in the healthcare sector is a resource not used in infrastructure, like roads and bridges, or in education, like hiring teachers or purchasing books and other materials.

Health Economic Evaluation Perspective

Health economic analysis can take on different perspectives, depending on which population groups are included in the analysis of the impacts or outcomes of the health decisions to be taken. This perspective then becomes the point of view through which the costs and outcomes of decision-making are deliminated. Two commonly used perspectives are healthcare and societal (Byford & Raftery, 1998):

1. A healthcare perspective only takes into account costs and benefits that are limited to the healthcare sector and excludes costs and benefits for any other potential stakeholder, such as employers, and other sectors in society (e.g., education or social services).
2. Societal is a broader perspective on costs and benefits stemming from the health intervention, considering the impact on society beyond the healthcare sector.

To exemplify, imagine that you are trying to perform an economic evaluation of a drug that seems to be more effective than standard care for eliminating parasites. If you were to assume a healthcare perspective, you would be interested in the medical outcomes of the drug (episodes of diarrhea, diagnosis of dehydration, hospitalization, quality of life, etc.) and concerned about the costs incurred providing healthcare, such as the amount and

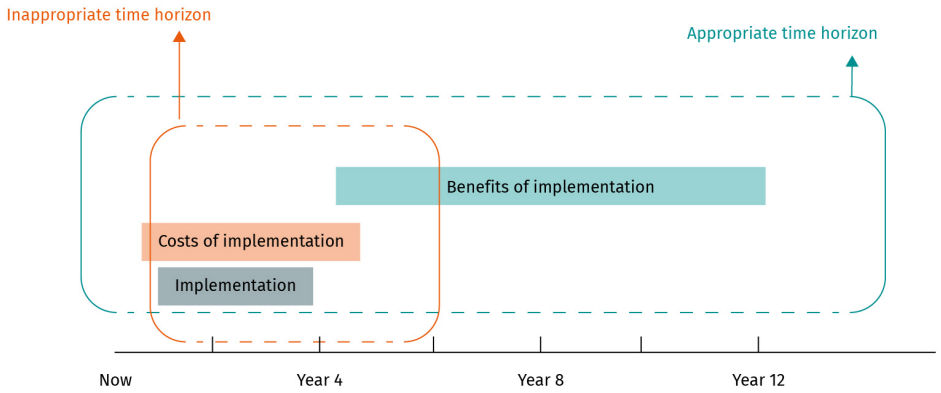
cost of drugs needed for treatment. Alternatively, assuming a societal perspective would include other outcomes, such as school attendance or sick leave, and costs, such as lost earnings.

However, as in most fields, demand always exceeds supply in healthcare. There are many reasons for this, but authors agree that the most important ones are as follows (Frankel et al., 2000):

- the slow but steady rise in life expectancy, which is a success in improving healthcare but also means that the number of older people that consume more healthcare is continuously growing as a consequence of chronic and degenerative diseases becoming more prevalent in old age
- technological innovations within healthcare, which provide better clinical outcomes but also come at a greater cost without necessarily producing a correlated increase in healthcare productivity
- patient expectations possibly resulting in a mismatch between what the healthcare sector offers and what the patient prefers as treatment (or demands)

Time Horizon for Measuring the Effects of an Intervention

Figure 16: Time Horizon



Source: Sergio Flores (2022).

Another aspect to consider within health economics, as with many other fields of economics, is the time horizon. The magnitude of the costs and benefits varies according to the point in time relative to the health intervention. For instance, an early health intervention in children might result in most of the health benefits occurring only during adolescence or adulthood and, if benefits were measured after a narrow window of time, they would be missed for the most part. As shown in the figure above, this would create the impression that those health interventions are not producing enough benefits to be worth the cost. However, as stated by the “time value of money” principle, the value of money also fluctuates over time: The value of a unit of currency today is worth more than the value of the same unit of currency in the future due to factors such as inflation and the opportunity costs of investing (Glied & Smith, 2011). Therefore, to correctly identify and collect all rele-

vant costs and benefits of each alternative, a time horizon must be set. In some instances, a very short time horizon of days or weeks might be warranted, whereas in others, the time horizon can be extended throughout the individual's whole life.

Because of all of the above, health economists deal with the challenge of making decisions in healthcare by using a set of analyses collectively referred to as health economic evaluations. There are several types of economic evaluations but, in essence, they are comparative analyses between different alternatives in terms of their costs and health outcomes.

5.2 Measuring Costs

Every health intervention attempts to improve the health of a population, albeit at a certain cost. As we strive to understand the most efficient balance between health benefits and the costs needed to obtain them, we need to look more closely at health costs.

A common method of categorizing costs is dividing them into direct, indirect, and intangible costs. Direct costs of a health intervention relate to costs directly attributable to patient care. From a provider's perspective, this may include drugs, used equipment and supplies, imaging, and health professional's salaries. From a patient's perspective, they may include medical costs, such as fees paid by the patients, and non-medical costs, such as traveling, accommodation, and meals in pursuit of the treatment. Indirect costs are not directly attributable to the provision of care. From a provider's perspective, these include general administration and maintenance costs. From a patient's perspective, they include loss of income (or opportunity cost) due to attending medical care (Ibrahim et al., 2015; Ernst, 2006). Intangible costs include things like pain, suffering, or discomfort that a person might be experiencing. They are very hard to put a value to and, for the most part, are not included in health economic evaluations. However, in many instances, they might be major factors in decision-making (Guinness & Wiseman, 2011; Špacírová et al., 2020).

To be able to compare different health interventions, a precise quantification of costs must be made. To do this, a simple three step process is usually followed (Raftery, 2000): (1) resource identification, (2) measurement of resources, and (3) valuing resources.

Resource Identification

Identify the resources the health intervention requires and the amounts of each you need. A common approach is to categorize them within seven different types of resources that may be used (Raftery, 2000):

1. Personnel: the health workforce
2. Buildings: the physical location in which health is delivered
3. Equipment: the tools used to assist health service provision
4. Supplies and drugs: single-use resources used to provide healthcare
5. Transportation: the logistic system to mobilize other health resources

6. Training: teaching skills to the health workforce
7. Socialization: the creation of policies and programs to promote health

To identify relevant resources, we need to establish the perspective used for the analysis. If the interest is on how much the healthcare sector needs to pay for an intervention, then the resources in which patients, their families, or their employers incur or forego would not be included in the analysis. Likewise, a societal extended perspective might require a wider identification range of resources, considering all the possible ripple effects in resource use an intervention might have (Raftery, 2000).

Measurement of Resources

Identified resources are quantified in physical units (not monetary units). In many instances, logs (in the case of randomized controlled trials) or questionnaires designed specially to collect this kind of data are used. Depending on the approach taken, the resources can be measured one by one in a very detailed manner or be assumed within a larger category. For example, to measure the amount of resources used to treat a case of heart failure at a hospital, we might either go to the logs of the hospital and check how many vials of medication, gloves, syringes, and hours of healthcare personnel, etc., were used to treat the patient. Or, we might use precalculated estimations of resource use for patients dependent on the particular disease and inherent characteristics, such as age and gender (Raftery, 2000).

Valuing Resources

Once the resources are identified, we need to assign a value to them. After knowing the amount of resources required for an intervention, we can use market prices as a proxy to the monetary value of these resources since any other measure of opportunity cost is difficult to ascertain. However, market prices can be inaccurate even as a proxy in situations in which prices are subsidized by the government or set by monopolies, so good judgement must be executed, and prices must be adapted to every unique circumstance. In the case of societal perspectives when we might want to include productivity losses, there are two different techniques that can be used: the human capital approach and the friction cost method. Once we know the total amount of resources needed for the intervention and the prices we might assign to them, the total cost may be calculated by adding the resources included and multiplying by the price (Raftery, 2000). Any estimation of costs carried out using these three steps must also address two questions (Glied & Smith, 2011):

1. How disaggregated is our approach when identifying and measuring resources?
2. What strategy was used to evaluate resources and cost components?

Micro-costing and gross costing pertain to the first question. Micro-costing aims to recognize resources at a very detailed level, whereas gross costing tends to aggregate resource items. In the case of costing hospital-delivered pregnancies during the last year, a micro-costing approach would identify and value each individual item used, such as gloves, syringes, plasters, oxytocin vials, and nurse's and doctor's time. The sum of these factors is the cost of the hospital -delivered pregnancy and the result would be extrapolated to obtain the yearly costs for the hospital. A gross costing approach for the same case would

rely on standardized measures of resources that are specific to the setting – in this case, the hospital. These standardized measures would already estimate the cost for the delivery at the hospital. Pertaining to the second question, the two approaches used to evaluate resources and cost components are either a top-down or bottom-up approach. A top-down approach assigns expenses to each organizational cost center, trickling down to units of activity by estimating costs for a whole set of services and products during a time period and assigning them to cost objects (e.g., patient, case). A bottom-up approach identifies and assigns value to the resources used per patient or case and then aggregates them to to a cost center (e.g., organizational, activity-based). Many authors recommend a bottom-up approach since it has more flexibility and precision (Špacírová et al., 2020).

5.3 Measuring Benefits

Historically, measuring health benefits has proven to be a difficult task. When talking about health, researchers are mostly interested in the maintenance or improvement of health in the population. However, this is not easy to measure or quantify, particularly at an administrative level where most records come from. Therefore, one of the most common ways to measure health benefits is by quantifying clinical outcome measurement (sometimes also referred to as natural health units), which can be either surrogate endpoints or clinical outcomes, depending on the focus of the analysis and the source of data (Guinness & Wiseman, 2011).

For example, in an intervention aimed at treating heart failure, some surrogate endpoints that can be used are functional capacity scores, such as peak oxygen values in cardiopulmonary exercises, points in a patient-reported outcomes test examining fatigue and depression dimensions, glycated hemoglobin (Hb1Ac) levels in diabetes, or estimated glomerular filtration rate (eGFR) in the case of kidney failure. Clinical outcomes could explore morbidity (such as hospitalizations or emergency visits) or mortality.

Using clinical surrogates and outcomes (such as the ones mentioned above) as the benefits of an intervention has both advantages and disadvantages. First, they are more intuitive because they are usually the same benefits researchers and healthcare professionals use in their everyday practice to measure the success or failure of their interventions. Furthermore, if different interventions are designed to alter the same clinical phenomenon, comparison between them is straightforward. For example, if we want to compare whether a diet program or a pill is more effective to reduce blood glucose levels, using glycated hemoglobin (Hb1Ac) as a medium-term metric allows us to compare their effectiveness because, ultimately, both treatments aim for the same result. The measuring of glycated hemoglobin (Hb1Ac) is also standard practice and is, therefore, a familiar metric understood by physicians, researchers, and patients alike. These units are also quite sensitive to changes: Weight, blood pressure, glucose levels, and other factors shift rapidly and dramatically around health-status fluctuations of the patients. Therefore, the relationship between costs and effects can change significantly in response to clinical outcomes (Whitehead & Ali, 2010).

A disadvantage of measuring clinical outcomes is that it is difficult to compare different health-improvement treatment paths. To illustrate, imagine you want to compare the clinical effectiveness of an exercise program aiming to reduce weight in obese patients and a public health intervention aiming to reduce the incidence of Dengue fever cases by removing breeding spaces for disease transmitting mosquitoes. While both interventions ultimately aim to improve the health of people and extend their lifetimes, how do you compare lost kilograms of weight to the number of hospitalizations with Dengue fever (Greco et al., 2016)?

Furthermore, suppose our health intervention provides several health benefits we might consider relevant to our evaluation. A daily training regime, for example, can improve weight control, immunological response, and mental health, among others. However, we cannot combine all of these different clinical endpoints and outcomes into one for our health economic evaluation. We would need to produce a health economic evaluation for every outcome, which potentially makes presenting results harder. Finally, if we settle on a single, specific clinical metric in our health economic evaluations to make comparisons possible, this disregards other potential benefits or outcomes of a program and thus makes our economic evaluation "incomplete" (Greco et al., 2016).

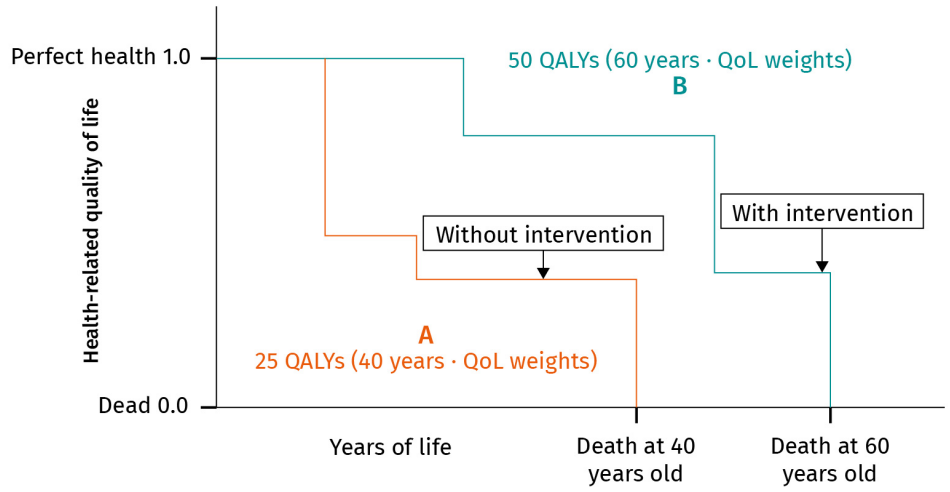
So, how do we overcome the disadvantages of using specific clinical metrics in our evaluation? How do we compare different interventions resulting in different clinical metrics? How do we encompass all potential health benefits resulting from an intervention without leaving any out?

Generic Measure of Health

The quality-adjusted life year (QALY) is the closest attempt to a standardized unit of health that combines mortality and quality of life. This type of outcome has become popular and its use is widespread among health economic evaluations. It has become an important outcome used in countries practicing evidence-based policies in healthcare. In its simplest definition, a QALY is a year of life in perfect health (Whitehead & Ali, 2010).

QALYs are determined by estimating the life expectancy of a patient after a health intervention and assigning a weight to each year with a quality of life score on a scale from 0 to 1. The quality of life score is gauged in relation to the person's ability to carry out the activities of daily life, as well as freedom from pain and mental disturbance (Prieto & Sacristán, 2003).

Figure 17: Quality-Adjusted Life Year Visualization



Source: Sergio Flores (2022).

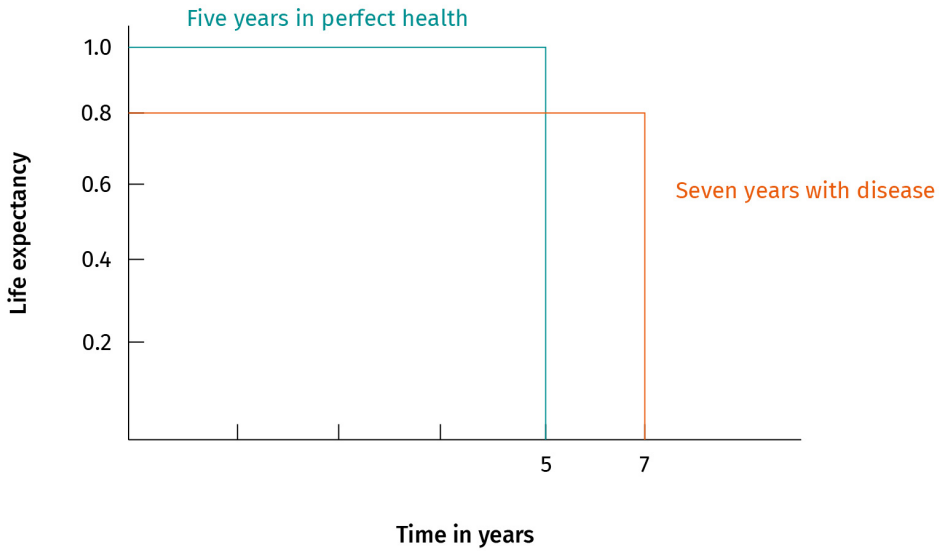
As you can see in the figure above, quality of life scores (indicates in the figure as QoL) change over a person's life, dependent on health deterioration, until death. However, changes in quality of life can vary greatly depending on whether or not an intervention is put in place. The figure could exemplify the difference between a person who, after a depression diagnosis, accepts treatment versus the same person who does not accept treatment. Assuming everything else is equal, without the intervention, the person is likely to live for fewer years at a reduced quality of life compared to if they received health intervention. Therefore, the A area in the figure represents the QALYs of the person if the health issue was left to progress untreated (i.e., 25 QALYs) and the B area represents the QALYs gained during the individual's lifetime due to the intervention (i.e., 50 QALYs).

As previously mentioned, a QALY is a generic measure of health resulting from combining life expectancy and quality of life weights, but how do we calculate these quality-of-life weights in a patient? There are direct and indirect methods to do so. The direct methods include several types of implied preference assessment tools, such as time trade-off and standard gamble.

Direct Methods to Assess Quality of Life

The time trade-off approach presents people with two potential outcomes and asks them which one they would prefer. One has the option of living the remainder of their lives in a condition of diminished health or living in perfect health for a shorter amount of time. The length of time spent in perfect health varies until the person is undecided between the two options. Participants are therefore asked how much time they would be ready to give up if it meant preventing a state of diminished health.

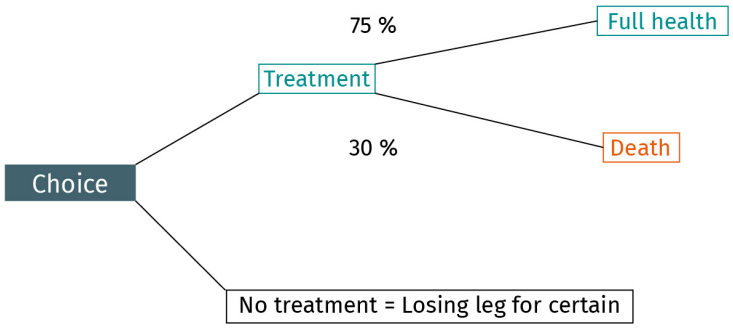
Figure 18: Time Trade-Off



Source: Sergio Flores (2022).

In the standard gamble, the decision is between taking a chance on either remaining at a specific level of health or risking death against remaining in a better state of health with assurance. The likelihood of dying varies until the person cannot tell the difference between certainty and chance. The danger of mortality that a patient would tolerate in exchange for a cure increases with the severity of their health condition.

Figure 19: Standard Gamble



Source: Sergio Flores (2022)

Indirect Methods to Assess Quality of Life

Indirect methods usually assess quality of life through specially designed and weighted questionnaires, such as the commonly used EuroQol-5 Dimension (EQ5D). The EQ5D is a simple and generic questionnaire that appraises five dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Patients provide feedback on each dimension. The questionnaire also includes a visual analog scale (that looks like a thermometer), which allows patients to mark their own health score from 1 to 100. Patients tick boxes to show how they feel in each of these dimensions, and these answers can be turned into a single numerical value called an index score that ranges between 0 and 1 – adapted to different contexts – giving us the quality of life value.

Figure 20: EQ5D Example

Under each heading, please tick one box that best describes your health today.

Mobility

I have no problems walking about.	☐
I have some problems walking about.	☑
I am confined to bed.	☐

Self-Care

I have no problems with self-care.	☐
I have some problems with washing or dressing myself.	☐
I am unable to wash or dress myself.	☑

Usual activities

I have no problems performing my usual activities.	☐
I have some problems performing my usual activities.	☑
I am unable to perform my usual activities.	☐

Pain and discomfort

I have no pain or discomfort.	☑
I have moderate pain or discomfort.	☐
I have extreme pain or discomfort.	☐

Anxiety and depression

I am not anxious or depressed.	☐
I am moderately anxious or depressed.	☑
I am extremely anxious or depressed.	☐

Source: Sergio Flores (2022).

There are several variants of this questionnaire, with more or fewer levels per answer to improve sensitivity or make it easier to fill out. Furthermore, there are a wide array of other instruments and questionnaires, such as the short-form health survey or the World

Health Organization (WHO) quality of life questionnaire. Others are tailored to different populations, such as the CHU9D for children and the Mental Health Quality of Life questionnaire (MHQoL) to assess mental health.

Using generic measures of health like QALYs has the biggest advantage in that it allows us to compare multiple health outcomes resulting from health intervention alternatives. They include morbidity and mortality, and some of them include quality of life, in one single metric. Furthermore, when considering how much the society is willing to pay for each QALY gained or **disability-adjusted life year** (DALY) averted in an intervention, an inbuilt value for money is assigned to these outcomes. Finally, these types of outcomes are widely present in many evidence-based policies, priority setting guidelines, and burden of disease studies.

Disability-adjusted life year
This is a year lost due to disability, ill health, or early death.

However, there are some criticisms of these benefit measures. Ethical concerns have been raised about whether they strongly favor those with the greater capacity to benefit. For example, the amount of QALYs to be gained by a terminal or elderly patient are very low compared to the ones gained by another kind of patient. Should that be the only determinant in prioritizing healthcare? As previously noted, there is not a single, straightforward way to obtain quality of life values. The quality of life values for many diseases can vary between countries due to several socioeconomic and cultural differences, and therefore require a lot of country-specific data and preference-based weights calculations. Finally, QALYs are much less sensitive to change because of the nature of answers in the questionnaires: You are asked to pick a value from a predefined set of answers varying between three and five options, whereas metrics such as blood pressure, weight, or glucose levels can detect small changes easily due to the continuous nature of the values. Choosing an option from five answers requires bigger changes to elicit different answers.

5.4 Practical Steps in Economic Evaluation

Economic evaluations are defined by two main features: comparative analysis and assessment of costs and benefits. A comparative analysis compares two or more options. When a new health intervention is to be introduced, standard therapy or even a “do nothing” approach are valid comparators. An assessment of costs and benefits is conducted for each alternative. Incremental analysis is used to compare the differences in cost and benefits between alternatives, meaning that the cost of each additional unit of benefit is ascertained for all alternatives. Within these two conditions, several types of economic evaluations exist. Some of the most used are as follows:

- cost-minimization analysis
- cost-benefit analysis
- cost-effectiveness analysis
- cost-utility analysis

Cost-Minimization Analysis

This type of analysis is best for situations in which the health benefits for the different alternatives are the same, meaning the only difference between them is cost. We then compare the costs of each alternative and, since the aim is to minimize the costs without affecting health benefits in any way, the alternative with the lowest costs is chosen. This might seem straightforward, but it relies on a robust costing approach for each alternative, which carries a degree of complexity. Cost-minimization analysis is no longer favored, as it assumes perfect clinical equivalence between choices, which is very hard to observe in a real-life scenario.

Cost-Effectiveness Analysis

This type of analysis works when comparing interventions that aim to produce changes in the same type of health outcome or natural health unit. For example, if the aim of a set of interventions is to improve appearance, pulse, grimace, activity, and respiration (APGAR) scores in newborn babies, then the cost and the changes in APGAR scores for each alternative will be taken into consideration for the analysis. The results of the analysis are presented in summary as a statistical measure called incremental cost effectiveness ratio (ICER). This is obtained by dividing the difference in total costs by the difference in the change to health benefits (such as APGAR score). As a result, the cost per additional unit of health benefit is obtained for the different treatment options, ergo, the ICER.

Cost-Utility Analysis

This type of analysis is very similar to, and considered a subgroup of, cost-effectiveness analysis. Instead of a natural health unit, it uses utility scores to compare alternative interventions. QALY is the most common utility measure used. Cost-utility analysis is ideal when comparing different alternatives that aim to improve the health of a population through different means. It is also the most appropriate type of analysis when comparing interventions that might extend life years at the expense of quality of life, e.g., due to side effects. As in cost-effectiveness analysis, an ICER summarizes the results, with the difference being that they represent the cost per additional QALY gained.

Cost-Benefit Analysis

This type of analysis converts all costs and outcomes (benefits) of an intervention to monetary terms. The difficulty of this approach is converting health outcomes (and all other non-monetary benefits) into monetary terms. This is tackled for the most part using two strategies: 1) exploring patients' willingness to pay for a health intervention or avoid the costs of an illness or 2) exploring the economic productivity gains resulting from the health gains caused by an intervention. Cost-benefit analysis is often used at executive levels of government decision-making since it can be used to compare interventions with very different outcomes. The results of this type of analysis are often summarized in two ways: through a net present value, which is obtained by subtracting the discounted cost of an intervention from its discounted benefits, or through a benefit cost ratio, which divides the benefits by the costs (Centers for Disease Control and Prevention, n.d.).

Whichever economic evaluation strategy is chosen, gathering data and evidence to appropriately quantify both costs and benefits of health interventions is not straightforward. Evidence synthesis to feed an analysis might sometimes require assumptions or the use of estimates. Therefore, it is always good practice to run a sensitivity analysis alongside economic evaluations. Sensitivity analyses work by toggling assumptions or inputs in our analysis, either one by one (one-way sensitivity analysis) or several at once (multiway sensitivity analysis), and reporting on how these variations affect the final result of the analysis. Sensitivity analysis adds robustness to the findings.

5.5 Economic Evaluation and Resource Allocation

The main purpose of economic evaluations is to identify the most efficient way to use resources within healthcare. This efficiency may be either technical or allocative. Technical efficiency aims to produce a certain output with the least possible amount of input. In terms of our healthcare setting, it would aim to identify the least costly way to achieve health gains. Allocative efficiency is more concerned with identifying the best possible allocation of resources within different alternatives to maximize outputs or health gains. In the case of healthcare, it aims to identify the right mixture of healthcare programs to maximize the health of society (Guinness & Wiseman, 2011; Palmer & Torgerson, 1999).

Cost effectiveness and cost-utility analysis are founded on the production function approach, which calculates the ratio of input to output to find the least costly way to produce services. Therefore, we can infer that these methods are more aligned with technical efficiency. They can partially provide us with allocative efficiency information, but it is constrained to the healthcare sector. Cost-benefit analysis is primarily concerned with allocative efficiency, as it is able to produce comparable results within the healthcare sector, as well as other sectors (Guinness & Wiseman, 2011).

Decision Rules in Economic Evaluations

Economic evaluation is meant to assist in decision-making within healthcare. For each type of economic evaluation, some decision rules have been proposed to come as close as possible to optimized decision-making (Glied & Smith, 2011). When different alternatives promise very similar or identical health gains (or manifest clinical equivalence), a cost-minimization analysis is warranted. A decision-maker should look at the cost of each and choose the one with the lowest costs. In reality, this is a rare scenario. When a decision-maker needs to choose between the current approach and a new intervention, a cost-effectiveness analysis is warranted (keep in mind that cost utility analyses are considered a subset of cost-effectiveness analysis). A cost-effectiveness analysis can be conducted via the following four steps (Guinness & Wiseman, 2011):

1. Identify cost and health gains for each alternative. All alternatives that are costlier and produce fewer health gains than the current approach (which may include doing nothing) must be discarded. This is referred to as the dominance principle.
2. If left with more effective (i.e., more health gains) but also more costly interventions, list all the interventions in order of effectiveness (the ones producing more health gains first) and then run a cost-effectiveness analysis. Each intervention is then compared to the next one using the ICER. Rule out the interventions that result in a higher ICER value than a more effective intervention (Kattan, 2009).
3. The decision-maker will choose the intervention that yields the greatest amount of health benefits at the lowest ICER value.
4. To decide whether or not to fund a new intervention, the decision-maker must compare the ICER of the chosen intervention to a monetary threshold, ideally derived from the national context of where the intervention is to be implemented, or their willingness to pay. For example, Sweden has set interventions that demonstrate more health gains at 500,000 Swedish kronor (around 55,000 euros) per QALY as cost-effective (Vallejo-Torres et al., 2016). The WHO classes interventions as highly cost effective if they can avert a DALY (a measure similar to QALY but calculating for health loss instead of health gain) for less than the per capita national gross domestic product (GDP).

Cost-benefit analysis may be used to either assess a single intervention or compare it to alternatives. It is also different from the other economic evaluations in the sense that it can be used to compare health and non-health outcomes. This analysis computes all costs and benefits for the intervention(s). When both are quantified, you can either subtract the cost of an intervention from its benefits to produce a net present value, or you can divide the benefits by the costs, which produces a benefit cost ratio. If the net present value yields a positive result, or if the benefit cost ratio is greater than one, then the intervention's benefits exceed its costs. If a choice has to be made between different interventions, then the one with the greatest net present value or benefit cost ratio is selected.

When alternative health interventions are not mutually exclusive and we can choose more than one, a similar economic evaluation process is performed. However, instead of discarding all other choices for the most cost-effective one, they should be sorted in order of either cost per health unit, cost per QALY, or net positive value/benefit cost ratio, depending on the economic evaluation used. Decision-makers would then select these ranked interventions based on the budget and willingness to pay thresholds.

Table 7: Decision Rules in Health Economic Evaluation Example

Options	Total cost	ICER
A	€400,000	€30,000 per QALY
B	€700,000	€40,000 per QALY
C	€600,000	€60,000 per QALY
D	€900,000	€20,000 per QALY

Source: Sergio Flores (2022).

For example, imagine your budget is set at two million euros and you are presented with the choices shown in the table above. You cannot fund all of them, but you can choose more than one. The order of priority in which you should fund these projects according to the decision rules would be as follows:

1. D
2. A
3. B

Option C, albeit cheaper overall than options B and D, is not as cost effective, as it provides fewer health benefits relative to the amount of money invested. Therefore, this option would be excluded, as our budget is not enough to fund them all.

SUMMARY

In every market, the principles of scarcity and opportunity cost dictate that people must make choices – the health market is no exception. To maximize the utility society perceives through these choices, a structured approach to decision-making concerning health services is necessary. In health economics, a specific toolkit of analyses is available for this purpose, which are collectively called health economic evaluations. These allow decision-makers to compare health interventions based on their cost and benefits.

The most commonly used health economic analyses are cost minimization, cost effectiveness, cost utility, and cost-benefit analysis. All of them have different characteristics that make them appropriate to use in different circumstances but share two key principles: They involve a comparative analysis of different alternatives and they run said analysis in terms of cost and health benefits.

Cost-minimization analysis has fallen out of favor due to assuming perfect clinical equivalence between choices, which is very hard to observe in a real-life scenario. Cost-effectiveness analysis allows us to compare different alternatives aiming at the same type of outcome, and cost-utility analysis makes use of special composite measures of longevity and health that can be used to compare different kinds of clinical outcomes, called QALYs. Cost-benefit analysis allows us to compare between health and non-health interventions by converting all costs and benefits to monetary terms.

UNIT 6

DISTRIBUTION

STUDY GOALS

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

- understand the concepts of equality and equity in the context of healthcare.
- explore the concept of distributional analysis within health economics.
- recognize the benefit incidence analysis as a useful tool to assess inequity within health systems.

6. DISTRIBUTION

Case Study

Maria is facing a dilemma. She has been laid off from work and cannot find a well-paid job easily without a degree. Her husband's job planting and harvesting coffee is not enough to cover the expenses of raising their six children. Their household finances are stretched out and they have been borrowing from family and friends for the last six months to make ends meet. Therefore, for the past month, Maria has been working as a cleaning aide at two factories, working double shifts, and asking her mother for help taking care of the children.

However, during the past week, Maria has developed a bad cough that has gotten worse. This morning, she fainted when showering and realized she started coughing up blood. She knows that not showing up for work means no food for the family, but she also feels unable to cope with double shifts in her condition. She also knows that visiting a doctor at the primary health clinic will result in getting referred to a specialist that is several hundreds of kilometers away and could take weeks in waiting times. She will need money to travel, pay for any tests and medications, and possibly even for accommodation.

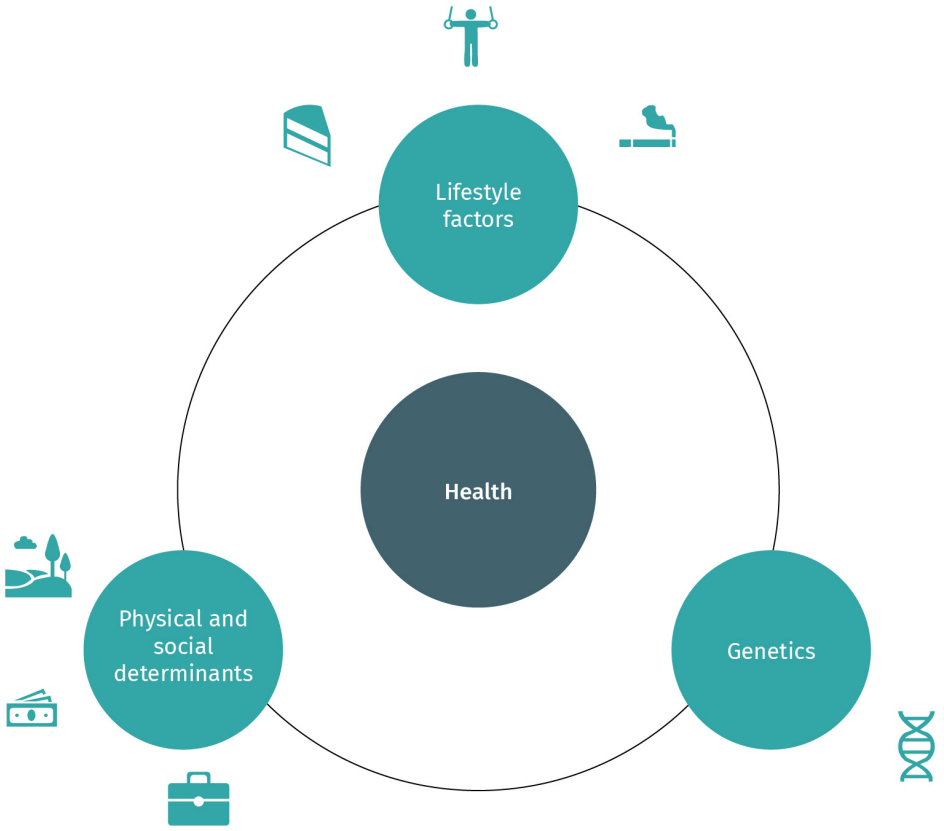
She decides to talk to her boss. Maybe she can secure some sort of payment in advance or a permit to see the doctor. When she arrives to the administration office, she is told that her boss has taken the week off to watch the final match of a prestigious soccer tournament in a different continent, but she can get an appointment next week when he is back. Maria excuses herself and goes to the bathroom. She resolves to push through without consulting a medic for her children's sake and wonders amidst tears what will happen now.

Maria's story is common. A wide combination of factors – within and out of an individual's control – place people in positions where they have considerable health needs but cannot access proper healthcare without compromising financial stability. This raises the question of equity in health economics.

6.1 Equity in Health and Healthcare

At any given point in life, human health is determined by many factors. The main determinants of health are shown in the following figure.

Figure 21: Main Determinants of Health II



Source: Sergio Flores (2022).

Determinants of Health

We can categorize determinants of health into three different types based on the extent to which individuals have control of them:

1. Genetics and the inherited risks to health: This is determined before birth and individuals have no control over this whatsoever; it is sometimes referred to as a health precondition.
2. Physical and social determinants: This includes cultural norms, working conditions, pollution, presence of sanitation, and social standing. These translate into the opportunities people may have to maintain or improve their health. At least during early stages of life, this is also uncontrollable for most people.
3. Health-related lifestyles: This refers to an individual's behavior relating to diet; exercise; and unhealthy habits, such as smoking and recreational drug use.

The interaction between these determinants dictates that differing health needs – or capacity to benefit from healthcare – are present among individuals. In other words, health inequalities are unavoidable without any intervention.

Equity

To address these inequalities, two types of policy interventions exist (Glied & Smith, 2011):

1. Policies targeting the aforementioned specific determinants, which require efforts from within and beyond the healthcare sector
2. Policies that aim to distribute healthcare in terms of equity

It is widely recognized that social inequalities are extremely hard to avoid due to underlying variables, such as natural variations from birth, health-related preferences, personality traits, and cultural and social norms. Therefore, most policies addressing inequalities aim to reduce rather than eliminate them (Glied & Smith, 2011). A distinction must first be made between equality and equity as concepts applied to healthcare. Equality refers to the equal distribution of healthcare and, in practical terms, equal access to healthcare for everyone, without making any distinctions to account for variations or health needs individuals may have. Equity refers to a distribution of healthcare that accounts for both health needs and inequalities in healthcare (Guinness & Wiseman, 2011; Glied & Smith, 2011). In practical terms, equity is often operationalized in both horizontal and vertical equity.

Horizontal equity

This refers to the effort to ensure that people with the same level of disadvantage are not treated any differently, meaning that if individuals have the same health need, all of them should have the same access to healthcare, the same amount of healthcare, and the same amount of healthcare funding allocated to them. For example, if two individuals have the same medical condition, such as kidney failure, but one individual is much richer than the other, both should still be able to get the exact same treatment.

Vertical equity

This refers to adjusting people's care to their level of need or lessening the gap in health needs between healthy and less healthy individuals. This also means improving healthcare access for those who might have difficulties accessing it, making sure they are able to consume healthcare according to their need regardless of factors such as age, gender, ethnicity, socioeconomic status, sexual orientation, nationality, or geographical location. For example, treatment for a cancer diagnosis should be different from the treatment of a small wound. In another example, people that cannot afford quality healthcare should be financially supported to do so, whereas people with the means to afford better access should not expect the same type of financial support.

Univeral Health Coverage

Under the equity framework, several initiatives have been put in place on a national and international level. One of the most noteworthy initiatives is the universal health coverage goal, which has the following aim: "People can access quality health services, to safeguard all people from public health risks, and to protect all people from impoverishment due to illness" (The World Bank, n.d., para. 3). A central part of this initiative is the idea of risk

pooling for crucial resource transference from healthy to the sick, from younger to older individuals, and from high- to low-income individuals. The most **progressive** types of health systems with regards to raising funds are the ones in which widespread pooling is in place. This is typical in social health insurance or tax-financed national health service systems. The most **regressive** health system are private-type systems funded by private health insurances or out-of-pocket payments. Within these systems, the more money you make, the smaller the share of your income that goes to health. Regressive systems increase the risk of catastrophic health expenditures and poverty for individuals.

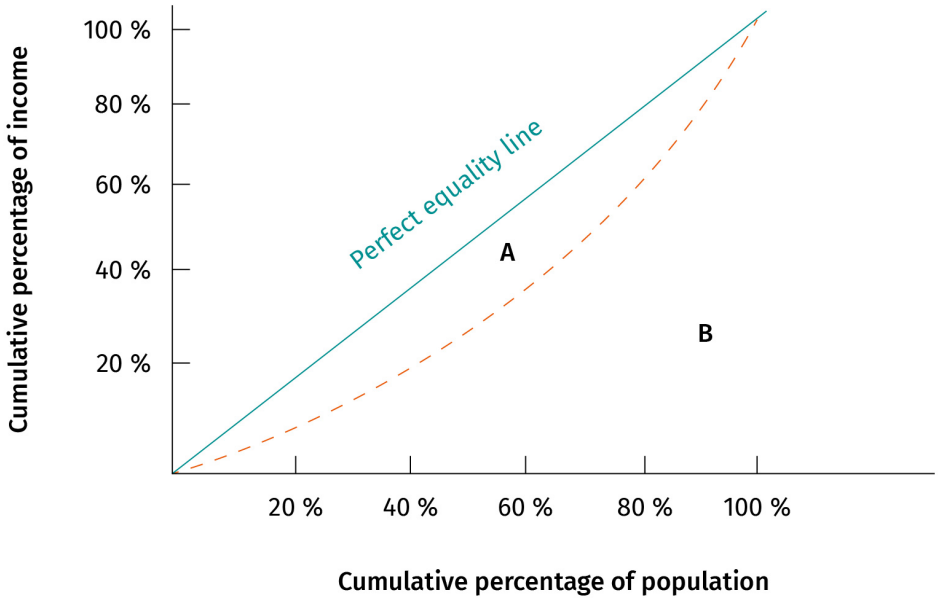
Progressive
This refers to the collection of taxes or contributions at a greater percentage for individuals with higher incomes and a lower percentage for individuals with lower incomes.

Regressive
This refers to the collection of taxes or contributions at a percentage that does not vary based on the amount of income an individual has, making it disproportionately more difficult for lower earners.

Measuring Inequity

There is no single category of empirical data that can be used to measure health inequities. However, since it is now a well-documented fact that health and illness are strongly correlated to social gradients (Braveman & Gottlieb, 2014; Donkin, 2014), one common way health systems evaluate their own equity performance is by performing subgroup analyses, i.e., keeping track of specific health outcomes on segments of populations disaggregated by quintiles of wealth and income. Some methodologies have been brought forward using these intuitions as its core, such as the concentration index, a modified Gini index, and the relative inequality index.

Figure 22: Gini Coefficient

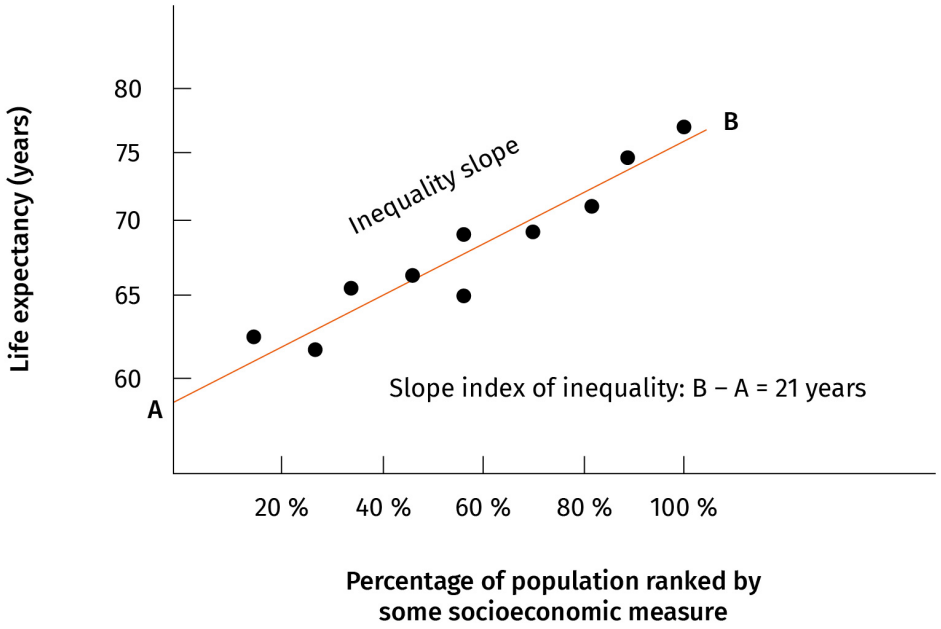


Source: Sergio Flores (2022).

The most common way to measure inequity at a country level is the Gini index, which is a type of concentration index specific to wealth or income. The figure above illustrates the way this is calculated. The dashed line is the Lorenz curve of the population, which is a visualization of the wealth distribution in the society. The more curved it is, the more inequality the population has. The solid line represents perfect equality in terms of wealth

distribution in a society. The Gini index is the result of dividing area A by the combined areas of A and B. This results in a number between 0 and 1. A value of 0 would mean that wealth is perfectly distributed, while a value of 1 would mean that one person holds all the wealth (Steinbeis et al., 2019).

Figure 23: Slope Inequality Index



Source: Sergio Flores (2022).

The slope inequality index and its derivation, the relative inequality index, are another way to measure inequity. To calculate this measure, we must first use some kind of socioeconomic index to rank the whole population. One common approach is to use wealth quintiles, as in other methods (such as the Gini index). However, some researchers use specific indices adapted to their areas. For example, Sweden has constructed the care needs index derived from nine different variables that include family composition, ethnic background, employment, education, and recent internal or external migration. After ranking the population, the health outcome we want to explore is chosen, such as life expectancy, vaccine coverage, or prevalence of stunting. These outcomes are then plotted on the graph and a regression analysis is performed to find the line of best fit for the health outcomes, as shown in the figure above. This allows us to identify the difference in health outcomes between the most and least socioeconomically deprived groups of a population (Steinbeis et al., 2019).

To find the linear relative risk inequality index, we simply divide the slope index of inequality by the value for the whole population. The linear risk inequality index can range between -2 and +2. If a population was ranked from most to least deprived, the closer the value to -2, the more concentrated health outcomes are on the most deprived group. How-

ever, the closer the value is to +2, the more concentrated the health outcomes are on the least deprived group. Therefore, the interpretation of the results depends on how the population is ranked and whether we are tracking a positive or negative health outcome.

6.2 Interdependent Utility and Equity

Individuals are embedded in society and come from different types of social groups, such as families or communities. This is the case for all individuals, whether suppliers or demanders of healthcare. Standard health economic theory treats individuals in the health market as autonomous and utility-maximizing actors, ignoring social positions structured by social relationships. The membership of individuals within different social groups creates a package of rights and responsibilities supported by the individual's collective intentions (Davis & McMaster, 2007). This, by extension, introduces an element of obligation into an individual's decision-making and constitutes one of the bases of interdependence principles. This provides a theoretical framework that explains why individuals, collectively, might pursue initiatives aimed at reducing health inequities even if they do not experience a direct utility.

The perception that an individual's use of healthcare also affects another person's utility is called interdependent utility. It was first identified when healthcare consumption that produced a large number of externalities, such as vaccination, was creating utilities not only for the autonomous individual making the decision but also for others. Since then, the concept of interdependence has expanded and three types of interdependent utilities have been identified (Labelle & Hurley, 1992):

1. Selfish interdependence occurs when an individual cares about others' consumption of healthcare because it might directly affect their health status. This is exemplified by someone cohabiting with a person who has an infectious disease.
2. Paternalistic interdependence occurs when an individual cares about others' consumption of healthcare because of the effect it may have on the other individual's heath status. An example of this is the highly restrictive alcohol purchase system in Sweden. The state has made this decision instead of the individual based on the potential health effects for the population, even though it could affect other variables (e.g., tax collection or political popularity).
3. Altruistic interdependence occurs when an individual cares about another's health status independently of how this was achieved. An example of this could be a young, healthy person who decided to vaccinate against COVID-19 early in the pandemic to protect older, less healthy individuals even if, at that point, vaccinating would not significantly improve their chances.

The existence of high levels of interdependent utilities can be observed by the high amounts of charities and contributions dispensed, the number of volunteers involved in different initiatives, and even the existence of large amounts of blood given at blood banks. All of these activities do not guarantee an increase of utilities to the giving party and are not mandatory, but they happen either way (Labelle & Hurley, 1992).

6.3 Benefit Incidence Analysis

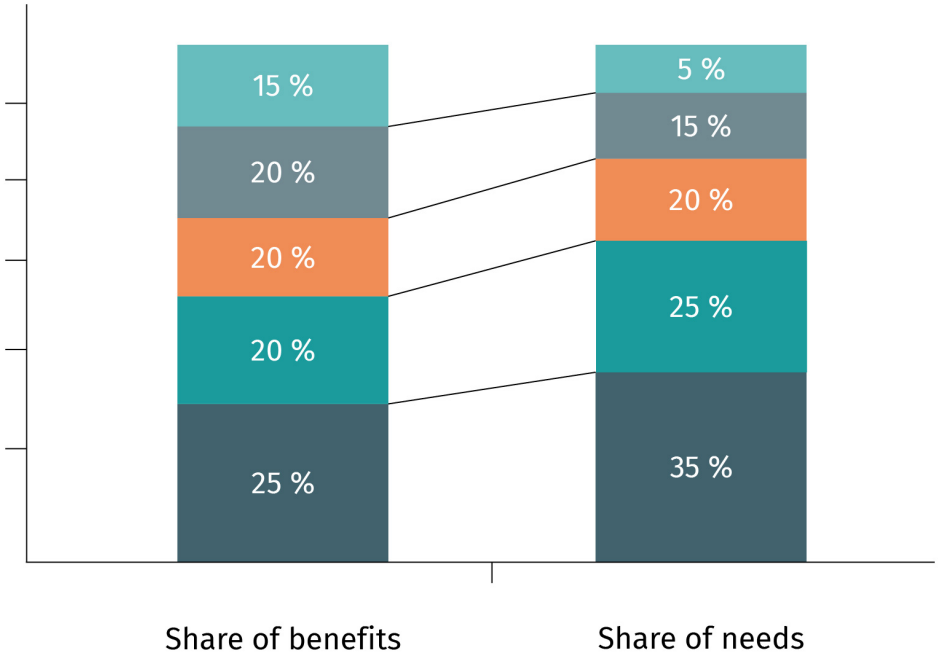
The benefit incidence analysis is a methodology that aims to assess the distribution and impact of health expenditure by the government with the understanding that health funds should benefit the lowest socioeconomic groups the most.

Most approaches to measuring health performance try to link health expenditure and several goals, such as average and proportional health status of the population and financial contributions. These approaches provide a broad overview of the effects of increasing health expenditure on the health outcomes of different groups within the population but miss how health services are delivered and to whom as the connection between health expenditure and health outcomes. The benefit incidence analysis lets us more precisely identify how health spending is assigned to the population (Pearson, 2002). A benefit incidence analysis is performed using the following six steps (McIntyre & Ataguba, 2011):

1. Rank the population from poorest to richest using a measure of socioeconomic status.
2. Estimate the utilization of health services among the different socioeconomic groups.
3. Calculate the cost of each unit of health service.
4. Multiply the resulting cost of healthcare services by the utilization rate for each socioeconomic group.
5. Add up all the calculated monetary utilization values of the previous steps across different types of health services for each socioeconomic group.
6. Compare how the distribution of health monetary utilization values benefits each of the socioeconomic groups.

The results of a benefit incidence analysis are usually presented as either a simple percentage (as shown in the figure below) or using concentration curves and indices. However, many studies using benefit incidence analysis compare the percentage of benefits to the share of the population at each quintile, which is 20 percent. Therefore, a 20 percent assignment of benefits for each quintile is often considered an acceptable distribution in these studies. Perhaps a more accurate way to approach a benefit incidence analysis is by pairing the distribution of benefits to the need of healthcare for each percentile, which is usually greater for those at the lowest quintile than those on top. This adds the necessity to measure the health needs of population within each quintile, and this can be addressed by the use of questionnaires that self-assess the health status, such as the EuroQol-5 Dimension [EQ5D]), or questions that self-assess illness in household surveys, such as the Living Standards Measurement Survey by the World Bank. Then, as shown in the figure below, different population segment needs can be more easily tracked, and the equity principle is fulfilled.

Figure 24: Distribution of Health Benefits to Health Needs



Source: Sergio Flores (2022).

Benefit incidence analysis is a powerful tool tailored to assess how well a health system is meeting the population health needs through health service delivery. It is most often deployed by governments and international organizations, such as the World Bank, to assess the performance of health systems within the equity dimensions (Demery, 2000; Lanjouw & Ravallion, 1999; Pearson, 2002; The World Bank, 2001). It does have a couple of caveats: It requires accurate and updated household data to create health utilization and needs parameters that feed into the analysis, making it harder to deploy in settings where data collection is of poor quality or sporadic.

SUMMARY

We are not all equally healthy. Health inequalities are inherently present among humans due to a mix of preconditions before birth (genetics) and conditions during early stages oflife (social, economic, and cultural environments) over which we have little or no control. On top of that, people engage in different lifestyles influenced by our surroundings but are ultimately individual choices. To address these inequalities, two main types of policies have been proposed: those that aim to modify specific determinants (which, for the most part, are so diverse and multidisciplinary that they are out of the scope of this book) and those that aim to redistribute healthcare among the population based on equity principles.

Equity and equality are related but not the same thing: Equality aims to distribute resources equally to everyone without any consideration for other factors, while equity in healthcare tries to assign resources based on heath need; ergo, more healthcare is assigned to those that have the most potential to benefit from it. Within this equity framework, a global initiative has been launched called universal health coverage. It aims to ensure people have access to the health services they need without risking financial hardship. Many countries are making steps towards this goal, but measuring health inequity is not straightforward. Many mechanisms have been brought forward to measure progress around health inequities, among them a modified Gini coefficient, a concentration index, a slope inequality index, and a benefit incidence analysis. Benefit incidence analysis aims to measure the distribution of health service provision among different quintiles of income or socioeconomic status.