BLOCK 2
FOUNDATIONS OF HEALTH ECONOMICS

BLOCK 2 INTRODUCTION

Block 2 is on Foundations of Health Economics. It has three units (Units 3 to 5). **Unit 3** is on Demand for Healthcare Services. The unit first makes a distinction between how health is regarded both as 'a consumption good' as well as an 'investment good'. The characteristics and measurement of healthcare services are then outlined. In the context of demand for healthcare services, the unit explains the concepts like utility maximisation, determinants of demand, price elasticity, income elasticity and market demand.

Unit 4 is on Supply of Healthcare Services. This unit deals with issues like: (i) factor market for healthcare services, (ii) equilibrium price for the services of physicians in the specific contexts of pure monopoly and price discriminating monopoly and (iii) production and supply of healthcare services.

Unit 5 is on 'measurement of health benefits'. It covers two areas viz. cost benefit analysis and impact evaluation. In the former, issues like cost of illness, willingness to pay, cost effective analysis (CEA) and cost utility analysis (CUA) are explained. In the latter, the method of constructing 'comparison groups' is explained. Different methods of impact evaluation like: randomised evaluation, propensity scoring matching (PSM), double difference (DD) method and instrumental variable (IV) method are also discussed.

UNIT 3 DEMAND FOR HEALTHCARE SERVICES

Structure

3.0 OBJECTIVES

After reading this unit, you will be able to:

- explain how health is both a consumption and an investment good;
- specify the functional form of health in a production function framework;
- delineate the four characteristics of health care services;
- specify the conditions for the utility maximisation of health care services;
- show that the demand for health care services is similar to that of any other commodity;
- discuss the factors determining the demand for health care services;
- outline the concept of Fuzzy Demand for health care services; and
- describe the concept of 'elasticity of demand' in health services.

3.1 INTRODUCTION

Being in good '*health'* is the most primary pre-requisite for the well-being of any individual and thereby for the society as a whole. In the preamble of the 1946 Constitution of the World Health Organisation, enjoyment of the highest attainable standard of health is acknowledged as one of the fundamental rights of every human being. Even in our daily life, we experience how much important it is to remain healthy. Ill health not only makes us suffer physically and mentally, but also acts as a hindrance to other sources of well-being. For instance, it may stop us from going to school, to work, or even to movies, thereby affecting our education, income, leisure.

3.2 HEALTH AS CONSUMPTION GOOD AND INVESTMENT GOOD

Economists view health as a durable commodity or a type of capital from which a flow of services are produced as well as consumed. As a consumption commodity, as sick days create disutility, it directly enters into the consumer's utility function. It gains cognizance as an investment commodity by reducing the loss of time for other market and non-market activities. The monetary value received from the activities with this increased time can be viewed as the rate of return on this investment (Grossman, 1972).

Each person is endowed with a given stock of health at the beginning of his life. This stock of health, like any other capital, depreciates with age. This may be replenished by investments in healthcare services. Death occurs when this stock of health falls below a critical minimum level. The initial stock of health and the rate of depreciation are different for each person. It depends on many economic and non-economic factors some of which are uncontrollable. For instance, a person may be born with some genetic disease (e.g. Thalassemia or Congenital heart defect). This means such persons would have a lower initial stock of health than others. Although a person has no control on his initial stock of health at birth, healthcare services can replenish the deficiencies to some extent. The rate at which stock of health depreciates depends on many factors like individual's age, physical structure, lifestyle, environmental factors and the amount of healthcare services consumed. For instance, the rate at which health depreciates in a person diagnosed with high blood sugar would depend on the amount of healthcare services received (e.g. does one take insulin regularly?), environmental factors (does one have a stressful occupation?), and lifestyle (does one consume alcohol or have an obesity problem or does one exercise regularly?). All these factors determine the person's stock of health at any point of time as also the rate at which it depreciates.

People demand good health (and education) as a commodity both for consumption and investment purposes. From consumption point of view, an individual desires good health because it gives psychological satisfaction by increasing the quality of life. As an investment commodity, good health means less time of sickness and more healthy days which a person can devote to more work-days or in leisure time. The investment aspect of health can be illustrated with a person preferring to exercise in the morning compared to sleeping for some more time. Here preference for having more healthy days in future outweighs the utility obtained from enjoying additional sleep. Therefore, a person will consume the commodity bundle, including health, which will give one maximum utility. Fig. 3.1 shows the relationship between a person's stock of health and total utility derived from it. The upward sloping total utility curve explains the positive relationship between a person's stock of health and his utility (assuming that the consumption of other commodities in the consumption bundle remains same). The curve is concave to the horizontal axis which is due to the law of diminishing marginal utility. This law tells that the extra utility received from the consumption of an additional unit of any commodity gradually reduces with more and more consumption of that commodity. In Fig. 3.1, an increase in the stock of health from H_0 to H_1 increases the total utility from U_0 to U_1, whereas, an increase in stock of health from H_2 to H_3 results in a smaller increase in utility from U_2 to U_3. However, this does not mean that every individual receives

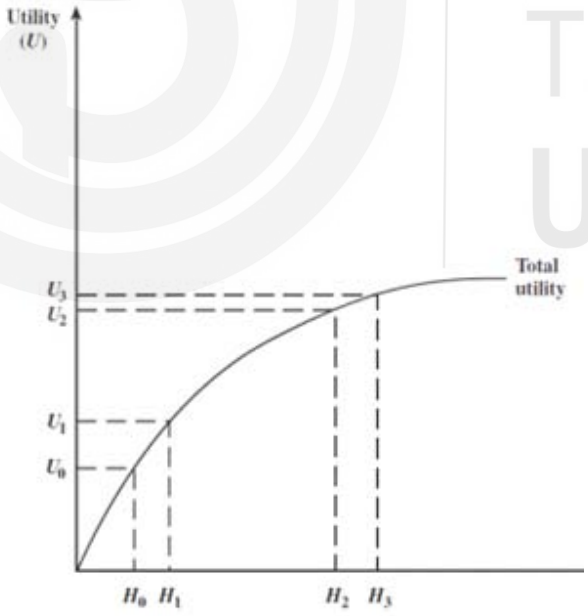

Fig. 3.1: Total Utility curve of Health

Source: Health Economics, Neun and Santerre, 5th edition.

same amount of utility from the same stock of health even though the law of diminishing marginal utility applies to everyone. In other words, every individual has a different total utility curve of health though with similar properties. Another way of expressing the *law of diminishing marginal utility*

is by looking at the marginal utilities for each unit of the stock of health. Marginal utility, being the addition in the total utility for each extra unit of stock of health, can be mathematically expressed as:

$$MU_H = \delta TU/\delta H > 0, \text{ but } \delta^2 TU/\delta H^2 < 0 \quad (3.1)$$

In Fig. 3.1, diminishing marginal utility is depicted by the bowed down (concave to the horizontal axis) shape of the total utility curve, whereas in Fig. 3.2, it is shown by the negative slope of the marginal utility curve.

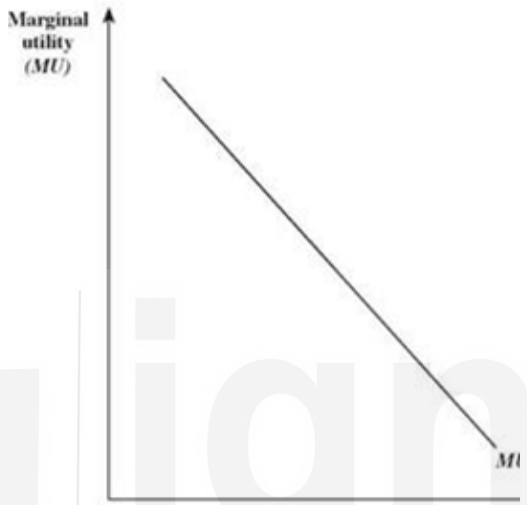

Fig. 3.2: Marginal Utility curve of Health

Source: Health Economics, Neun and Santerre (5th edition).

3.2.1 Production of Health

Success of modern medical science and treatment has led the health economists to the idea that good health can be both produced and restored with appropriate means. This is therefore like a production process in which certain inputs are used to get a state of good health. The question that arises is 'who is the producer here'? Since whatever medicines and treatment are applied, the ultimate healing and recovery process occurs inside the body of the individual, the individual is both the producer as well as the consumer of good health. In other words, both the production and consumption occur within the human body itself. Hence, just like a firm uses various inputs to produce an output, an individual can use many inputs such as medical assistance, healthy lifestyle, etc. to maintain/produce good health. A health production function can then be defined as the production of maximum amount of health an individual can achieve from a specific set of inputs within a given period of time. It can be expressed in the following form:

Health = H (healthcare services, lifestyle, socio-economic status, technology, profile, environment, etc.) (3.2)

where *Health* is the level of health at a point of time, *healthcare services* are the quantity of healthcare services consumed, *lifestyle* represents a set of lifestyle variables (e.g. diet, exercise, etc.), *socio-economic status* represents the joint effect of social and economic factors like education, income, etc., *technology* refers to the status of medical technology available at a given point of time, *profile* means the individual's mental and physical status at a point of time and *environment* accounts for a variety of factors like air, water quality, etc. Lifestyle affects the production of health through habits like smoking, drinking, consumption of alcohol, being overweight or under nourished, etc. Likewise, a person with good education can appreciate the effect of balanced diet on health. However, the causation between education and health could be negative too (e.g. a child with poor health can miss education) as health status itself determines the level of education received.

Thus, an individual would demand the inputs for better health. However, while some of these inputs are quantifiable in nature, some others are not. In other words, health being intangible in nature, a person's demand for health is difficult to determine. But an idea on the extent to which health as a commodity is being produced and consumed by an individual can be had by analysing the demand for its inputs.

3.2.2 Healthcare Services: Characteristics and Measurement

There is a difference between *'health care'* and *'healthcare'*. *Health care* refers to a set of actions by one individual to improve ones health or of some individuals to improve the health of a patient. But *healthcare* refers to a system or industry that facilitates with the logistics and delivery of health care for patients or consumers. Healthcare services can be divided into three categories viz. primary health care [i.e. prevention of health problems before they occur (e.g. disease prevention, immunisation, etc.)] which is also known as preventive care; secondary health care defined as early detection of diseases and intervention (hence also termed as curative care); and tertiary health care referring to correction and prevention of deterioration and rehabilitation of terminally ill patients. It must be noted that although health care and medical care are used interchangeably they are two distinct concepts. While medical care is applied to a person by a doctor (or other health personnel when ill), health care is applied to a person as a preventive measure (e.g. vaccination of a child, proper diet plan for an adult, etc.).

As a commodity, health care services have some distinct characteristics. Firstly, there are many types of health care services heterogeneous in nature making it very difficult to measure them. Some of them are *intangible* i.e. they cannot be felt by seeing, hearing, touching, tasting or smelling. Sometimes, they are also *inseparable* in nature i.e. production and consumption has to take place simultaneously (e.g. doctor's consultancy). A patient often acts as a producer and a consumer simultaneously (e.g. awareness about health and hygiene learnt from internet and then applied to oneself). Without active participation of a patient, health care services cannot

be produced and consumed. With services that are inseparable, there is no scope of *inventory* or stock piling. Finally, health care services are often characterised by *inconsistency* i.e. same amount of service may generate different outcomes for different people (e.g. same hours of doctor's service may result differently for two different patients). These characteristics are referred to as the *four I's* of health care services.

Although the aspect of quality is of crucial importance in health care services, quality is difficult to measure. This is because, there may be differences in *structural quality* (i.e. type and age of medical equipments, training level and experience of personnel, etc.) or differences may lie in *process quality* (i.e. waiting time, nurse's behaviour with patients, etc.) or there may be differences in *outcome quality* (i.e. post-treatment mortality rate, patient's satisfaction about treatment, etc.). Due to these features of highly non-quantifiable nature of health care services, health economists often measure health care services by either *availability* or *utilisation*. *Availability* considers measures like number of beds or doctor per thousand population while *utilisation* reflects how often the service is actually delivered (e.g. number of inpatient days of a hospital, number of hospitalisation cases in a year, etc.).

Check Your Progress 1 [answer within the space given in about 50-100 words.

1) How is health an investment good?

.....

.....

.....

.....

.....

2) What are the factors that determine the health stock of a person?

.....

.....

.....

.....

.....

3) Define (and specify) a health production function.

.....

.....

.....

.....

4) Why is health said to be 'intangible' in nature? What is its consequence? How can this be overcome?

.....

.....

.....

.....

.....

5) Distinguish between the terms: health care and healthcare.

.....

.....

.....

.....

.....

6) How is health care different from medical care?

.....

.....

.....

.....

.....

7) State the four distinct characteristics of 'health care services' as a commodity.

.....

.....

.....

.....

.....

8) What are the three identified difficulties of measuring the quality of health? Due to these factors, what is the approach generally followed by health economists?

.....

.....

.....

.....

.....

.....

3.3 DEMAND FOR HEALTHCARE SERVICES

Health care services are critical inputs for the health production function. Health is a durable good (an investment good) which enhances the utility level of the consumers. Thus, health care services, like any other commodity, maximises utility even though not directly but indirectly.

3.3.1 Utility Maximisation

In the context of health care services, applying the standard neo-classical utility function, we can express the total utility function of a consumer as:

$$U = U(X_1, X_2, ..., X_n, HC) \quad (3.3)$$

Equation (3.3) shows that utility depends on the amount of commodities consumed in which health care services (HC) also figures as an input. With this, we shall state the demand curve for health care services exactly as in a Marshallian approach. We are aware that the Marshallian demand curve rests on the basic proposition that given the market prices and fixed income of consumers, a consumer purchases that bundle of commodities which maximises his total utility. The total utility is maximised when the marginal utility from the last rupee spent on a commodity is equal to that of any other commodity. This is the *utility-maximising rule* which implies that total utility becomes maximum when the marginal utility per unit of money spent is equal for each or any commodity. In other words:

$$MU_{HC}/P_{HC} = MU_X/P_X \quad (3.4)$$

where MU_{HC} represents the marginal utility received from the last unit of health care service purchased and MU_X is the marginal utility received from the last unit of any other good and P_{HC} and P_X are the respective fixed unit prices. Equation (3.4) thus gives the equilibrium condition for an individual consumer. If $MU_{HC}/P_{HC} > MU_X/P_X$, then the last rupee spent on health care services will generate higher utility than the last rupee spent on other commodities. If this happens, the consumer can increase his total utility by increasing his spend on health care services than on other commodities. Due to law of diminishing marginal utility, marginal utility of health care services falls as consumption of the same increases and marginal utility of all the other commodities will rise as their consumption falls. This continues till, eventually, they become equal i.e. condition (3.4) is attained. At this point, the marginal utility generated from the consumption of last unit of each commodity is equal and the total utility is maximum.

3.3.2 Hicksian Analysis

For the consumer to attain his equilibrium condition, let us assume that there is an initial equilibrium in which the condition (3.4) holds so that the consumer consumes the optimal mix of health care services and all other commodities. Let there be a rise in the price of health care services. As a

result, MU_{HC}/P_{HC} will become less than MU_X/P_X (since P_X remains constant). Therefore, the marginal utility received from last rupee spent on other commodities is now greater than that of health care services. The consumer will react by reducing the consumption of health care services and increasing that of other commodities. This will again continue until MU_{HC}/P_{HC} becomes equal to MU_X/P_X so that the equilibrium condition is restored. This means, there exists an inverse relationship between the price and the consumption for the quantity demanded of the health care services like any other good. This can be illustrated graphically (Fig. 3.3). In Fig. 3.3, price of health care services is measured along the vertical axis and the quantity demanded measured along the horizontal axis. As the price of health care service falls from P_0 to P_1, the quantity demanded increases from Q_0 to Q_1. This inverse relationship can be explained by Hicksian analysis of price effect as per which the quantity demanded changes as an effect of a price change due to two effects viz. substitution effect and income effect. As price falls, both these effects lead to a rise in quantity demanded. Due to substitution effect, a fall in the price of the health care services will lead the consumer to substitute his other high priced commodities by the now relatively cheaper health care services. As a result, quantity demanded of health care services increases. By income effect, the lower price of health care services will increase the real income of the consumer. Since health care services is considered to be a normal good (because people increase its purchase if income increases), the consumer increases his demand for health care services.

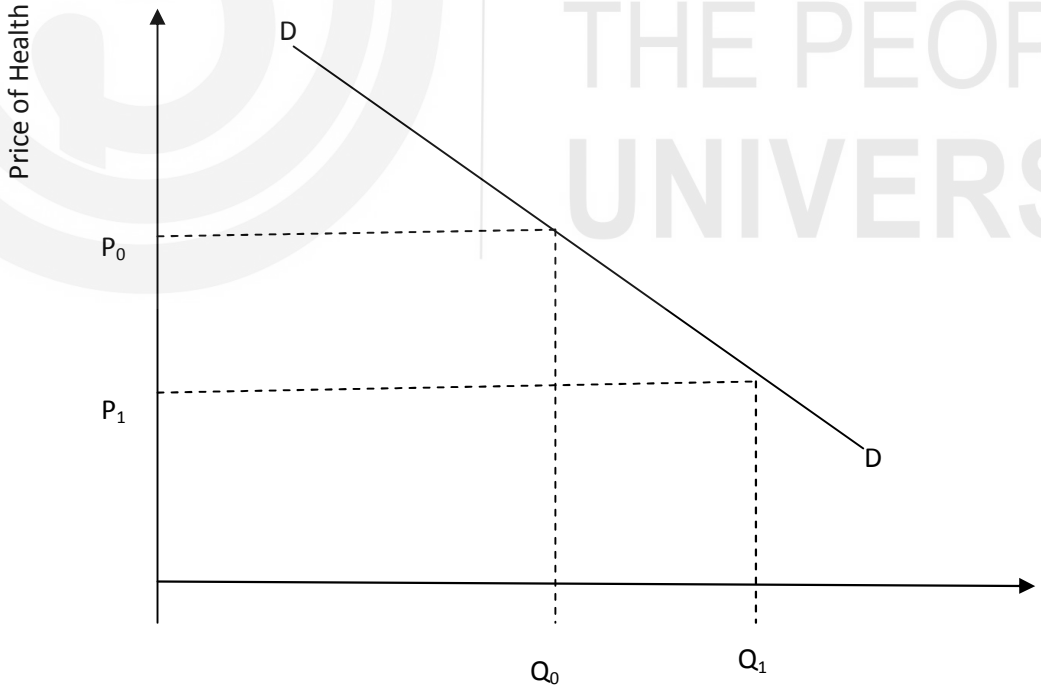

Fig. 3.3: Demand for Health Care Services

Taken together, these effects imply that the consumer will increase his demand for health care services whenever price of the same falls. This inverse relationship between the price and the demand for health care

services is nothing but the *law of demand*. However, this demand for health care services is a derived demand because it depends on the person's demand for good health.

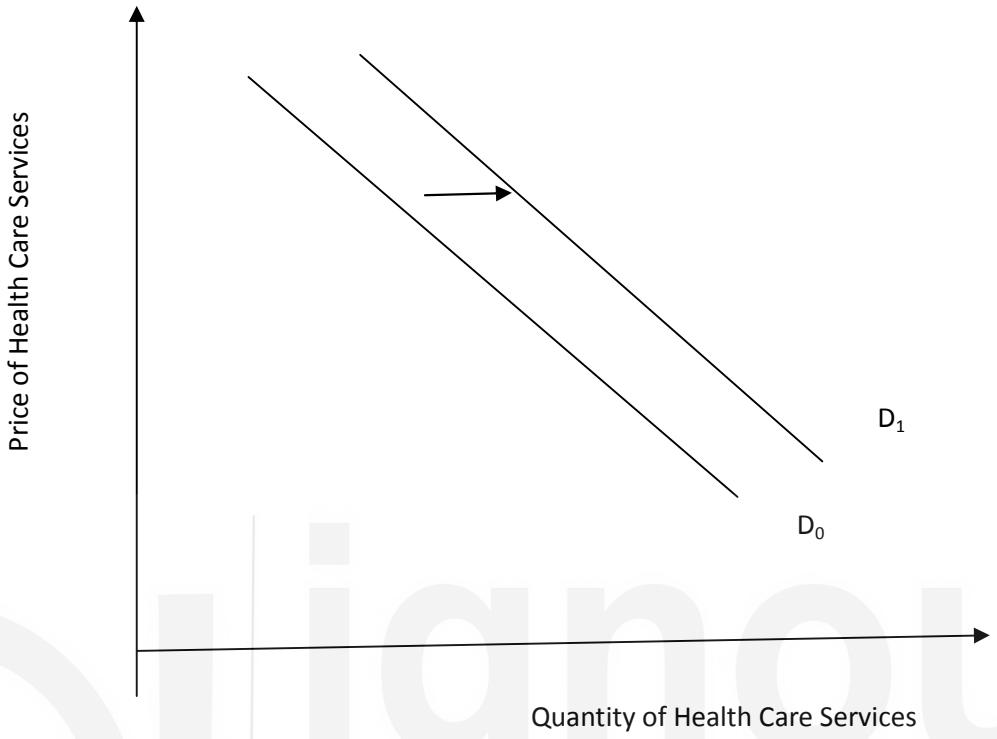

Fig. 3.4: Shift in Demand Curve for Health Care Services due to Increase in Income

3.3.3 Determinants of Demand for Health Care Services

Among the economic factors determining the demand for health care services, apart from the price of the health services itself, the prices of the complementary and substitute commodities and consumer's *income* are the main factors. Since, health care services are supposed to be like other normal commodities, a rise in income will result in a rightward shift in the consumers' demand curve (Fig. 3.4). With increased income, at every price point, the consumer will demand more quantity of health care services. Further, like income, the prices of other related commodities, both *substitutes and complements*, would also influence the demand for health care services. For instance, the *social care* provided to the society is one of the substitutes of health care services. If the government decides to reduce old age pension, it leads a person to spend more on health care. Commodities like *food* acts as complement to health care services in the sense that increase in the prices of vegetables could lead a consumer to change his consumption and lifestyle habit. This rise in food prices might have the effect of causing a reduction in the health care services availed. Among the other economic factors which affect the demand for health care services, *time* is an important factor. We can measure the time cost economically by taking into account the travel cost for visiting the hospital, waiting time at the hospital, the loss of wage due to this, etc. As these increase, people tend to reduce their demand for health

care services. Likewise, *health insurance* plays a crucial role in determining the demand for health care services. It acts as a subsidy thereby reducing the price and increasing the demand for it. However, in case of full insurance coverage, health care services become practically free good.

Among the non-economic determinants of demand for health care services, *taste and preferences, individual profile and state of health* are important. Taste and preference depends on factors like education, lifestyle, etc. *Education* has a positive effect on demand for health care services as a person with more education is likely to seek health care more frequently due to his awareness of its importance. *Lifestyle* factors like smoking or drinking habits affects the demand for health care services. A person with these habits tend to seek more health care services to compensate the bad effects of these habits on his/her health. *Profile* factors like gender, race, age, etc. also affect the demand for health care services. For instance, women may need more health care services during child bearing. Also, certain diseases (e.g. osteoporosis, rheumatoid arthritis, etc.) are found to have more prevalence in women. *Age* is a determining factor because as a person ages, one tends to need more health care services. *State of health* affects the demand for health care services in the sense that a person demands more health care services when sick which increases with the severity of the illness. Considering both the economic and non-economic factors, the individual demand function for health care services can be functionally expressed as:

Quantity demanded of health services = *f* (income, time, prices of substitute and complements, taste and preferences, profile, state of health, etc.)

3.3.4 Fuzzy Demand or Market Demand

The market demand for health service is the total of the individual demand by all consumers in a market. Graphically, the market demand curve is the horizontal summation of the individual demand curves for health care service. For instance, if the price per visit of a doctor is 300 INR and patient A is willing to pay five visits and patient B is willing to pay four visits to the doctor in a year, then the market demand for visits to that doctor is nine per year. Further, it is also downward sloping. Therefore, eventually, the factors which cause a shift in the individual demand curve also shifts the market demand curve in the same direction. In other words, the market demand curve also shifts rightwards or leftwards depending on whether the number of consumers in the market increases or decreases.

However, in reality, the market demand curve for health care services does not look like the standard market demand curve since the demand and price relationship in the health market is not precise. The demand for health care services being a derived demand, the relation between health and health care services is far from exact. There may be lack of knowledge among the health

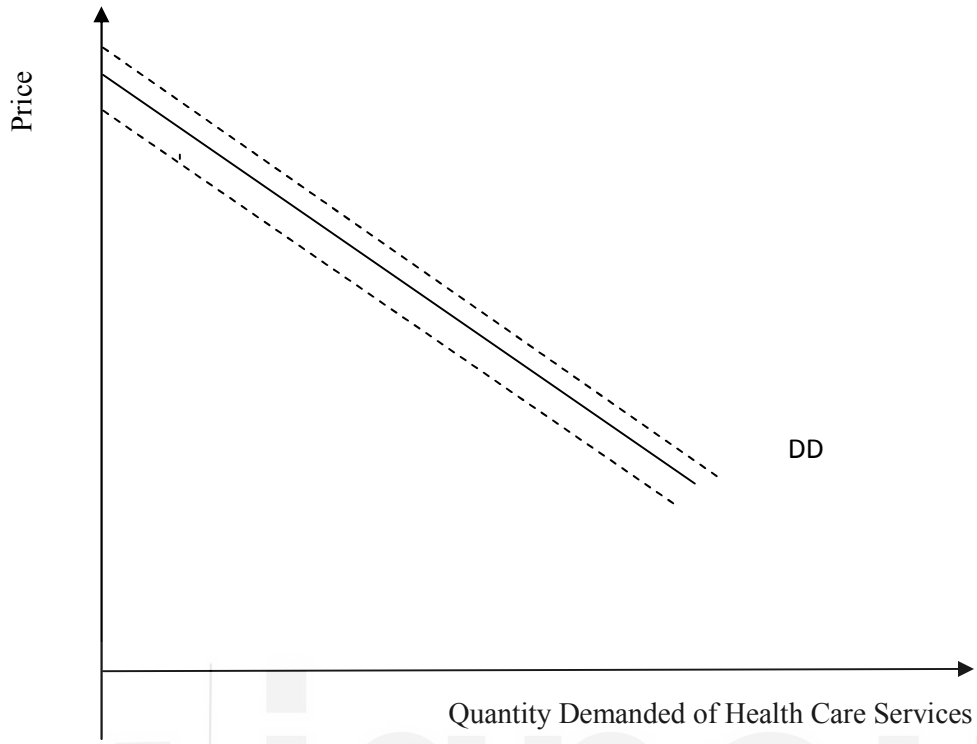

Fig. 3.5: The Fuzzy Demand Curve for Health Care Services

care providers about the efficacy of some health care services. Hence, there may be disagreement about the treatment required. For instance, doctors may disagree about the need for surgery for elderly patients. There may also be lack of information among the consumers leading to heavy dependence on doctors. Further, in some cases, health care itself may be difficult to measure. All these aspects make it extremely difficult to accurately delineate the price and demand relationship. Hence, a fuzzy-looking market demand curve for health care services (Fig. 3.5) results where the fuzziness of the relation between price and demand is depicted by a band rather than a well-defined line.

3.4 DEMAND ELASTICITY

Equation (3.5) tells us what factors influence the demand for health care services. By how much it will rise can be measured by demand elasticities. In other words, elasticities tells us the extent by which these factors influence the demand.

3.4.1 Price Elasticity

Price elasticity of demand measures the influence of price (i.e. price of the commodity itself) on its demand i.e. the extent by which consumers change their consumption of it, if its own price changes. It is defined as the percentage change in demand caused by one percent change in own price. That is:

$$E_p = \frac{\%\ change\ in\ quantity\ demanded}{\%\ change\ in\ own\ price}$$

$$= \frac{\left(\frac{\Delta Q}{Q}\right) \times 100}{\left(\frac{\Delta P}{P}\right) \times 100}$$

$$= \frac{\left(\frac{\Delta Q}{Q}\right)}{\left(\frac{\Delta P}{P}\right)}$$

$$= \frac{\Delta Q \times P}{\Delta P \times Q} \qquad (3.6)$$

where E_p is the price elasticity of demand, Δ is the change, Q is the quantity demanded and P is the price of the commodity. Since E_p is a ratio, it is scale free. This makes it easier to compare elasticity across different commodities. For instance, we can compare price elasticity of demand for doctor's visit with the hospital care although the former is measured in terms of number of visits and the latter in terms of number of inpatient days.

Due to the inverse relationship between price and quantity demanded, the value of E_p is generally negative. Hence, generally, the absolute value of the elasticity of demand is considered i.e. $|E_p|$. If $|E_p|$ is equal to zero, the demand is said to be *perfectly inelastic* i.e. there will be no change in quantity demanded for a price change. Extremely necessary commodities, or those with no substitutes, have perfectly inelastic demand (e.g. life saving drugs or emergency surgery after an accident). The demand curve will be a vertical straight line in this case. If $|E_p|$ is greater than zero but less than one ($0 < |E_p| < 1$), the price elasticity of demand is *inelastic.* Here, the demand will change as price changes but the percentage change in quantity demanded is less than the percentage change in price. For instance, if the price elasticity of demand for doctor's visit is 0.4, it means that if the doctor's fee per visit increases by 100 percent (i.e. it doubles), the demand for the same will fall by 40 percent. The demand curve could be steep in this case. If $|E_p|$ equals one, it means that the percentage change in quantity demanded will be exactly the same as the percentage change in price. Here, the price elasticity of demand is said to be *unitary elastic*. A rectangular hyperbolic demand curve represents this case. For an *elastic* demand, $|E_p|$ should be greater than one i.e. the percentage change in quantity demanded is greater than that of price. Health care services that are not so necessary (e.g. cosmetic surgery) or with a number of substitutes falls into this category. The demand curve in this case would be flat. Price elasticity is said to be *perfectly elastic*, when $|E_p|$ tends to infinity. In this case, the demand is extremely price sensitive and the demand curve is a horizontal straight line.

3.4.2 Income Elasticity

The *income elasticity of demand* measures the extent to which the demand changes as income of the consumer changes. It is defined as the percentage

change in quantity demanded due to percentage change in consumer's income. That is:

$$\eta = \frac{\%\ change\ in\ quantity\ demanded}{\%\ change\ in\ income}$$

$$= \frac{\left(\frac{\Delta Q}{Q}\right) \times 100}{\left(\frac{\Delta M}{M}\right) \times 100}$$

$$= \frac{\left(\frac{\Delta Q}{Q}\right)}{\left(\frac{\Delta M}{M}\right)}$$

$$= \frac{\Delta Q \times M}{\Delta M \times Q} \qquad (3.7)$$

If η is zero, it means quantity demanded does not change as income changes. However, generally, η is positive as people tend to increase their demand for health care services as income increase. The commodity is said to be *normal*, if η is greater than zero, but less than one ($0 < \eta < 1$). In this case, the percentage change in quantity demanded is less than the percentage change in income. If η is greater than one, the percentage change in quantity demanded is greater than that in income and the commodity is said to be *superior.* Demand is *perfectly income elastic,* if η tends to infinity. When consumers reduce the demand for a commodity with an increase in income, η is negative and the commodity is referred to be *inferior*. Visit to an MBBS general practitioner may be considered as inferior to that of a specialist doctor as consumers may switch from the former to the latter as their income increases.

3.4.3 Cross-Price Elasticity

Cross-price elasticity of demand measures the extent to which demand for a commodity changes as a result of a change in the price of its related commodities. It is:

$$E_{xy} = \frac{\%\ change\ in\ quantity\ demanded\ of\ X}{\%\ change\ in\ price\ of\ Y}$$

$$= \frac{\left(\frac{\Delta Q_X}{Q_X}\right) \times 100}{\left(\frac{\Delta P_Y}{P_Y}\right) \times 100}$$

$$= \frac{\left(\frac{\Delta Q_X}{Q_X}\right)}{\left(\frac{\Delta P_Y}{P_Y}\right)}$$

$$= \frac{\Delta Q_X \times P_Y}{\Delta P_Y \times Q_X} \qquad (3.8)$$

If as a result of the change in the price of a substitute commodity Y, the quantity demanded of *X* changes in the same direction, then E_{xy} is positive. This is the case when the price increase of the medicine of a particular brand

increases the demand for itself. E_{xy} is negative if a change in a complementary good changes the quantity demanded of a good in the opposite direction (e.g. if the price for eye glasses increases, the demand for optometrist's service may drop). If E_{xy} is zero, then the commodities are unrelated and the change in one's price does not affect the quantity demanded of the other. For instance, increase in the price of the eye glasses does not influence the demand for a dentist's service.

Check Your Progress 2 [answer within the space given in about 50-100 words]

1) State the two conditions of Hicksian analysis in the context of health care services.

.....

.....

.....

.....

.....

2) Apart from the price of health care services itself, which are the other two major factor which influence the demand for health care services?

.....

.....

.....

.....

.....

3) State the four economic and three non-economic factor which together determines the demand for health care services.

.....

.....

.....

.....

.....

4) Why is demand for health care services called as fuzzy demand? What factor contributes to it being so?

.....

.....

.....

.....

3.5 LET US SUM UP

Health, as a commodity, has both the components of a product and a capital i.e. for an individual it is both a product to consume as well as a means to invest in health. Neoclassical demand function cannot fully explain health. Health status because health status or level of health is not easily measurable. Hence, demand for health care service is considered which is a 'derived demand' of demand for health. Besides its own price, both economic and non-economic factors influence the demand for health care services. One can also measure the extent to which these factors influence the demand for health care services with the help of price, income and cross-price elasticities. However, the market demand curve for health care services is different than the standard market demand curve and is fuzzy in nature.

3.6 KEY WORDS

Derived Demand	**:**	Refers to a demand for a commodity which is a consequence of demand for something else.
Fuzzy Demand Curve	**:**	Refers to a demand curve when price and demand do not have exact relationship. In this case, the demand curve looks like a thick band rather than a line.

3.7 SOME USEFUL BOOKS AND REFERENCES

1) Santerre R.E. and Neun S.P. (2010). *Health Economics: Theories, Insights and Industry Studies (5th ed.),* South-Western Cengage Learning, USA.

2) Zweifel P., Breyer F and Kifmann M. (1997). *Health Economics (2nd Ed.),* Springer, OUP.

3.8 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) Sickness is disutility since it reduces time available for many marketable and non-marketable activities. Any investment on health, ensures the reduced loss of time. Thus, investment on health can be viewed by the rate of return it provides in terms of increased availability of time.

2) Health stock at any point of time is linked to factors that allows it to depreciate. These are: age, physical structure, lifestyle and environmental factors.

3) It is defined as the production of maximum health from a specific set of inputs within a given time frame. Functionally, it can be expressed as Equation (3.2) in Sub-section 3.2.1.

4) Health is said to be intangible as it cannot be felt by touching, seeing, hearing, etc.). In other words, it cannot be easily quantified. A consequence of this is the difficulty in its measurement. This can be overcome by treating health as a commodity and then considering the various inputs that goes into its production.

5) Health care refers to a set of action to improve one's health. It can be by an individual by devoting time required to keeping himself healthy or a person or a group of persons who individually or collectively work to improving the health status of a person. Healthcare, on the other hand, refers to a system or industry which supplies the logistics and delivery of health care (e.g. surgical instruments industry, pharmaceutical industry, etc. are part of healthcare sector).

6) Health care has a preventive character to thwart the setting-in of a disease (e.g. vaccination). Medical care is provided by a doctor (and other qualified persons like nurses) after the setting-in of a disease.

7) Intangible, inseparable and hence no scope for inventory or piling up (i.e. they cannot be stored) and inconsistency (i.e. they produce different results for same input in terms of health output). These are called as the '4 Is' of health care services.

8) Differences in structural quality, process quality and outcome quality are the three identified factors. Availability and utilisation are used as proxies to measure quality of health services by professionals in the field.

Check Your Progress 2

1) The Hicksian analysis relating to the effect on price of a commodity consequent to change in price of the commodity, is explained by two effects viz. the substitution effect and the income effect. Taken together, the effect of these two factors simply means that the consumer will increase his demand for health care services whenever price of the same falls.

2) Price of substitute and complementary products and the income of the consumer.

3) Social care provided by the government (e.g. pension), food, time and health insurance are the four economic factors. Tastes and preferences, individual profile and state of health are the three non-economic factors.

4) The demand and price relationship in the health care sector is not precise. It is actually 'derived demand', derived from factors both economic and non-economic extremely heterogeneous in their character. Due to these factors, the demand curve is not a well defined line and hence needs to be depicted as a band.

UNIT 4 SUPPLY OF HEALTHCARE SERVICES

Structure

4.0 OBJECTIVES

After reading this unit, you will be able to:

- state the basic features of factor markets of healthcare services;
- specify the equilibrium condition for non-physicians in the input market structure of healthcare sector distinguishing them from that of physicians;
- discuss the consequences of monopsony power in the input market of healthcare sector;
- describe the equilibrium conditions of a physician under: (i) pure monopoly; and (ii) price discriminating monopoly;
- outline the advantages of competition among price discriminating among physicians;
- indicate how monopolistic competition among physicians services lead to long run stability in the price for physicians services;
- explain the features of production function of healthcare services;

- derive the average and marginal cost function for healthcare services both for short-run and long-run; and
- distinguish between the concepts of 'economies of scale' and 'economy of scope' in the context of market for supply of healthcare services.

4.1 INTRODUCTION

Healthcare encompasses a lot more than just medical care. One can avoid medical care (or medical expenses) if basic healthcare services are available (e.g. different immunisation programmes initiated by a government). Healthcare includes a host of social welfare programs contributing to keeping people healthy. Healthcare falls under the category of merit good (or near public good) and therefore calls for government intervention or regulation. The market price of healthcare is difficult to determine due to imperfect market characteristics of healthcare sector. These are: (a) healthcare is a heterogeneous product as a range of outcomes could be there for different patients; (b) patients who are insured by third-party payers would behave differently than those without such coverage; and (c) the feedback mechanism on a standard 'market price' is missing due to which there is scope for malpractice (e.g. a hospital which has invested in machinery and equipments might order treatments with the sole objective of recovering the sunk cost).

Healthcare services follow certain characteristics like (a) intangibility (i.e. healthcare services are incapable of being assessed by the five senses), (b) inseparability (i.e. consumption and production of healthcare services take place simultaneously) and (c) inconsistency (i.e. the composition and quality of healthcare services consumed vary widely across medical events). Thus, unlike normal goods, consumers seeking healthcare services are less informed in general. The production and supply of healthcare is also therefore quite complex as compared to normal goods and services.

4.2 FACTOR MARKET OF HEALTHCARE SERVICES

Supply of healthcare services is multi-product and interdependent. This is because it includes the labour time of multitude of trained professionals like physicians, surgeons, specialists, nurses, medical technicians and pharmacists. Lack of coordination of any one of these critical services places the patient's life in danger making the healthcare services mutually interdependent and complex in its character. In other words, healthcare services are not homogeneous as they are composed of interdependent multiple inputs.

4.2.1 Structure of Input Market

In standard microeconomics, if the product and input market is perfectly competitive, then the firm (i.e. a nursing home or hospital) will use input up

to that level where the value of marginal product is equal to wage rate. This means:

$$\pi = TR - TC \tag{4.1}$$

where π stands for total profit, TR = total revenue, TC = total cost. Since the total cost comprise of labour (L) and capital (K), if we assume that the two input markets are perfectly competitive, then Equation (4.1) becomes:

$$\pi = P.f(X) - wL - rK \tag{4.2}$$

where, P is the unit price of healthcare services, $f(X)$ is the production function of healthcare, w stands for total wage, r is the interest rate on capital, L and K are the amount of total labour and capital invested respectively. Thus, in order to maximise π, the first order condition yields:

$$\frac{\delta\pi}{\delta L} = P.\frac{\delta f}{\delta L} - w = 0 \tag{4.3}$$

Equation (4.3) implies that $P.MP_L = w$ (where MP_L is the Marginal Productivity of Labour given by $\frac{\delta f}{\delta L}$). The healthcare provider will hire physicians up to that level where the 'value of marginal product' (VMP) of labour (physician) is equal to wage rate. Graphically, the equilibrium condition is attained at VMP as in Fig. 4.1. The healthcare provider (hospital) will employ OL^* number of labour units in order to maximise profit. At OL^*, the value of marginal product ($P.MP_L$) is equal to the wage rate OW. Under perfect competition in both the product (healthcare) and factor (labour) market, the exploitation of labour will not occur.

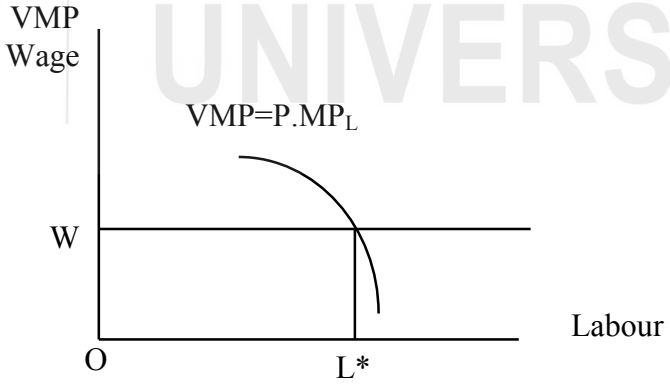

Fig. 4.1: Equilibrium under Perfect Competition

The above equilibrium condition is valid in case of hiring nurses and other medical staff but not in case of doctors or physicians. This is because the input market of physician is more heterogeneous in terms of skill, experience and expertise. Physicians further enjoy market power unlike nurses and other non-medical staff who cannot provide services independently of doctors. However, there may be monopsony power in markets for nurses in case of single buyer hospital.

4.2.2 Monopsony Power in Input Market

Monopsony occurs when there is a single buyer in an input market. This is similar to the concept of monopoly when there is a single seller in a product market. Buyers may possess monopsony power in an input market if the mobility of the input is restricted to that market. Thus, if a hospital enjoys monopsony power in the input market for nurses, then the wage rate of nurses would be lower than what would prevail under perfect competition. This is indicated by the downward sloping curve EF in Fig. 4.2 which represents the 'marginal revenue product' MRP which is obtained as: MR.MPP [i.e. product of marginal revenue (MR) of healthcare services and 'marginal physical product' (MPP)]. Since the price of healthcare (Px) is higher than MR in an imperfectly competitive market, MRP is less than 'value of marginal product' (VMP). But under perfect competition, VMP = MRP.

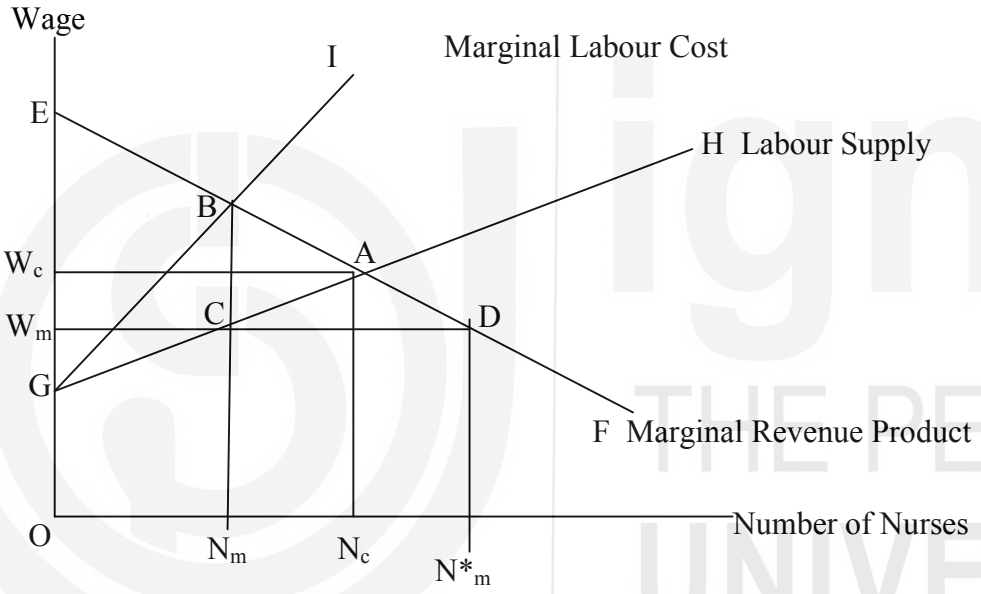

Fig. 4.2: Monopsony in Health Market

The lower upward sloping curve (GH) is the supply curve of labour (viz. nurses). The single employer (hospital, in our case) realises that to attract more nurses, it has to offer a higher wage rate. Thus, the 'marginal labour cost' (MLC) curve viz. GI lies above the supply curve. The MLC (GI) curve determines how many nurses are to be employed by the hospital. Since wage rate *w* is a function of labour L, from Equation (4.2), we can write the profit function as:

$$\pi = P.X - w(L).L - rK \qquad (4.4)$$

To maximise π, with respect to *L*, the first order condition gives us:

$$\frac{\delta\pi}{\delta L} = P.\frac{\delta X}{\delta L} + X.\frac{\delta P}{\delta X}.\frac{\delta X}{\delta L} - \left(\left(w(L) + L.w'(L)\right)\right) = 0$$

$$\Rightarrow P.MP_L + X.\frac{\delta P}{\delta X}.MP_L = MC_L \qquad (4.5)$$

$$\Rightarrow \left[MP_L \left(P + X.\frac{\delta P}{\delta X} \right) \right] = MC_L \qquad (4.6)$$

Since, $P = f(X)$, $X = X(L)$, assuming only L to be the variable cost and identifying $MR = \frac{\delta}{\delta X}(TR) = P + X.\frac{\delta P}{\delta X}$, (4.6) can be written as:

$$MP_L * MR_X = MC_L \qquad (4.7)$$

This means that under equilibrium the firm (hospital) will hire labour (nurse) up to the level where MRP is equal to marginal cost of input for nurses. Therefore, the hospital will employ N_m number of nurses at wage rate W_m. But, at W_m, the demand for nurses is N^*_m, and there is a shortage of labour equal to $(N^*_m - N_m)$. If the monopsonist hospital desires to hire additional nurses, it cannot do so until it raises the wage beyond W_m which reduces its profit. Therefore, in a monopsonist market, the equilibrium is not efficient because: (i) vacancies persist; (ii) employment would be less than that in a competitive market; and (iii) the equilibrium wage W_m would be less than MRP as shown by the vertical distance between points B and C in Fig. 4.2. The gap between MRP and wage under monopsony is:

$$\text{Gap (G)} = (\text{MRP} - W_m)/W_m = 1/\varepsilon \qquad (4.7)$$

where ε stands for elasticity of supply of labour to the hospital with respect to wage W_m. 'G' is thus interpreted as a measure of 'exploitation' or hospital's market power in the input market.

Check Your Progress 1 [answer within the space given in about 50-100 words].

1) What is the basic characteristic of healthcare sector making it both different and complex from other normal goods and services?

.....

.....

.....

.....

.....

2) Why is attainment of equilibrium in numbers hired and wages paid possible for nurses but not for physicians in the healthcare market?

.....

.....

.....

.....

.....

3) Why is it in case of a monopsonist market, equilibrium conditions are not met?

.....

.....

.....

.....

.....

4.3 DETERMINATION OF EQUILIBRIUM PRICE FOR PHYSICIANS SERVICES

Physicians play a dual role: to a hospital as inputs and independently as supplier of healthcare services to patients directly. A physician thus enjoys monopoly power in some respects. In this section, we discuss the following three scenarios pertaining to physicians.

4.3.1 Pure Monopoly

The monopolist can set the market price. The profit function of the monopolist is:

$$\pi = TR - TC \tag{4.8}$$

where Total Revenue $TR = P.Q$ with P as the price per unit of supply of healthcare and Q as the quantity of healthcare supplied. The first order condition of profit (π) maximisation requires that MR = MC. Thus:

$$\frac{\Delta\pi}{\Delta Q} = \frac{\Delta TR}{\Delta Q} - \frac{\Delta TC}{\Delta Q} = 0$$

$$P + Q.\frac{\Delta P}{\Delta Q} = MC \tag{4.9}$$

$$P\left(1 + \frac{Q}{P}.\frac{\Delta P}{\Delta Q}\right) = MC$$

Since the price elasticity of demand is given by:

$(E_P) = (-).\frac{\Delta Q}{\Delta P}.\frac{P}{Q}$, Equation (4.9) can be re-written as:

$$P\left(1 - \frac{1}{-E_P}\right) = MC \text{ or } P\left(1 - \frac{1}{|E_P|}\right) = MC \tag{4.10}$$

Thus the change in total revenue is influenced by two opposing forces when a monopolist physician lowers the fees (price) of his healthcare services. In other words, while on the one hand, total revenue increases because the physician is selling more healthcare services, on the other hand, total revenue (TR) falls because the physician has lowered the price. This happens because

of the downward sloping demand curve. The equilibrium of the physician can be graphically illustrated as in Fig. 4.3.

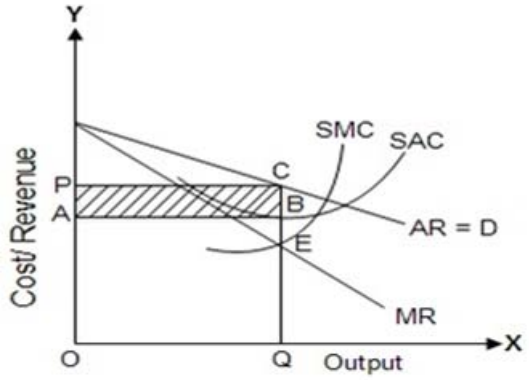

Fig. 4.3: Equilibrium Price and Output of a Monopolist Physician

In Fig. 4.3, the equilibrium condition is at E where MR = MC. The physician charges the price/fee OP and supplies OQ amount of healthcare services, earning ABCP profit (π). Since the equilibrium amount of healthcare services is determined by the intersection of MR and MC, there is no unique supply curve of a monopolist. Further, since π = MR – MC, the second order condition of profit maximisation requires that:

$$\frac{\Delta^2\pi}{\Delta Q^2} = \frac{\Delta MR}{\Delta Q} - \frac{\Delta MC}{\Delta Q} < 0 \qquad (4.11)$$

Thus, slope of MR < Slope of MC or slope of MC > slope of MR. Since at point E, both the first and second order conditions for profit maximisation are satisfied, the rising portion of the marginal cost curve (MC) can be considered as supply curve of a monopolist for healthcare services. In the long-run, the physician earns normal profit when price (P) is equal to the Average Cost (AC). The physician can enjoy absolute monopoly power if he is the sole supplier of healthcare services. This is shown by the Lerner's Index (*LI*) of degree of monopoly power given by:

$$\begin{aligned} LI &= \left(\frac{P - MC}{P}\right) \\ &= \frac{P - MR}{P} \qquad \because MR = P\left(1 - \frac{1}{E_P}\right) \\ &= \frac{1}{E_P} \end{aligned} \qquad (4.12)$$

The Lerner's Index of monopoly power is therefore nothing but the reciprocal of the numerical coefficient of price-elasticity of demand for the product or services. Thus, the less elastic is the demand for the product or services, more would be the degree of monopoly power and vice versa. The economic meaning of this idea is that 'smaller the price-elasticity of demand, smaller would be the response of demand for the product (in response to a

change in its price) and larger would be the power of the monopolist to charge a price in excess of *MC*'.

4.3.2 Price Discriminating Monopoly

We consider here the case of physician as 'price discriminating monopolist'. Price discrimination being the practice of charging a different price for the same good or service, there can be following three types of price discrimination.

- *First-degree* price discrimination, alternatively known as *perfect* price discrimination, which occurs when a physician charges a different price for every unit consumed.
- *Second-degree* price discrimination where a different price is charged for different quantities (such as discounts for bulk purchase of healthcare services). The physician can provide discount to patients in case of multiple medical tests.
- *Third-degree* price discrimination means charging a different price to different consumer groups. For instance, the patients can be divided into two groups on the basis of health insurance. In the same way, patients can also be divided into two groups viz. poor and non-poor. This type of discrimination is the commonest type.

The equilibrium condition of the discriminating monopolist physician can be explained as follows. Suppose, the physician divides the market based on health insurance with TR_1 and TR_2 as the revenues earned by the physician from two markets (patients having health insurance and others not insured). The profit function can be written as:

$$\pi = TR_1 + TR_2 - TC \qquad (4.13)$$

To maximise profit (π), the first order condition yields:

$$\frac{\delta\pi}{\delta Q_1} = \frac{\delta TR_1}{\delta Q_1} + \frac{\delta TR_2}{\delta Q_1} - \frac{\delta TC}{\delta Q_1} = 0 \Rightarrow MR_1 - \frac{\delta TC}{\delta Q}.\frac{\delta Q}{\delta Q_1} = 0 \Rightarrow MR_1 - MC.\left(\frac{\delta(Q_1+Q_2)}{\delta Q_1}\right) = 0$$

$$\Rightarrow MR_1 = MC(1+0) = MC$$

$$\frac{\delta\pi}{\delta Q_2} = \frac{\delta TR_1}{\delta Q_2} + \frac{\delta TR_2}{\delta Q_2} - \frac{\delta TC}{\delta Q_2} = 0 \Rightarrow MR_2 - \frac{\delta TC}{\delta Q}.\frac{\delta Q}{\delta Q_2} = 0 \Rightarrow MR_2 - MC.\left(\frac{\delta(Q_1+Q_2)}{\delta Q_2}\right) = 0$$

$$\Rightarrow MR_2 = MC(0+1) = MC$$

Assuming Q_1 and Q_2 are the quantity of healthcare services supplied by the physician in Markets 1 and 2 respectively so that $Q = Q_1 + Q_2$, the price discriminating physician in equilibrium will equate $MR_1 = MR_2 = MC$. Using the relationship between AR (Price), marginal revenue (MR) and elasticity of demand (E_p) as shown in Equation (4.10), we can write:

$$P_1\left(1-\frac{1}{E_1}\right)=P_2\left(1-\frac{1}{E_2}\right) \quad (4.14)$$

where P_1 and P_2 are the prices charged by the monopolist in the 1st and 2nd markets and E_1 and E_2 are the price elasticities of the 1st and 2nd markets respectively. Noting that if $P_1 > P_2$ then $E_1 < E_2$, we observe that discrimination is only worth undertaking if the profit from keeping the markets separated is greater than from keeping the markets combined. Since this will depend upon the relative elasticities of demand in the sub-markets, consumers in the relatively inelastic sub-market will be charged the higher price and those in the relatively elastic sub-market will be charged the lower price.

Thus, under price discrimination, the physician is able to discriminate between different markets in respect of the price of his healthcare services. From a physician's perspective price discrimination can offer some advantages. These include: profit maximisation, economies of scale and positive externalities. For instance, matching prices to the specific characteristics of the market, and its various segments, is a profit maximising strategy where the physician can extract some of the consumer surplus into producer surplus (i.e. profits). Likewise, given that charging different prices can increase sales volume, especially as a result of new consumers entering into the market attracted by the discounted prices, physicians can benefit from 'economies of scale' which arise from increased output and production. Economies of scale are thus the cost advantages that physicians derive due to their scale of operation with cost per unit of healthcare services decreasing with increasing scale of supply of healthcare services. This line of argument can be extended to consider the role of price discrimination in reducing market failure (e.g. enabling wider consumption of merit goods like healthcare services). For instance, if a private hospital charge relatively high medical fees for those who can afford them where demand is inelastic, the revenue generated allows them to cover their costs and run the hospital. With fixed costs covered, they can then offer services at discounted fees (to cover only the variable costs) to poor patients. Given that the demand for private healthcare by less well-off patients is likely to be price elastic, the lower price will encourage greater demand for healthcare. The benefit to 'society' is therefore that more healthcare services are 'consumed' and more positive externalities are generated.

4.3.3 Competition among Physicians

Monopolistic competition is a market structure which combines elements of monopoly and competitive markets. Essentially, a monopolistic competitive market is one with freedom of entry and exit with physicians providing differentiated services. The physicians therefore have an inelastic demand curve on which they can set prices. However, because there is freedom of

entry, supernormal profits in the short run will encourage more firms to enter into the market leading to normal profits in the long term.

Healthcare service by physicians, in reality, is characterised by monopolistic competition. As a result, each physician's demand curve is negatively sloped. This is mainly because of certain reasons. Firstly, search among doctors is costly for patients. So, if the patient is satisfied with a doctor, the cost of further searching is not worthwhile. Thus, a patient may not leave his or her physician even if the physician's price is slightly above another physician's. A second reason for monopolistic competition among physicians is that people may be willing to pay more for the convenience of visiting a physician who is geographically closer to the patient. Further, there being no barriers to entry, as more sellers enter the market, the demand curve shifts until each seller earns zero profit (Fig. 4.4).

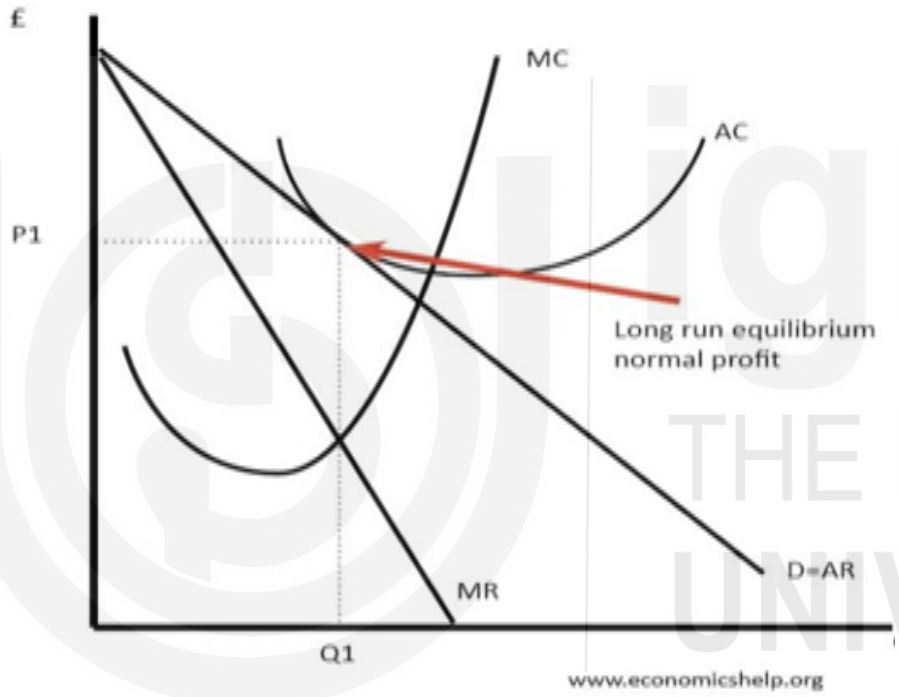

Fig. 4.4: Monopolistic Competition in Physicians Healthcare Market

Thus, under monopolistic competition, since physicians produce differentiated services, they are not at complete liberty to set the prices. In other words, they have inelastic demand. Physicians compete among themselves both for quality of healthcare as also on price. Differentiated healthcare services being a key element of the market, monopolistic competition remains a middle ground in which physicians have the same but relatively low degree of market power even though they are all price makers. In the long run, demand becomes elastic and ensures sensitivity to price changes. In other words, it is only in the short run that economic profit is positive and it approaches zero in the long run.

Check Your Progress 2 [answer within the space given in about 50-100 words].

1) How is the 'total revenue' for a monopolist physician influenced by two opposite forces?

.....

.....

.....

2) What does Lerner's index indicate? On which factor, does this basically depend?

.....

.....

.....

.....

3) What is meant by 'third degree discrimination'?

.....

.....

.....

.....

4) What factor enables the physician to practice 'price discrimination' in different markets? What advantages accrue to physicians by practicing this?

.....

.....

.....

5) How would you distinguish for the short run and long run characteristic of price setting physicians in a market? Why?

.....

.....

.....

4.4 PRODUCTION AND SUPPLY OF HEALTHCARE SERVICES

Medical firms like hospitals earn revenue from producing and selling some type of medical or healthcare output. Production and retailing activities occur regardless of the form of ownership. A comprehensive healthcare production and supply is performed by a well organised hospital.

4.4.1 Production Function

A healthcare production function can be written as:

$$HC = f(P, N, NM, K) \quad (4.15)$$

where HC = Output of healthcare service, P = number of physicians, N = number of nursing staff, NM = number of non-medical staff and K = amount of physical capital. P, N and NM are the inputs referred to as human capital and K is the physical capital (that includes building, medical equipments and machines). If we assume that a given level of output of healthcare (HC) can be produced by using two inputs [human capital (H_1) and physical capital (H_2)], graphically, the efficiency frontier is the lower envelope of the set of production possibilities (Fig. 4.5). This would be associated with a particular type and volume of output of healthcare. The collection of all these input combinations yields the concept of production function.

The efficiency frontier, or isoquant, represents all possible efficient ways to achieve a certain production level. Fig. 4.5 shows two distinct types of isoquants based on two types of production technologies represented by BB*B** and LO*L* (associated with two different types of production functions). If we move along OO*O**, output of healthcare (HC) increases. Therefore, health output produced along the isoquant BB*B** is lower than the output produced along the isoquant LO*L*. The marginal rate of technical substitution (MRTS) is defined as: keeping the total output of health desired to be produced as constant, how much of H_1 has to be decreased if H_2 increases by one additional unit. In other words, it shows the relationship between the two inputs, and the trade-offs amongst them, without changing the level of total output of health. Technically, MRTS amounts to:

$$MRTS_{H_1,H_2} = (-\frac{dH_1}{dH_2}) = \frac{MP_{H_1}}{MP_{H_2}} \quad (4.16)$$

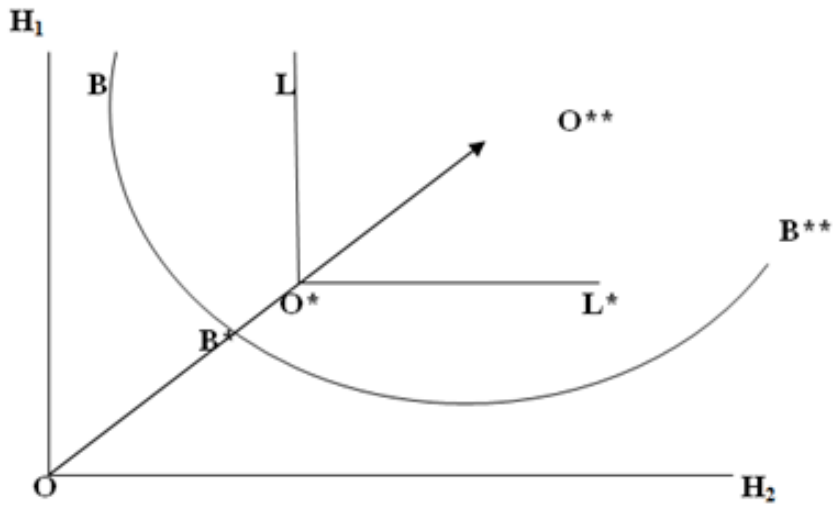

Fig. 4.5: Efficiency Frontier for Differing Production Levels

In Equation (4.16) $\frac{dH_1}{dH_2}$ is the slope of the isoquant and MP_{H_1} and MP_{H_2} are the marginal productivity of H_1 and H_2 respectively. MRTS is thus the negative slope of the isoquant. If $HC = \overline{H}$, the healthcare production function $HC = H(H_1, H_2)$ for fixed level of healthcare output becomes: $\overline{H} = H(H_1, H_2)$. Taking the total differential on $\overline{H}$, we have: $d\overline{H} = \frac{\partial H}{\partial H_1}.dH_1 + \frac{\partial H}{\partial H_2}.dH_2$. Since there is no change in output of health along a fixed isoquant, $d\overline{H} = 0$. Thus, we have:

$$d\overline{H} = \frac{\partial H}{\partial H_1}.dH_1 + \frac{\partial H}{\partial H_2}.dH_2 = 0 \text{ or } \frac{dH_1}{dH_2} = -\frac{\frac{\partial H}{\partial H_2}}{\frac{\partial H}{\partial H_1}} = -\frac{MP_{H_2}}{MP_{H_1}}$$

If H_1 and H_2 are assumed to be perfect substitutes of each other, then the isoquant will be a straight line having a negative slope of –1. This means, if we add one unit of H_1, we must sacrifice the same unit of H_2 (in order to remain on the same isoquant H). In case of the isoquant BB*B**, substitution that takes place between inputs is imperfect. As a result, MRTS tends to zero while diminishing the quantity of H_1 and to infinity while diminishing the quantity of H_2.

If we assume that the healthcare output is produced along the L-shaped isoquant LO*L*, there is no scope of input substitution with the two inputs (H_1 and H_2) being treated as complementary to each other i.e. the two inputs are always used in a fixed proportion. In this case, the horizontal segment of LO*L* has a MRTS = 0 and along the vertical segment of LO*L*, the MRTS = ∞. Since MRTS measures the slope of an isoquant, along LO*, the slope is infinity (since H_1 changes but H_2 remains same along LO*). As a result, $dH_1/dH_2 = \infty$ (infinity) [since $dH_2 = 0$]. Thus, along the isoquant LO*L*, the optimum health supply $H = Min\left(\frac{H_1}{\alpha}, \frac{H_2}{\beta}\right)$, where, α and β are two positive constants. In Fig. 4.5, if the angle L*O*O** is 45º then α = β. In our present case of fixed coefficient type production, the production path is along the locus of line OO*O**. Thus, in case of LO*L* isoquant, no substitution is possible between the two factors H_1 and H_2 i.e. the two factors of production are perfect complement to each other. In other words, if the two factors are used in fixed proportion, we can avoid wastage of inputs either H_1 or H_2.

Practically, the production of healthcare service is fixed coefficient type. For instance, a medical operation requires a specialised surgeon, one anaesthetic and one nurse along with surgical equipments. Therefore, an additional surgeon or nurse becomes redundant. Similarly, production (medical operation) becomes impossible if at least one of the required input combinations is lacking.

4.4.2 Cost Function

Following the discussion in 4.4.1, we now proceed to analyse the cost function of a hospital assuming again that only two inputs are used: human resources (H_1) and physical capital (H_2). With this, we derive the average and marginal cost functions for short-run as well long-run. In the short-run, the 'law of variable proportions' operate (where only one input is considered as variable with the other assumed to be constant), but in the long-run, both the inputs (H_1 and H_2) would vary and the 'law of returns to scale' would operate. For the short-run, we assume that the production function of healthcare can be written as $H = H(H_1)_{\overline{H}_2}$ in which input H_2 is fixed at a certain level and only first input (human capital), H_1 varies.

Law of variable proportion implies that $\frac{\partial H}{\partial H_1} > 0, \frac{\partial^2 H}{\partial H_1^2} < 0$. The total product of healthcare curve looks like inverted U-shaped which means that as H_1 increases total healthcare increases at a diminishing rate and once it reaches a maximum level, it declines. The corresponding marginal productivity also initially increases, reaches a maximum level and then sharply declines. In other words, the marginal productivity of H_1 is zero when the total healthcare product is maximum. The short run average total cost (SRATC) can therefore be written as:

$$SRATC = \frac{C}{H} = \frac{SRTVC + SRTFC}{H} = \frac{SRTVC}{H} + \frac{SRTFC}{H} = SRAVC + SRAFC \quad (4.17)$$

where C = Total Cost, H = Amount of Healthcare, $SRTVC$ = Short-Run Total Variable Cost, $SRTFC$ = Short-Run Total Fixed Cost, $SRAVC$ = Short-Run Average Variable Cost, $SRAFC$ = Short-Run Average Fixed Cost. Graphically, the short-run ATC, AVC, AFC and marginal cost (MC) would be as shown in the Fig. 4.6.

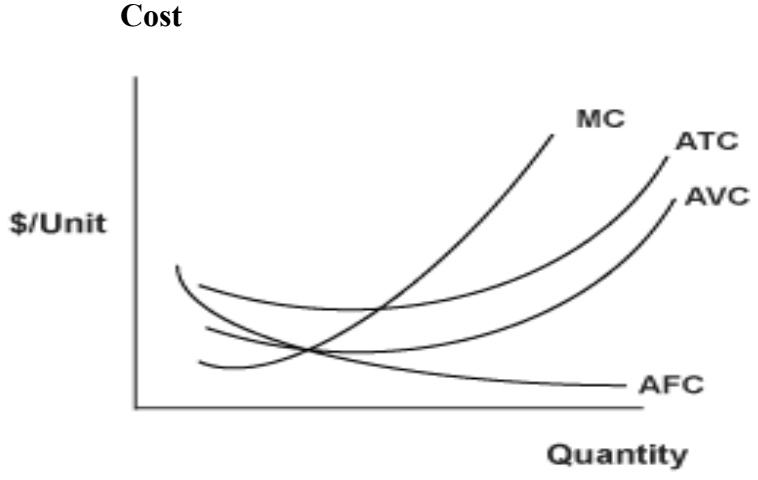

Fig. 4.6: Short Run Marginal/Average/Fixed/Total Cost Curves

As H increases, AFC declines continuously becoming asymptotic to the quantity of healthcare axis (X axis). Given H_2, due to law of variable proportions, in the short-run, the AVC of healthcare reduces because of higher productivity. Since the total cost *(TC)* function consists of total variable cost (*TVC*) and total fixed cost (*F*), in the short-run, the average cost (*AC*) is given by:

$$AC = \frac{TVC + F}{H} = \frac{W.H_1 + F}{H} = W.\frac{1}{\frac{H}{H_1}} + \frac{F}{H} = \frac{W}{AP_{H1}} + AFC \qquad (4.18)$$

where, W = unit price of H_1, H_1 is the number of units of human resources used in producing healthcare services (treated as variable input). The second input H_2, assumed as fixed, requires F amount of cost. The law of variable proportion gives us that initially average productivity of H_1 (AP_{H_1}) increases, reaches to a maximum and thereafter decreases as the unit of healthcare (H) increases. This means that the first term of right hand side of (4.18) falls, reaches to a minimum level then rises. This is the reason why in the short run, the hospital faces a U shaped AC curve. We can establish the relationship between *SRAC* and *SRMC* as:

$$\frac{d}{dH}\left[\frac{C}{H}\right] < 0 \Rightarrow \frac{H.\frac{dC}{dH} - C\frac{dH}{dH}}{H^2} < 0 \Rightarrow \frac{MC}{H} - \frac{1}{H}\frac{C}{H} < 0 \Rightarrow MC < ATC\,.$$

In the same way, we can show that when *ATC* is minimum, $MC = ATC$ and when *ATC* rises, $MC > ATC$.

The long run consists of variable inputs only. The hospital can therefore increase the size of the plant in the long run. Long run total cost refers to the minimum cost of production or the least cost of producing a given level of output. It can therefore be less than or equal to the short run average costs at different levels of output. Similarly, 'long run average cost' (LRAC) can be defined as the average of the LTC curve or the cost per unit of output in the long run. It can be calculated by the division of LTC by the quantity of output.

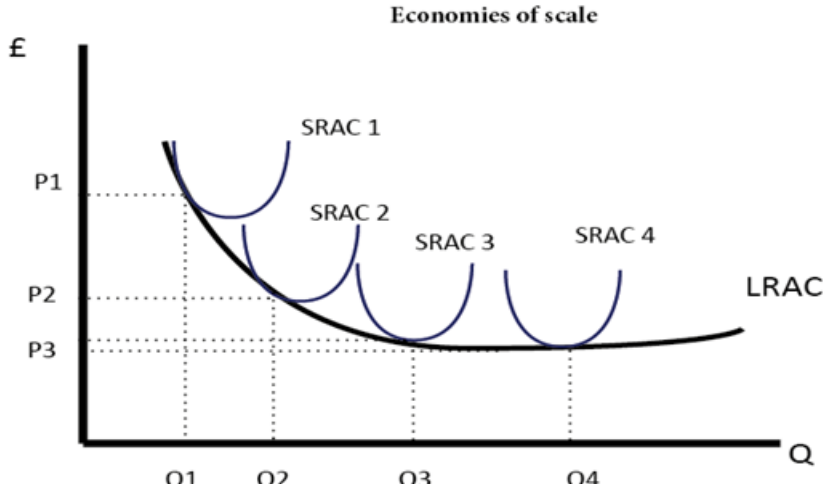

Fig. 4.7: Long Run Average Cost Curves

Graphically, LAC can be derived from the Short run Average Cost (SAC) curves. While the SAC curves correspond to a particular plant (the *plant being fixed in the short-run)*, the LAC curve depicts the scope for expansion of plant by minimising cost. The long-run cost curves are also U shaped due to 'economies of scale'. If a hospital has high fixed costs, increasing output will lead to lower average costs as in the long-run 'returns to scale' operates. This means that in the initial stage, hospital enjoys increasing returns, later constant returns but finally decreasing returns. Thus, the 'long-run average cost' (LRAC) curves can be seen as a lower 'envelope' of short run average cost curves (SRAC) i.e. the curve lies below the short run curves. Any change in inputs in the short run will carry over into the long run as well. It is therefore impossible for average costs to be driven down more in the short run than in the long run. But since the factors that are fixed in the short run can be changed in the long run, average cost (AC) reductions resulting from changing these inputs in the long run will not be felt in the short run. Hence, the long run average costs are lower. For this reason, they are also sometimes called 'envelope curves'.

4.4.3 Economies of Scale and Scope

Economies of scale for a hospital involves reduction in the **average** cost per unit arising from increasing scale of production for a **single** product type. Economies of scope involve lowering the **average** cost by producing more types of products. Economies of scope exist if a firm can produce several product lines at a given output level more cheaply than a combination of separate firms each producing a single product at the same output level. Economies of scope differ from economies of scale in that a firm receives a cost advantage by producing a complementary variety of products with a concentration on a core competency. While economies of scope and scale are often positively correlated and interdependent, strictly speaking, the benefits from scope have little to do with the size of output. Economies of scope and economies of scale are thus two different concepts used to help cut a company's costs. Economies of scope focuses on the average total cost of production of a variety of goods, whereas, economies of scale focuses on the cost advantage that arises when there is a higher level of production of one good.

Economies of scale states that the unit cost decreases as the volume of output increases. This connection can be explained in terms of the effects of volume on fixed cost, specialisation of resources and market position toward suppliers. Economies of scale arise since the initial investment of capital is spread over a larger output with increases in output. Examples are medical equipment such as CT Scan or a computer system for medical records of patients. Another source of economies of scale is that larger firms may have more market power for inputs they purchase. Economies of scope exist when the provision of one type of output makes it possible for another type of output to be produced more efficiently. For instance, consider a radiologist

practising with a general surgeon. There may be efficiency in the diagnosis and treatment of a disease if an accurate diagnosis can be followed by a timely surgical procedure. Cost complementary effect and scale effect are the explained mechanisms behind the scope effect. Cost complementary means that once the inputs have been used for producing one good they become available, at no additional cost, for use in the production of other goods. The scale effect refers to the situation where it is profitable to produce different outputs in one large unit instead of in several specialised production units.

Check Your Progress 3 [answer within the space given in about 50-100 words]

1) In practice, what does it mean when we say that the production of healthcare is of 'fixed coefficient' type? Why?

.....

.....

.....

.....

.....

2) Why are 'long run average cost (LRAC) curves' also called as 'envelope curves'?

.....

.....

.....

.....

.....

3) Distinguish between 'economy of scale' and 'economy of scope' in the context of production of healthcare services.

.....

.....

.....

.....

.....

4.5 LET US SUM UP

The unit has discussed three inter-related issues of healthcare sector viz. input market in healthcare services, physician's supply of healthcare services and cost structure of a hospital supplying healthcare services. The dual role of physician is discussed where the physician can independently supply

healthcare services and also get employed by the hospitals. The behaviour of a physician as price discriminator is explained. Exploitation of nurses (in the input market of healthcare) is analysed as a special case of monopsony power of a hospital. Similarly, the supply of healthcare service as determined by costs of inputs is discussed both for its short-run and long-run cost impacts on the sector. The concepts of 'economies of scale' and 'economies of scope' is distinguished towards the end.

4.6 KEY WORDS

Value of Marginal Product (VMP) : VMP is defined as the product of 'price' and 'marginal physical productivity' (MPP) of an input.

Marginal Revenue Product (MRP) : MMP is defined as the product of 'marginal revenue' (of the product) and MPP of an input.

Isoquant : Refers to the efficiency frontier representing all possible efficient ways to achieve a certain production level.

Monopsonistic Exploitation : Refers to the gap between MRP and the wage paid. It is given by: Gap (G) = (MRP – W_m)/W_m=1/ε where ε stands for elasticity of supply of labour with respect to wage.

Lerner's Index of Degree of Monopoly Power : Lerner's Index gives an indicator of the degree of monopoly power. It is the reciprocal of the numerical coefficient of price-elasticity of demand for the product or service. The less elastic is the demand for the product or services, the more would be the degree of monopoly power and vice versa.

Marginal Rate of Technical Substitution (MRTS) of Healthcare Production : MRTS is defined as, 'keeping the total output of health desired to be produced as constant, how much of input 1 (H_1) has to be decreased if input 2 (H_2) increases by one extra unit'? It shows the relation between inputs, and the trade-offs amongst them, without changing the level of total output of healthcare services. MRTS is the negative slope of the isoquant.

Envelope Curve : The 'long-run average cost' (LRAC) curve is a lower 'envelope' of short run average cost curves (SRAC). This means that since any change in inputs in the short run will carry over into the long run as well, it is impossible for average costs to be driven down more in the short run than the long run.

4.7 SOME USEFUL BOOKS AND REFERENCES

1) Barros Pedro. and Xavier Martinez-Giralt (2012). Health Economics: An Industrial Organization Perspectives, Routledge.

2) Sloan Frank. A and Chee-Ruey Hsieh (2012). Health Economics, The MIT Press, 55 Hayward Street, Cambridge, MA, 02142.

3) Rexford E Santerre and Stephen P Neun (2010). Health Economics, Theories, Insights and Industry Studies, Mason, Ohio: South-Western Cengage Learning.

4.8 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) Many trained persons contribute to it. The chain of services needs maintained failing which services will be sub-optimal. The sector is thus non-homogeneous with each patient's response needing to be followed individually. Lack of homogeneity in the response of patients is thus the basic characteristic of the sector making it both different and complex.

2) The services of nurses are marked for a higher degree of homogeneity whereas that of physicians is the opposite i.e. it is heterogeneous. This makes it difficult for the market to experience equilibrium in case of the latter as compared to the former.

3) Due to the gap that exists between MRP and W_m as shown in Equation (4.7).

Check Your Progress 2

1) The revenue falls due to lowered price effect. The revenue increases due to more services and fee earned. The overall or gross revenue is thus influenced by two opposing forces. The economic reasoning is the downward sloping demand curve.

2) The index shows under what conditions the monopolist physician can enjoy absolute monopoly. It depends upon the price-elasticity of demand for his services.

3) Refers to physicians charging differential prices for two segments e.g. insured, non-insured; poor, non-poor.

4) The factor of 'price elasticity of demand' in the two markets. The advantages to physicians are: profit maximisation, economies of scale and positive externalities.

5) Due to free entry and exit, entry of more physicians would keep the prices to fluctuate yielding only profits in the short run to the physicians.

In the long run, prices would become sensitive and elastic responding only to factors of search cost and geographic proximity.

Check Your Progress 3

1) This means that additional inputs will not increase efficiency just as the absence of any critical input will collapse the functioning of the healthcare sector.

2) The factors that are fixed in the short run can be changed in the long run. In view of this, the average cost (AC) reductions resulting from changing inputs (in the long run) will not be felt in the short run. Thus, the SRCCs which are all U shaped (when considered for several short stretches of time periods) provides the limits to LRAC. Fig. 4.7 illustrates this.

3) Economies of scale relates to decrease in the unit cost of production with increase in the volume of production. Economies of scope relates to lowering the average cost by producing more types of products.

UNIT 5 MEASUREMENT OF HEALTH BENEFITS

Structure

5.0 OBJECTIVES

After reading this unit, you will be able to:

- outline the method of 'cost benefit analysis' (CBA) using the four quadrant cost-benefit axes zones;
- distinguish between the 'cost of illness approach' (CIA) and the 'willingness to pay (WTP) approach' (WTPA) of measuring the value of benefits of health projects;
- differentiate between 'cost effectiveness analysis' (CEA) and 'cost utility analysis' (CUA) for evaluating the outcomes of health programs;
- discuss the method of 'Impact Evaluation' (IE) for evaluating the actual benefits from health projects; and
- describe the four methods of selecting the participants for an Impact Evaluation experiment.

5.1 INTRODUCTION

There have been a number of initiatives and policy interventions from public agencies and social planners, especially in developing countries, aimed at improving the health benefits to groups of citizens. It has also therefore been of crucial importance to the agencies involved to measure the extent of health benefits that occur after a specific period of intervention. Whether the policies or programmes improved the health status in the medium and long run; whether the policies came with a rider of counter-productive health losses to a specific group; whether the programmes could justify the small benefits emerging from too large investments in terms of health improvements, etc. are some of the key issues to which answers were required. Benefits from health programmes are difficult to measure directly and hence special approaches are needed for them.

Two distinct schools of thought have emerged on measuring the health benefits to communities. The first looks at the measurement of benefits vis-à-vis the costs or expenditures incurred. Rice (1967) was one of the first one to use cost-benefit analysis in the health sector by estimating the monetary value of health improvements on the basis of 'cost of illness' averted. This approach was based on human capital theory of Becker (1964) which identified the value of healthy life with labour earnings or wages. While this approach moved from cost-benefit analysis to cost-effectiveness analysis in due course, they primarily centred around identifying the cost of the programme in money terms and the health benefits accruing from it also in money terms. In this connection, various methods of calculation of measurements of willingness to pay were developed, either based on contingent valuation approach or by estimating the demand curves to measure the consumer surplus. The other school of thought looked at the programme implementation and the health benefits accruing to the target beneficiaries. They focused on impact evaluation which aimed to quantify the benefits for the people that received the programme in comparison to those who did not receive it. This did not aim to identify the costs incurred by implementer agency because it was assumed that the expenditures on healthcare of citizens is primarily a states' responsibility. Thus, what is more important is the actual impact on the treated beneficiary vis-à-vis the non-treated control persons. Thus, the impact analysis school centred itself around examining the relative impacts between the treated and the non-treated groups for assessing the benefits of health programmes. The cost-benefit analysis and its extended family of 'cost effectiveness analysis and cost utility analysis' considered primarily the accounting concerns from the supplier side whereas the impact evaluation studies gave more importance to the self-selection into the programme by the beneficiaries. In short, while the former is of more interest to managerial economists, the latter is of significance to the welfare or development economists.

5.2 COST BENEFIT ANALYSIS

The basic emphasis of cost benefit analysis is to find the monetary value of inputs used to measure the health production and then compare with a corresponding measurement of the value of its benefits. Services of doctors, nurses, paramedical staff; drugs, vaccines, etc. are considered as inputs to produce health benefits. For instance, in a programme to introduce new drug for ailing persons, the cost of the drug measures the cost component while the decreased DALY losses are used to measure the health benefits. However, in all programmes no such direct link can be identified. For instance, the cost of prevention services used by the high risk population groups to reduce the incidence of HIV AIDS disease might result in fall in incidence risks among those groups. Again, under the programme to immunise all children below one year of age, the costs of vaccines, services of health workers and information-education campaign (IEC) are used as inputs. The benefit of children in terms of 'increased life years' is compared with the input costs. But such improvements might also come from increased general awareness from higher education of the population concerned and hence it is difficult to attribute the impact due to more usage of inputs alone.

An alternative approach is to quantify, but not value, the health outcomes of programme interventions. The quantities are used to construct a measure of cost-effectiveness approach (CEA) which is then use as a useful tool for choosing between alternative programmes. This can be extended to cost-utility analysis (CUA) which allows comparison of programmes with qualitatively different health outcomes. For instance, for analysing the efficacy of different anti-cancer drugs in increasing the health benefits, both CEA and CUA can be used. CEA measures the incremental cost of achieving an incremental health benefit in which varies cost is measured in monetary terms and effectiveness is measured in terms of a clinical outcome like number of lives saved, complications prevented or diseases cured, etc. On the other hand, CUA expresses the value for money in terms of a single type of health outcome, usually expressed as the incremental cost to gain an extra quality-adjusted life-year (QALY). This approach incorporates both increases in survival time (extra life-years) and changes in quality of life (with or without increased survival) into one measure. Both the CEA and the CUA are, therefore, techniques which focus on health outcomes measured in different units. Fig. 5.1 shows the important zones in a four quadrant diagram wherein, in the first two axes, we measure incremental costs and incremental benefits. Zone I shows the increase in both benefits and costs due to a programme, while in Zone IV, costs increase but benefits fall. Likewise, in Zone II incremental costs are negative, but incremental benefits are positive. Thus, it is a zone to be most preferred. In Zone III, both costs and benefits fall. Thus, for a social planner, it might not be interesting to operate in Zone III, because of his ultimate aim being to increase health benefits, but since it decreases total costs, a private agency might take some

interest. The zones that are crucial are Zone I and III. Here we draw a 45 degree line to divide the zones into two halves. In IA, incremental costs are higher compared to incremental benefits and the opposite is true for IB. Thus, a manager would like to put the intervention in zone IB, where benefits increase more than the costs. Zone II is always welcome where costs reduce while benefits increase. In Zone IIIA, the fall in benefit is more profound than fall in costs and hence would be avoided. If a reduction in health benefit is acceptable, then Zone IIIB would be preferred. Zone IV would be completely avoided.

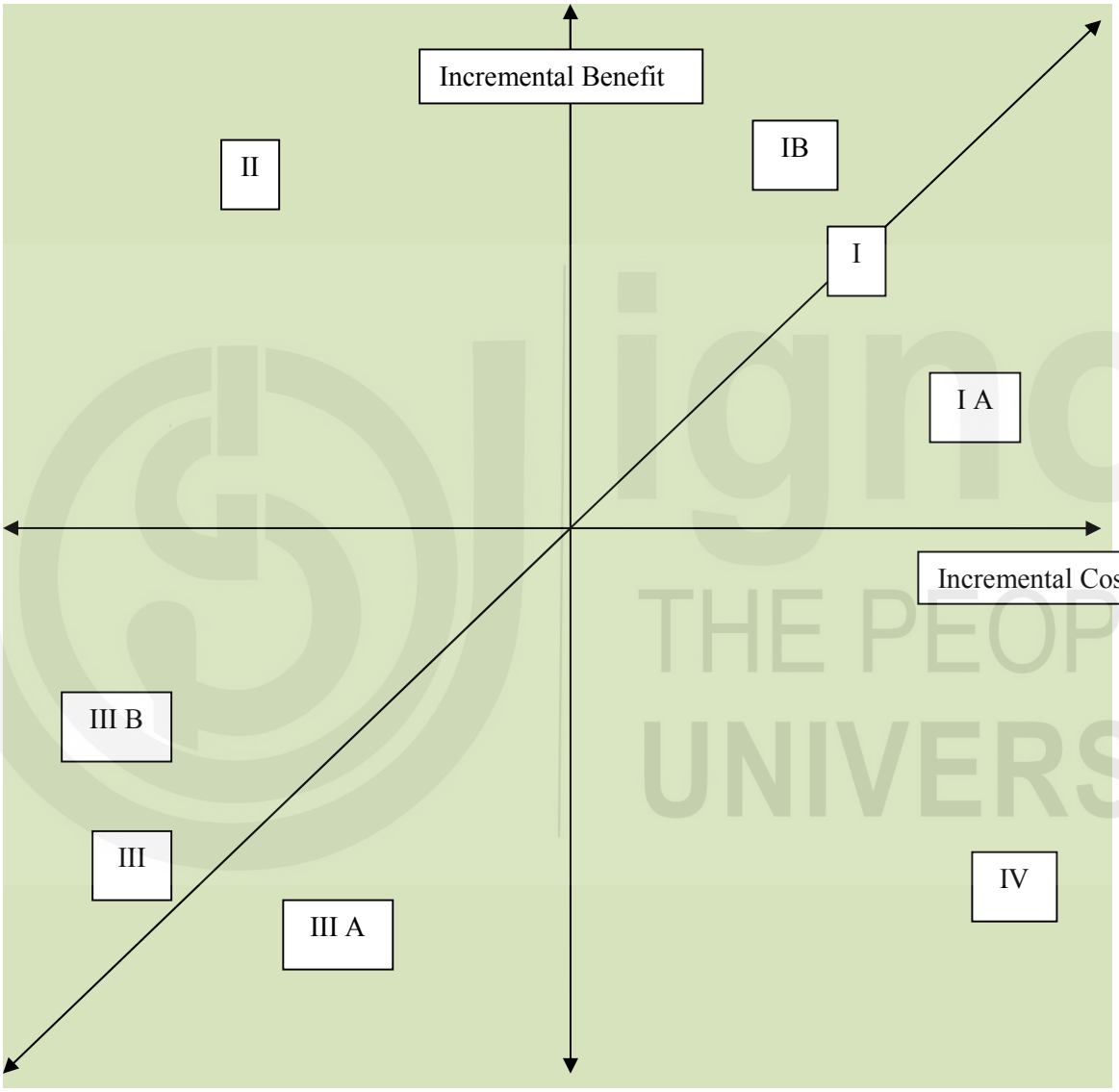

Fig. 5.1: Cost-Benefit Axes Zones

The most difficult part of this approach is how to quantify the benefits in value terms. Important health outcomes that need to be addressed are life expectancy, morbidity, disability, quality-adjusted life, future medical costs, etc. all of which are challenging to estimate. Also, there are often non-health benefits along with health benefits from a programme. For instance, with environment quality, valuation of property could increase. In practice, both the valuation of costs and benefits are difficult. Valuation of costs in health projects are relatively more difficult to capture (compared to other

programmes) as besides the non-monetary costs (like travel time, waiting time, etc. i.e. the opportunity cost) of availing the healthcare, another type of cost that needs to be considered in case of health programme analysis is the 'future medical care costs'. For instance, while a programme might reduce the immediate chance of death from a particular disease, such postponing of death would invariably be associated with continuous and sustained investment in medical care due to the frailty effect on the patient. In such cases, the costs of future medical care should be ideally treated as a cost of the programme. Valuation of benefits can be done in two approaches viz. cost of illness approach (CIA) and the willingness to pay approach (WTPA). We shall now turn to briefly outline these two methods.

5.2.1 Cost of Illness

Under this approach, the cost of continuing the existing treatment and the value of foregone production due to illness are estimated. Only the direct treatment of costs of the disease is considered although the indirect costs could often far outweigh the direct costs. For instance, in case of a death of a youth by suicide, although the immediate direct cost is zero, the death would impose a huge loss of earning to the family besides the cost of trauma care for the members of the family left behind. On the other hand, for death by heart attack at a matured age, the preceding direct medical care costs are normally higher, while the indirect costs after death is low. Another problem that arises with the CIA is that, when the health benefits are not produced within a framework of market production, the method does not yield estimates of the expected loss or gain from the health loss or benefits. This is typically the case in developing countries where home care and care received in non-formal institutions constitute a large part of treatment. Thus, the CIA is often of little value in estimating the social loss to the economy on account of health.

5.2.2 Willingness to Pay

The willingness to pay approach for a health programme is another method to calculate the health benefits. With the option of either becoming members or not for the programme available to choose from, determining the individuals' willingness to opt for the programme is the basis of this approach. To assess this, the approach proceeds to determine the willingness of potential subscribers to pay for a health programme. The approach is as follows. Starting from an initial amount of payment, the surveyor gradually increases the amount (Δm_i) to ascertain the highest amount that the respondent is willing to pay. The enquiry continues in graduated scaled amounts till the person shows indifference or unwillingness, at which point it is taken as the maximum amount of money that the individual is willing to pay (Δm_i^*). The WTPA, thus, aims at determining this highest amount on which the health programme (or the health insurance programme) is then designed. One of the standard ways to estimate the WTP is to conduct surveys of individuals by

'contingent valuation method' using the existing market estimates of prices for health services. The willingness to pay is directly asked to be quoted after revealing that the respondents are not actually required to pay. This approach is therefore likely to lead to quoting escalated figures as WTP. On the other hand, if the respondents are actually required to pay at least a part of their own valuation, they would tend to under-report. Even if they report honestly, they might change their mind when it comes to actual payment. The other way is to measure the willingness to pay by estimating the underlying demand curves. This approach involves many econometric issues associated with the price-quantity matches and hence is prone to result in biased estimates.

5.2.3 Cost Effective Analysis (CEA) and Cost Utility Analysis (CUA)

When project outcomes are difficult to quantify, alternative methods of evaluation are needed. One such method is the Cost Effectiveness Analysis (CEA) which is used when we need to choose between alternative production techniques that result in some quantitative and some qualitative output. For instance, when a CEA is done for administering a new medicine, the effectiveness can be measured on reduced loss of working days which is a quantitative outcome or the changes in the perception of health status of the patient can be measured in terms of perceived health status (i.e. how one feels) which is a qualitative outcome. Cost Utility Analysis (CUA) is another method which is used when the qualitative nature of the output differs between projects (e.g. one programme increases immunisation and the other reduces incidence of committing suicide).

The CEA involves assessing the health benefits (effectiveness) and resource requirements (costs) for alternative ways of achieving a specific objective, usually expressed in terms of costs per unit of effectiveness. The alternative programme with the least cost-effectiveness (C-E) ratio is then chosen for implementation. This way, the choice of technology, choice of targets, etc. can be compared across an array of alternatives. For instance, suppose that the policy maker needs to choose from two alternatives: whether to distribute more condoms or to improve awareness campaign for reducing AIDS incidence. The C-E ratio for them would be costs per infection prevented and that alternative would be chosen for which the ratio of C-E is least. The CEA is criticised on the ground that effectiveness for all projects may not be similar. For instance, for awareness campaign against smoking, the effectiveness might be reduced incidence of lung diseases, but along with it one needs to consider the irritability and unpleasantness that sudden withdrawal from smoking can cause. Normally, there are five steps that are required to be performed for cost-effectiveness analysis. These are: (i) defining the programme's objectives; (ii) identifying the alternative ways to achieve these objectives; (iii) identifying and measuring the costs of each alternative; (iv) identifying and measuring the effectiveness of each

alternative; and (v) calculating the cost-effectiveness ratio of each alternative and then choose the one with the least CEA.

The Cost Utility Analysis, on the other hand, is applied when the outcome is measured in a manner that allows comparisons of qualitatively and distributionally different outputs. For instance, it is used when the analysis is made for years of life saved or adjusted for quality i.e. per QALY (Quality Adjusted Life Years). The C-U ratio is thus in the form of money per QALY. Thus, being an average cost concept, it suffers from the difference in programme size. Too large programmes can have an economy of scale reflecting a lower C-U ratio, with no significant difference in utility. However, CUA has a possibility of including the distributional issues as QALY across age cohorts, gender and region would usually be available. Thus, comparing QALY gains between two dissimilar groups (e.g. an informal sector worker and a rich businessman) could be avoided. CUA enables the calculation of such distributional aspects and hence is preferred at times.

We must remember here that each evaluation is based on some assumptions on the values for some variables. It is therefore required to check the robustness of the results, by altering the assumptions to see if the results within the permissible limits. This is called the sensitivity test.

Check Your Progress 1 [answer within the space given in about 50-100 words]

1) Distinguish between the CBA and IA.

.....

.....

.....

.....

.....

2) Differentiate between the CEA and CUA.

.....

.....

.....

.....

.....

3) What is the thrust of CIA and what are its limitations?

.....

.....

.....

4) What is the essence of the WTPA?

.....

.....

.....

.....

.....

5) What are the limitations of WTPA?

.....

.....

.....

.....

.....

6) Give an example where CUA is used? State its advantages and limitation.

.....

.....

.....

.....

.....

5.3 IMPACT EVALUATION

The cost-benefit (C-B) type of analysis gives an idea on the possible benefits accruing from the project, but not the actual benefits. The de-facto impact would depend upon the fact that the beneficiaries themselves choose to be a part of the programme for health benefits. The CBA groups of studies assume that once a programme is floated, the beneficiaries would automatically choose it. But this is not true especially in developing countries owing to different social taboos and gendered equation within a household.

Government programmes, both in developed and developing countries, target specific beneficiaries for a specific outcome (e.g. increased income, improved enrolment, better health outcome, etc.) in spite of market imperfections. They are specifically crucial in markets for goods that are in the nature of either public good or merit good (e.g. infrastructure, health, education, gender equality). Whether or not these changes are actually achieved is a crucial question but one that is not *often* examined. More commonly, monitoring agencies and policy makers focus on measuring the immediate outputs of a programme (e.g. how many bicycles were distributed, how much money was spent, etc i.e. the C-B ratio). In contrast, an impact

evaluation assesses the changes in the well-being of the targeted individuals that can be attributed to a particular programme. The focus is thus on the ***impact*** i.e. the changes ***directly attributable*** to a programme. In other words, the central challenge of an effective impact evaluation is to identify the causal relationship between the programme and the outcomes. Thus, examples of the typical questions raised on 'impact evaluation' (IE) are like: (i) has the Grameen Bank succeeded in lowering consumption poverty among the rural poor in Bangladesh? (ii) can the conditional cash-transfer programmes in Mexico and other Latin American countries improve health and schooling outcomes for poor women and children? (iii) has the National Rural Health Mission in India improved institutional delivery of babies?

5.3.1 Construction of Comparison Groups

In a context in which funding agencies and civil society are demanding results and accountability from public programmes, impact evaluation can provide credible evidence on performance and, crucially, on whether a particular programme has achieved its targeted outcomes. Additionally, Impact evaluations are also increasingly being used to test innovations in programme design or service delivery. A well-designed representative Impact Evaluation (IE) can provide convincing and comprehensive evidence useful in making informed policy decisions, shape public opinion and improve programme efficiency. This involves counterfactual analysis i.e. a comparison between what actually happened and what would have happened in the absence of the intervention. Impact evaluations seek to answer cause-and-effect questions i.e. they look for the changes in outcome that are directly attributable to a programme. The counterfactual is actually not observed since one individual cannot be both within and outside a policy simultaneously and hence it has to be created statistically. There are two common methods of constructing comparison groups viz. before-and-after comparison and with-and-without programme comparisons which are discussed below.

5.3.1.1 Before and After Comparisons

Also known as pre-post comparisons, they compare the outcomes of the same group before and after participating in a program. A before-and-after comparison attempts to establish the impact of a programme by tracking changes in outcomes of programme participants over time. For instance, for the programme of offering incentives in the form of *Janani Suraksha Yojna* (JSY) to increase institutional delivery in India, the pre and post comparison for the share of institutional deliveries in total delivery might be misleading because in the meantime certain other changes might have occurred outside the particular programme which could have contributed to increase the institutional deliveries (e.g. improvement in road connectivity across the country). Thus, the entire change might not be only on account of JSY.

5.3.1.2 Self Selected Comparisons

Also known as enrolled-and-non-enrolled comparisons, self-selected comparisons compare the outcomes of a group that choose to participate in a programme with those of a group that chooses not to participate i.e. a comparison group that *self-selects* out of a programme will provide a counterfactual estimate. Selection occurs when programme participation is based on the preferences, decisions or unobserved characteristics of potential participants. For the earlier example on JSY, in spite of JSY, the most vulnerable women from rural background, belonging to religious and ethnic backward classes, with low education might not self-select themselves into the programme of delivering child in institutions. Thus, the difference between the participants and the non-participants of the programme might not entirely be due to the programme effect, but also due to the differences in socio-economic characteristics.

The biggest problem of IE is the 'treatment assignment' of the programme which must be random. However, this is *often not random* because of: (a) ***Purposive*** programme placement; and (b) ***Self-selection*** into the programme i.e. programs are placed according to the need of the communities and individuals, who in turn self-select, given the programme design and placement. Self-selection could be based on observed characteristics, unobserved factors, or both. If the impact of a particular programme is assessed by comparing the observed difference in the status of participants vis-à-vis that of non-participants, then the evaluation would be incomplete and biased. This is because assessing the impact of any intervention requires making an inference about the outcome that would have been observed for treated had they not been treated. Since it is not possible for the same person to be a participant and a non-participant simultaneously, it is not possible to observe the *variable of with and without programme* for same individual. In practice, impact evaluation requires that the evaluation team find a comparison group to estimate what would have happened to the programme participants without the programme and then make comparisons with the treatment group that has received the programme. Thus, the challenge of an impact assessment is to create a convincing and reasonable comparison group for beneficiaries in the light of this missing data. Ideally, one would like to compare how the same household or individual would have fared with and without an intervention or 'treatment'. But one cannot do so because at a given point of time a household or an individual cannot have two simultaneous existences i.e. a household or an individual cannot be in the treated and the control groups simultaneously. Finding an appropriate counterfactual therefore constitutes the main challenge of an impact evaluation. Formally, therefore, the problem of observing the outcome variable Y, across treated and non-treated individuals 'i' can be stated as:

$$Yi = \alpha Xi + \beta Ti + \varepsilon i \qquad (5.1)$$

where, *X* is a set of observed characteristics of the individual and T is a dummy equal to 1 for those who participate and 0 for those who do not participate and ε is an error term reflecting unobserved characteristics that also affect *Y*. The problem with Equation (5.1) is that T_i is not independent of X_i.

Let D = 1 for the group of individuals who participated in the programme (i.e. the treated group) and D = 0 for the group of individuals who did not participate in the programme (the control group). We denote by Y_1 the outcome conditional on treatment and Y_0 the outcome conditional on non-treatment. The problem then reduces to finding out the **average treatment effect** i.e.

$$ATE = E((Y_1 - Y_0) | D = 1) \quad (5.2)$$

Equation (5.2) can be further decomposed as:

$$ATE = E(Y_1|D=1) - E(Y_0|D=1)$$

A B

While expression A can be estimated from the sample observations on the treated group, expression B cannot be observed and hence one needs to define a suitable proxy measure. In other words, the need for the proxy leads to a typical non-random sample problem where a part of the population will have no chance of being included in the programme rendering them not comparable. This is the 'truncation bias', a special kind of selection bias, to avoid which more emphasis is given to estimate the 'Treatment Effect on the Treated' (ATT) with:

$$ATT = E[Y_i(1) - Y_i(0)|W_i = 1]$$

where $W_i = 1$ isolates those who are likely to participate in the programme. This means the average value of the outcome variable will be compared across two groups: (a) one containing those who are likely to participate and participated; and (b) the other containing those who are also likely to participate but did not participate.

5.3.2 Methods of Impact Evaluation

We discuss four types of Impact Evaluation methods viz. (i) randomised evaluations, (ii) propensity score matching (PSM) methods, (iii) double difference (DD) methods and (iv) instrumental variable (IV) methods.

5.3.2.1 Randomised Evaluations

These involve a randomly allocated initiative across a sample of subjects (e.g. communities or individuals) where the progress of treatment and control subjects exhibit similar pre-programme characteristics who are then tracked over time. Randomised experiments have the advantage of avoiding selection bias at the level of randomisation. Randomised assignment of treatments is

considered the golden standard of impact evaluation as it uses a random process to decide who is granted access to the programme and who is not. Under the randomised assignment, every eligible unit (e.g. an individual, household, business, school, hospital or community) has the same probability of being selected for treatment by a programme.

5.3.2.2 Propensity Score Matching (PSM) Methods

In the absence of an experiment, PSM methods compare treatment effects across the participant and the matched non-participant units, with the matching conducted on a range of observed characteristics.

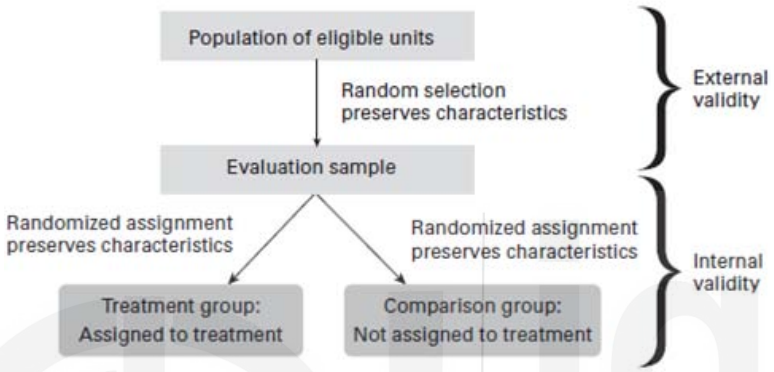

Fig. 5.2: Random Sampling and Random Assignment of Treatment

Source: World Bank, 2016

PSM methods assume that the selection bias is based only on observed characteristics and that they cannot account for unobserved factors affecting participation. Finding a good match for each programme participant requires approximating the characteristics that explain the individual's decision to enrol in the programme. However, if the list of observed characteristics is very large i.e. each characteristic takes on many values, it would be hard to identify a match for each of the units in the treatment group. This is when the method of *propensity score matching* (Rosenbaum and Rubin, 1983) becomes helpful. In this approach, we no longer need to try to match each enrolled unit with a non-enrolled unit, instead, for each unit in the treatment group and in the non-enrolled group, we compute the *probability* that the unit will enrol in the programme (i.e. the propensity score*)* based on the observed values of the explanatory variables. This score is a real number between 0 and 1 that summarises the influence of all the observed characteristics on the likelihood of enrolling in the programme. In short, we first calculate the propensity of each individual to receive the treatment in the sample and then club one treated individual with one individual from the control group. The average difference of the outcome variable is then taken as the average treatment effect on the treated. The main feature of the technique is that it removes the selection bias by balancing the sample on the characteristics that potentially determine the possibility of selection of cases into the treatment group.

5.3.2.3 Double Difference (DD) Methods

DD methods assume that unobserved selection is present and that it is time invariant i.e. the treatment effect is determined by taking the difference in outcomes across treatment and control units like in a before and after the programme intervention. DD methods can be used in both experimental and non-experimental settings. The *difference-in-differences method* compares the *changes* in outcomes over time between a population that is enrolled in a programme (the treatment group) and a population that is not (the comparison group). The difference in the before-and-after outcomes for the enrolled group (i.e. the first difference) controls for factors that are constant over time in that group, since we are comparing the same group with itself. But we are still left with the factors that vary over time (*time-varying factors*) for this group. One way to capture these time-varying factors is to measure the before-and-after change in outcomes for a group that did not enrol in the programme but was exposed to the same set of environmental conditions. This difference from the earlier first difference is *the second difference*. The difference-in-differences approach combines the two counterfeit estimates (of the before-and-after comparisons and comparisons between those who choose to enrol and those who choose not to enrol) to produce a better estimate of the counterfactual. This method actually culls out the impact created by a policy change, taking out the development-initiated automatic changes in the outcome. For instance, suppose the impact of *Sarva Siksha Abhiyan* (SSA) on child enrolment through setting up new schools is to be measured. Apart from the effect of this programme, we should not forget to consider that there is always an increase in child enrolment in schools due to other factors like the very process of development, improvement in awareness among parents, etc. The impact due to these factors, therefore, need to be deducted from the impact of SSA. Thus, one needs to first of all calculate the difference in enrolment in those sub-regions where more new schools have come up from the corresponding estimates in those sub-regions where less new schools have come up, both before and after SSA. Then one needs to take the second difference of those two first differences to get the actual impact of SSA.

5.3.2.4 Instrumental Variable (IV) Methods

IV methods are used with cross-section or panel data. In the latter, it allows for selection bias on unobserved characteristics to vary with time. In the IV approach, selection bias on unobserved characteristics is corrected by finding a variable (called instrument) that is correlated with participation but not correlated with unobserved characteristics affecting the outcome. This instrument is used to predict participation.

Check Your Progress 2 [answer within the space given in about 50-100 words]

1) What is an Impact Evaluation? How is it different from a CBA?

.....

.....

.....

.....

.....

2) What is meant by 'truncation bias' in IAs? How are they avoided?

.....

.....

.....

.....

.....

3) What is a special feature of 'randomised evaluations'?

.....

.....

.....

.....

.....

4) What is the essence of the 'double difference' method? How is it useful?

.....

.....

.....

.....

.....

5.4 LET US SUM UP

The unit discusses different methods of measuring health benefits of a particular programme implementation. Two specific groups of methods viz. the family of cost-benefit analysis and the family of impact evaluation are discussed. Cost analysis based on impact evaluation evidence exists not only for a particular programme but also for a series of programmes or programmes. Such analysis assists the makers to assess which programme or alternative is most cost effective in reaching a particular goal. Cost benefit analysis is a relative concept, while impact evaluation is not necessarily so.

It is important that impact evaluation is complemented with information on the cost of the project, programme, or policy being evaluated. Once impact and cost information are available for a variety of programmes, cost-effectiveness analysis can identify which investments yield the highest rate of return allowing policy makers to make informed decisions on which intervention to invest in.

5.5 KEY WORDS

Cost Benefit Analysis : Based on the human capital theory of Becker (1964), it aims at estimating the monetary value of health improvements on the basis of cost of illness averted.

Cost Effectiveness Analysis (CEA) : This is an alternative approach in which instead of valuing the method aims at quantifying the health outcomes of programme interventions.

Cost Utility Analysis : This is an extension of CEA which permits comparison of programmes with qualitatively different health outcomes.

Cost of Illness Approach : Under this, both the cost of continuing the existing treatment and the value of the foregone earnings due to illness are taken into account.

Impact Analysis : Unlike the CBA based approaches which assume that the beneficiaries opt for a health programme announced, IA aims at providing evidence on performance i.e. whether a particular programme has achieved its targeted outcomes. They are thus useful to test innovations in programme design or service delivery.

Counter-factual : Epistemologically, counterfactual refers to a conditional statement, the first clause of which expresses something contrary to the fact (e.g. 'if I had known'). In impact evaluation, the unit in one state, if would have belonged to the other state, is a counterfactual. This one unit cannot be observed simultaneously in two different states (in other words, with and without the programme). It remains unobserved, though it can be constructed statistically.

5.6 SOME USEFUL BOOKS AND REFERENCES

1) Creese, A and Parker, D (1994). *Cost Analysis in Primary Health Care: A Training Manual for Programme Managers*, WHO Publications Centre, NY 12210.

2) World Bank & IDB (2016). Impact Evaluation in Practice.

3) Jack, W (1999). *Principles of Health Economics for Developing Countries,* World Bank Publications.

5.7 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) CBA focuses on the inputs used, measured in money terms, in a health program. Based on the measure of total cost incurred, the CBA compares this with a measure, again measured in money terms, of benefits. IA, on the other hand, assumes the cost of health delivery as a responsibility of the State and focuses on a measure of the 'health outcome' i.e. it focuses on the target beneficiaries. The IA compares the benefits of health accrued to the 'treatment group' with that of the non-treated group (i.e. the controlled group).

2) The CEA quantifies the health outcomes of programme interventions. These are then used to construct a measure of CEA which can then be used as a tool for choosing between alternative programmes. CUA, on the other hand, makes comparison of programmes with qualitatively different outcomes.

3) The cost of illness approach (CIA) takes into account both the cost of treatment plus the loss of income. However, under costs, only the direct costs are taken into account. Further, the approach requires the health benefits to be produced within a market framework i.e. it cannot accommodate home care or care received in non-formal institutions. Thus, the CIA is not useful in estimating the social loss to the economy.

4) This is a survey based approach where the members to the health programme are voluntary subscribers. The approach determine the highest amount that a group of individuals are prepared to pay and then proceeds to design the programme based on this input.

5) The method results in over-reporting when the respondents are told that they are not required to pay any amount but have to merely indicate their willingness to pay. It tends to under-report, when the respondents are revealed that they are required to pay at least a part of the payment.

6) CUA is used when the outcome measured requires a comparison to be made for qualitative and distributionally different outputs (e.g. money per QALY). Thus, being a 'average cost concept', it suffers from the difference in programme size. But the merit of the approach is that it allows its computation for seemingly different cohorts as the CU ratio would generally be available by gender, age, region, etc.

Check Your Progress 2

1) Impact evaluation assesses the changes in the well-being of the targeted individuals that can be attributed to a particular programme. The focus is on the *impact* i.e. the changes *directly attributable* to a programme. CBA, on the other hand, aims at measuring the immediate outputs of a programme relating it mainly to the costs incurred in the programme. Impact evaluation thus aims at identifying the causal relationship between the programme and the outcomes.

2) The 'average treatment effect' [Equation (5.2)] is decomposed into two parts, the first one of which can be estimated but the second one of which cannot be estimated. For the latter, it therefore becomes necessary to choose a proxy. This introduces a special kind of bias which is termed as the 'truncation bias'. This is avoided by estimating the 'treatment effect on the treated'. The PSM method discussed is also a method to deal with this.

3) That it accords equal probability of being selected for each and every one of the unit of selection for the treatment by a programme.

4) The difference in the before-and-after outcomes for the enrolled group is the first difference. Variations due to time-varying factors or environmental conditions also needs to be captured. This is done by measuring the difference in the before-and-after change in outcomes for a group that **did not enrol in the programme** but was exposed to the same set of environmental conditions. This difference from the earlier first difference is then *the second difference*. It is useful by way of producing a better estimate of the counterfactual.