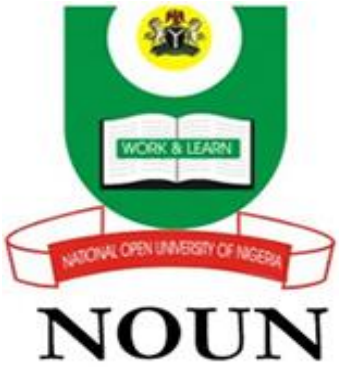


NATIONAL OPEN UNIVERSITY OF NIGERIA

HEALTH ECONOMICS

ECO 725

FACULTY OF SOCIAL SCIENCES

COURSE GUIDE

Course Developer:
Dr. Saheed O. OLAYIWOLA
Department of Economics
Federal University of Technology, Akure

Course Editor:
Professor Ishmael OGBORU
Department of Economics
University of Jos
Plateau State.

CONTENT

Introduction

Welcome to ECO 725: HEALTH ECONOMICS

ECO 725: Health Economics is a two-credit and one-semester course. The course is made up of twelve units spread across fifteen lectures weeks. This course guide provides details of issues involved in health economics and provides necessary information for those who are new to the course. It gives information about the course materials and how to work through it. It also suggests general guidelines on the time required to achieve each unit aims and objectives. Answers to tutor marked assignments (TMAs) are provided within the contents of the material.

Course Content

Health Economics is a fascinating subject. This course is meant to give both economics and non-economics student the basic economics principles and their application to the health sector. Hence, this course material should be regarded as an introduction to health economics rather than to economics. The overall purpose of the course is to introduce the basic concepts of economics and their application to the health sector and not to fully present all that is important about the subject matter of health economics. Thus, the need for supplementary reference books could be of paramount importance. The concepts and the analyses presented in this course material will help to serve as working material so that students and others could understand and apply basic ideas of economics to the health sector. The topics covered include special features of the healthcare market, the four basic questions, the main features of the health care service and its relation with economic development, economic models and analysis, the market for health insurance, the issue of equity as it relates to heath and heath care, the technique of health economic evaluation and health economics and sustainable development.

Course Aims

The aims of this course are to give you in-depth understanding of health economics as regards:

- the meaning and purpose of health economics,
- the basic instruments of economic analysis of the health sector,
- health as one of the social sectors with economic implication,
- the specific nature of the health care service,

- the importance of economics to resource allocation, planning and management of the health sector
- the implications of economic development to the health care services and
- economics of health care financing

Course Objectives

To achieve the aims of this course, there are overall and set out objectives which the course is set to achieve for each unit. The unit objectives are included at the beginning of a unit; you should read them before working through the unit. You may want to refer to them during your study of the unit to evaluate your progress. You should always look at the unit objectives after completing a unit. This is to assist the students in accomplishing the tasks entailed in this course. In this way, you can be sure you have done what was required of you by the unit. The objectives serve as study guides, such that a student could know if he is able to comprehend the knowledge of each unit through the sets of objectives in each one. At the end of the course period, the students are expected to:

- understand health as one of the social service sectors with economic implication.
- understand the specific nature of the health care service in implementing economic principles and techniques.
- know the implications of economic development to the health services.
- understand the effect of some economic factors on the health status of the society.
- identify the ways through which improvement of the health system can create conducive conditions for sustainable economic development and vice versa.
- be introduced to the possibilities of using cost benefit analysis and cost effectiveness analysis in assessing the performance of health care activities.
- outline the methods needed for costing in an economic evaluation and to give examples of costing methods and cost data types.
- identify the factors that influences the choice of a financing system.
- explore the different sources of financing the health service sector.

- understand the strong and weak points of different financing mechanisms.
- understand the role of government as affecting the resource allocation pattern in health and the extent to which it can influence the overall performance of the sector.

Working through the Course
To successfully complete this course, you are required to read the study units, referenced books and other materials on the course.

Each unit contains self-assessment exercises called Student Assessment Exercises (SAE). At some points in the course, you will be required to submit assignments for assessment purposes. At the end of the course there is a final examination. This course should take about 15weeks to complete and some components of the course are outlined under the course material subsection.

Course Material
The major component of the course, what you have to do and how you should allocate your time to each unit in order to complete the course successfully on time are listed as follows:

1. Course Guide
2. Study Unit
3. Textbook
4. Assignment File
5. Presentation Schedule

Study Unit
There are 12 units in this course which should be studied cautiously and meticulously.

MODULE ONE: INTRODUCTION TO HEALTH ECONOMICS
Unit 1: The Economics of Healthcare
Unit 2: The Four Basic Questions
Unit 3: The Main Features of Health Care Service and Its Relation with Economic Development
Unit 4: Health and Medical Care: An Economic Perspective

MODULE TWO: THE DEMAND AND SUPPLY OF MEDICAL CARE
Unit 1: The Demand for Medical Care
Unit 2: The Supply of Physician Services and other Medical Services
Unit 3: Medical Care Production and Costs
Unit 4: Hospital Services and Efficiency

MODULE THREE: COST CONCEPTS, ECONOMIC EVALUATION AND HEALTH FINANCING

Unit 1: Cost Concepts and Economic Evaluation
Unit 2: Health Care Financing
Unit 3: The Role of Government in Health Care

Each study unit will take at least two hours, and it include the introduction, objective, main content, self-assessment exercise, conclusion, summary and references. Other areas concern the Tutor-Marked Assessment (TMA) questions. Some of the self-assessment exercise will necessitate discussion, brainstorming sessions and arguments with some of your colleagues.

There are also textbooks under the reference and other (on-line and off-line) resources for further reading. They are meant to give you additional information. You are required to study the materials; practice the self-assessment exercise and tutor-marked assignment (TMA) questions for in-depth understanding of the course.

Textbooks and References

For further reading and more detailed information about the course, the following materials are recommended:

Culyer J.A. & J.P. Newhouse (2000) Eds, Handbook of Health Economics: Vols 1A & 1B, Elsivier, North-Holland.

Donaldson Cam and Karen Gerard (1993) Economics of Health Care Financing: The Visible Hand. Macmillan Press Ltd. London.

Folland S., A. Goodman & M. Stano (2010) The Economics of Health & Health Care, Sixth Edition, Prentice Hall, New Jersey.

Jacobs, P. (1991) The Economics of Health and Medical Care Maryland: Aspen Pub Inc.

Jack, Williams (1964) Principles of Health Economics for Developing Countries. WBI Development Studies. The World Bank, Washington D. C.

Jones Andrew (2007) Applied Econometrics for Health Economists: A Practical Guide, 2nd Edition OHE

Phelps Charles E. (1992) Health Economics, New York: Harper Collins Pub Inc.

Santerre E. & S.P. Neun (1996) Health Economics: Theories, Insights & Industry Studies, Irwin, Chicago.

Zweifel P., F. Breyer & M. Kifmann (2009) Health Economics, Second Edition, Springer Verlag Heidelberg.

Assignment File

Assignment files and marking scheme will be made available to you. This file gives you the details of the work you must submit to your tutor for marking. The marks you obtain from these assignments shall form part of your final mark for this course. Information on assignments will be found in the assignment file and later in this course guide in the section on assessment.

There are three assignments in this course. The three course assignments will cover:

Assignment 1 - All TMAs questions in Units 1 – 4 (Module 1)

Assignment 2 - All TMAs questions in Units 5 – 8 (Module 2)

Assignment 3 - All TMAs questions in Units 9 – 12 (Module 3)

Presentation Schedule

The presentation schedule included in the course material gives you the important dates for the completion of tutor-marking assignments and attending tutorials. You are required to submit all your assignments by due date. You should guard against falling behind in your work.

Assessment

There are two types of assessment of the course. First are the tutor-marked assignments; second is a written examination.

In attempting the assignments, you are expected to apply knowledge and techniques gathered during the course. The assignments must be submitted to your tutor for formal assessment in accordance with the deadlines stated in the presentation schedule and the assignments file. The works you submit to your tutor for assessment constitute 30 % of the total course mark.

At the end of the course, you will need to sit for a final written examination of two hours duration. This examination will constitute 70% of your total course mark.

Tutor-Marked Assignments (TMAs)

There are three tutor-marked assignments to be submitted in this course. The TMAs constitute 30% of the total score. You are encouraged to work all the questions thoroughly. Assignment questions for the units in this course are contained in the Assignment File. You will be able to complete your assignments from the information and materials contained in your set books,

reading and study units. However, it is desirable that you demonstrate that you have read and researched more widely than the required minimum. You should use other references to have a broad understanding of the subject and also to give you a deeper understanding of the subject.

When you have completed each assignment, send it, together with a TMA form, to your tutor. Make sure that each assignment reaches your tutor on or before the deadline given in the Presentation File. If for any reason, you cannot complete your work on time, contact your tutor before the assignment is due to discuss the possibility of an extension. Extensions will not be granted after the due date unless there are exceptional circumstances.

Final Examination and Grading

The final examination will be of two hours duration and have a value of 70% of the total course grade. The examination will consist of questions which reflect the types of self-assessment practice exercises and tutor-marked problems you have previously encountered. All areas of the course will be assessed.

Revise the entire course material using the time between finishing the last unit in the module and that of sitting for the final examination. You might find it useful to review your self-assessment exercises, tutor-marked assignments and comment on them before the examination. The final examination covers information from all parts of the course.

Course Marking Scheme

The Table presented below indicates the total marks (100%) allocation.

Assignment	**Marks**
Assignments (Best two out of three marked assignments)	30%
Final Examination	70%
Total	**100%**

Course Overview

The Table presented below indicates the units, number of weeks and assignments to be taken to successfully complete the course, Health Economics (ECO 725).

Units	Title of Work	Week's Activities	Assessment (end of unit)
	Course Guide		
MODULE 1: INTRODUCTION TO HEALTH ECONOMICS			
1	The Economics of Healthcare	Week 1& 2	Assignment 1
2	The Four Basic Questions	Week 3 & 4	Assignment 1
3	The Main Features of Health Care Service and Its Relation with Economic Development	Week 5	Assignment 1
4	Health and Medical Care: An Economic Perspective	Week 6	Assignment 1
MODULE 2: THE DEMAND AND SUPPLY OF MEDICAL CARE			
1	The Demand for Medical Care	Week 7	Assignment 1
2	The Supply of Physician Services and other Medical Services	Week 8	Assignment 2
3	Medical Care Production and Costs	Week 9	Assignment 2
4	Hospital Services and Efficiency	Week 10	Assignment 2
MODULE 3: COST CONCEPTS, ECONOMIC EVALUATION AND HEALTH FINANCING			
1	Cost Concepts and Economic Evaluation	Week 11 & 12	Assignment 3
2	Health Care Financing	Week 13	Assignment 3
3	The Role of Government in Health Care	Week 14 &15	Assignment 3
	Total	**15 Weeks**	

How to get the most from this Course

In distance learning the study units replace the university lecturer. This is one of the great advantages of distance learning; you can read and work through specially designed study materials at your own pace and at a time and place that suit you best. Think of it as reading the lecture instead of listening to a lecturer. In the same way that a lecturer might set you some reading to do, the study units tell you when to read your books or other material, and when to embark on discussion with your colleagues. Just as a lecturer might give you an in-class exercise, your study units provide exercises for you to do at appropriate points.

Each of the study units follows a common format. The first item is an introduction to the subject matter of the unit and how a particular unit is integrated into the other units and the course as a

whole. Next, is a set of learning objectives. These objectives let you know what you should be able to do by the time you have completed the unit. You should use these objectives to guide your study. When you have finished the unit you must re-check whether you have achieved the objectives. If you make a habit of doing this you will significantly improve your chances of passing the course and getting the best grade. The main body of the unit guides you through the reading from other sources. This will either be from your books or from a readings section.

Self-assessments are interspersed throughout the units, and answers are given at the end of the units. Working through these tests will help you to achieve the objectives of the unit and prepare you for the assignments and the examination. You should do each self-assessment exercises as you come to it in the study unit. Also, ensure to master some major historical dates and events during the course of studying the material.

The following is a practical strategy for working through the course. If you run into any trouble, consult your tutor. Remember that your tutor's job is to help you. When you need help, don't hesitate to call and ask your tutor to provide it.

1. Read this Course Guide thoroughly.

2. Organize a study schedule. Refer to the `Course overview' for more details. Note the time you are expected to spend on each unit and how the assignments relate to the units. Important information, e.g. details of your tutorials, and the date of the first day of the semester is available from study centre. You need to gather together all this information in one place, such as your diary or a wall calendar. Whatever method you choose to use, you should decide on and write in your own dates for working each unit.

3. Once you have created your own study schedule, do everything you can to stick to it. The major reason that students fail is that they get behind with their course work. If you get into difficulties with your schedule, let your tutor know before it is too late for help.

4. Turn to Unit 1 and read the introduction and the objectives for the unit.

5. Assemble the study materials. Information about what you need for a unit is given in the `Overview' at the beginning of each unit. You will also need both the study unit you are working on and one of your set books on your desk at the same time.

6. Work through the unit. The content of the unit itself has been arranged to provide a sequence for you to follow. As you work through the unit you will be instructed to read sections from your set books or other articles. Use the unit to guide your reading.

7. Up-to-date course information will be continuously delivered to you at the study centre.

8. Work before the relevant due date (about 4 weeks before due dates), get the Assignment File for the next required assignment. Keep in mind that you will learn a lot by doing the assignments carefully. They have been designed to help you meet the objectives of the course and, therefore, will help you pass the exam. Submit all assignments not later than the due date.

9. Review the objectives for each study unit to confirm that you have achieved them. If you feel unsure about any of the objectives, review the study material or consult your tutor.

10. When you are confident that you have achieved a unit's objectives, you can then start on the next unit. Proceed unit by unit through the course and try to pace your study so that you keep yourself on schedule.

11. When you have submitted an assignment to your tutor for marking do not wait for its return `before starting on the next units. Keep to your schedule. When the assignment is returned, pay particular attention to your tutor's comments, both on the tutor-marked assignment form and also written on the assignment. Consult your tutor as soon as possible if you have any questions or problems.

12. After completing the last unit, review the course and prepare yourself for the final examination. Check that you have achieved the unit objectives (listed at the beginning of each unit) and the course objectives (listed in this Course Guide).

Tutors and Tutorials

There are some hours of tutorials (2-hours sessions) provided in support of this course. You will be notified of the dates, times and locations of these tutorials together with the name and phone number of your tutor, as soon as you are allocated a tutorial group.

Your tutor will evaluate and comment on your assignments, keep a close watch on your progress and on any difficulties you might encounter during the course. You must mail your tutor-marked assignments to your tutor well before the due date (at least two working days are required). They will be marked by your tutor and returned to you as soon as possible.

Do not hesitate to contact your tutor by telephone, e-mail, or discussion board if you need help. Contact your tutor if:

• you do not understand any part of the study units or the assigned readings

• you have difficulty with the self-assessment exercises; or

• you have a question or problem with an assignment, with your tutor's comments on an assignment or with the grading of an assignment.

You should try your best to attend the tutorials. This is the only chance to have a face-to-face contact with your tutor and to ask questions which are answered instantly. You can raise any problem encountered in the course of your study. To gain the maximum benefit from course tutorials, prepare a question list before attempting them. You will learn a lot from participating in discussions actively.

Summary

The course, Health Economics (ECO 725), expose you to health economics, the four basic questions in health economics**,** the main features of tshe health care service and its relation with economic development, economic models and analysis, the economic perspective of health and medical care, the demand for medical care, the supply of physician services and other medical services, medical care production and costs, cost concepts, economic evaluation, health care financing and the role of government in health.

On successful completion of the course, you would have developed critical thinking skills with the material necessary for efficient and effective discussion on issues related to health economics and economics of healthcare.

I wish you success in the course and hope that you find it interesting and convenient.

MODULE ONE: INTRODUCTION TO HEALTH ECONOMICS

Unit 1: The Economics of Healthcare
Unit 2: The Four Basic Questions
Unit 3: The Main Features of Health Care Service and its Relation with Economic Development
Unit 4: Health and Medical Care: An Economic Perspective

Unit 1: The Economics of Healthcare

CONTENTS

1.0 INTRODUCTION

Economists in all sectors are concerned with the allocation of resources between competing demands. Demands are assumed to be infinite – there is no end to consumption aspirations. Resources like labour, raw materials, production equipment and land are always finite. Thus scarcity of resources becomes the fundamental problem to which economists address themselves. In the health sector, such scarcity can be recognised in a host of questions that concern all who work there or use its services. Why has the volume of resources absorbed by the sector increased so fast over the last four decades worldwide? Why does it seem that no matter how many nurses and doctors are employed, new technologies adopted, new drug therapies introduced, that even the rich countries of the world are not able to provide the highest quality of care for all citizens? Why do economists work in health? The health sector is not the first place people associate with economists. In principle, economists are concerned with better choices and in making the best use of existing resources and growth in the availability of resources. As economists started to work on problems in the health sector, the new discipline of health

economics emerged. Many of the concerns in health economics are also those of other health scientists – how can we improve survival, quality of life and fairness in access to services? However, economics brings a different framework that offers important and useful insights for analysing such questions. Therefore, understanding the modern economy requires an appreciation of the special economics of healthcare.

This module starts by examining various views on the definition of health economics within the various schools of thoughts in economics. It also examines various economic problems in the health sector. It looks at the special features of the health sector and the four basic questions on the allocation and distribution of resources in the health sector. This module further looks at the measurement of health status and the possible future challenges of the health sector. The proper scope of government intervention in the healthcare system is a topic of continuing political debate. The basic introduction to the economics of healthcare should help you become a more informed participant in what will be an on-going national discussion for many years to come.

2.0 OBJECTIVES

At the end of this unit, you should be able:

- To know the definition of health economics
- To understand the main characteristics of the healthcare market.
- To understand various problems of the health sector

3.0 MAIN CONTENT

3.1 The Discipline of Health Economics

Health economics as an independent scientific discipline started more than seven decades ago, in the sense that the specific treatment of topics relating to the economics of the health care sector has become common. The field is now well established that it has appeared in the ordinary curriculum of most universities, and academic health economists are found in the medical departments, the connections to economics proper are being strengthened, and the methodology applied is refined. That there is a need for concern about the economic situation of the healthcare

sector no longer demand long explanations. One of the main themes of health economics– and consequently one that we will be concerned about mostly – is to find out what is obtained in terms of outcome from the quite significant outlays on healthcare, and – in a slightly more sophisticated version – to develop methods which secure as much as possible the maximum outcome by the given means. Health economics has a wider scope than the study of a particular sector of the economy; the health care sector is not just another sector (as agriculture, industry or say, financial services); its output is somewhat elusive – but it certainly goes beyond what can reasonably be measured in monetary terms, due to the fact that the final output is individual health, or to be more specific, improvements in individual health conditions. That is quantities which are not readily comparable between individuals and not measureable in monetary terms.

This special nature of the sector gives rise to many fundamental problems, which by themselves represent challenges to economic theory. The economics of health economics is by no means trivial. Many applications of health economics are of a kind, where the deeper theoretical considerations are not heavily used (even though they perhaps should have been); indeed, much of the present interest in health economics grew out of specific needs to be satisfied here and now (such as the advantage or disadvantage of using a new treatment which may be more expensive than previous treatments but may reduce other types of cost). The day-to-day nature of such considerations implies that too much theorizing should be avoided. On the other hand, much of the methodology applied in such cases depends on conventions which ideally should have a theoretical foundation, and the need for such a foundation will surface from one time to another. One of the examples of this phenomenon is the controversy between supporters of human-capital versus frictional methods in assessing the "production gain" of a treatment: If the treatment makes the patients able to work more, how should this be measured, by wages earned or by the amount saved from not having to call in an unemployed and giving her the necessary instructions (the method chosen makes a big difference in the result). At a closer look, it turns out that health economics cannot easily be defined; as we argued above, it is not just the economics of the established health care sector, which, by the way, is not a very well defined concept, since health care institutions, providers, and financing differ among countries. Also an

attempt to define the field by its output, that is "health", seems to be largely unsuccessful, perhaps due to the ambiguities in the concept of "health", which in many contexts is interpreted in a very wide sense so as to become synonymous with "welfare" or "happiness".

A well-known definition of "health" proposed by the World Health Organisation (WHO) characterizes it as a state of perfect physical, moral and social well-being. It is doubted whether this way of looking at health will be productive in an analysis of how much health we get from the health expenditure. Indeed it seems that using this definition, there is no longer a specific field of health economics, since it has become synonymous with economics as such. Fortunately there is no need for a precise definition of health economics, since our primary interest will be in the specific topics whether they happen to be typical of health economics or not. In that case, even "health" is a term which we do not need very much. Rather we shall repeatedly consider models where we try to grasp one out of the presumably very many different aspects of health. Our lack of enthusiasm in measuring health or even discussing "health" has of course, to do with the fact that we are doing economics, trying to be so precise as possible about the concepts entering the model, and also trying as far as possible to construct the models so that only concepts which make immediate sense – which "health" does only very rarely – enter as variables to be studied. In order to understand the role of economics in relation to health care we have to return to the basic structure of economic science and its function. Economics is concerned with describing the interrelations between different individuals and organizations related to production and consumption of goods and services. The main point of the study of these interrelationships is to explain how the institutional framework, the rules of behaviour specified for the individuals and organizations influence the final outcome. Classical economic disciplines as price theory and welfare theory investigate the market mechanism; industrial organization focuses at the consequences of imperfect competition for prices, welfare, and incomes. The theory of international trade investigates the workings of different rules for international commodity exchanges, etc. On this level it should come as no surprise that health economics may be viewed as the economic discipline which deals with the institutional frameworks for health care (consumption, provision, financing) and the interconnections

between rules and institutions on the one side, and the resulting health condition of the population on the other side. This still remains a somewhat loose description of the field, and it seems difficult to get closer in a few words.

3.1.1 Economic Problems in the Health Care Sector

The problems in the health care sector will allow you to get much around in economic theory. Below are some of the relevant fields:

(i) Consumer substitution is one of the topics taught in economics – commodities compete with each other for the consumer's budget, and changes in the initial conditions (prices, budget, and tastes) will produce responses in the demand for all commodities. Substitution is a fundamental phenomenon in economics, in the medical profession, the viewpoint that health should be absolutely evenly distributed in the population is very firmly rooted. Although there seems to be no similar quest for equality in incomes; that the two are interrelated comes as a big surprise. A striking example of substitution with unexpected health effects may be provided by an investigation of teenager behaviour with respect to the use of mobile phones and smoking: while the use of mobile phones has increased dramatically, smoking habits have changed so that there are fewer smokers. A possible explanation is that both types of consumption have the main goal of signalling adulthood, but once the teenagers engage in buying a mobile phone and using it, the budget no longer allows smoking which is consequently reduced. The classical model of long-term consumption and individual health behaviour by Grossman (1972a) is a story about substitution. You can invest in your own health (by choosing the right diet, workout and frequent visits to the gym), and this investment will give you a payoff in terms of less time wasted on treating and curing your illnesses, but you will have to compare with other investments, such as buying shares, which may or may not give you a better payoff.

(ii) In the health care sector both consumption and production is subject to externalities. It matters to us what other people do or perhaps do not do. First of all, there is simple externality connected with infectious diseases, where the treatment of any patient has an effect on the number of possible future cases, thus on the probability of any other person

getting the disease. But the consumption externalities go beyond this. We experience disutility from seeing that other people do not get the same treatment for illness as we do ourselves, which means that our satisfaction depends on the consumption of other people besides ourselves. This is not in itself outstanding; traffic economists deal with congestion effects: the fact that so many people use their car has a detrimental effect on the pleasure that others get out of using their car. Also, the utility of conspicuous consumption (derived from showing other people that you can afford goods which they cannot) is reduced the more people engage in it. But in the health care sector the externality is other way round and it is a factor to be considered in the design of a system of health care financing.

(iii) On the production side of the economy there is an element of natural monopolies – hospitals need a certain minimal size to function and the cost structure is characterized by the presence of large fixed costs. There are other types of monopolies which are perhaps less based on technological characteristics and more on tradition and political expediency. Pharmacies have a monopoly on sale of prescribed drugs, the medical industry produces under monopoly based on patent rights. It is easily seen that market failure must be a central theme in any discussion of the economic performance of the health care sector.

(iv) Uncertainty is an important aspect of almost all economic behaviour, but in some situations (actually, most of the situations treated by economic theory) it is acceptable to disregard it when investigating the basic patterns of behaviour. However, when dealing with problems of illness and treatment for illness, uncertainty is central to the problem. Consumption of this type of health care is consumption under uncertainty, and as such it must be considered in the proper perspective. It has been argued that the presence of user payments for treatment does not reduce demand once the need for treatment is established – a broken leg must be treated whether the treatment is cheap or not. However, this argument neglects that consumption under uncertainty should be considered as contingent consumption (depending on whether an illness occurs or not) and that there is a wide spectrum of choice available to the individual in determining the proper contingent contract (insurance). The notion of user payments cannot be understood separately from insurance

and the types of market failure pertaining to insurance contracts, related to asymmetric information in one of its several forms.

We can go on with this, showing that the diverse fields of economic theory come into play in health economics, but it's better to proceed directly to health economics proper, where we shall consider the details with these and many other problems. The goal is not only to identify the problems and their theoretical content, but also to relate to the field of regulation and control. This is in many cases quite clear, since markets for health care often do not regulate themselves; there is a need for a regulation in the interest of society. Indeed, the health care sectors are highly regulated in most countries. Control and regulation is a central aspect of the economic organization of the health care sector. When the varying degree of direct public engagement in health care provision is added, it becomes clear that it is something that matters much. We will consider merits and demerits of government engagement versus decentralized market solutions, and since our discussion will have another point of departure (namely economic theory) than its counterpart in the public debate, the conclusions may not always be the same.

SELF ASSESSMENT EXERCISE

Discuss some of the economic thinking applicable to the health sector.

3.2 The Special Characteristics of the Market for Healthcare

The standard theory of how markets work is the model of supply and demand. This model has several notable features:

(i) The main interested parties are the buyers and sellers in the market.

(ii) Buyers are good judges of what they get from sellers.

(iii) Buyers pay sellers directly for the goods and services being exchanged.

(iv) Market prices are the primary mechanism for coordinating the decisions of market participants.

(v) The invisible hand, left to its own devices, leads to an efficient allocation of resources.

For many goods and services in the economy, this model offers a good description. But none of these features of the standard model reflects what goes on in the health care market. Like other

markets, the health care market has consumers (patients) and producers (doctors, nurses, etc.). But various features of this market complicate the analysis of their interactions. In particular:

(i) Third parties — insurers, governments, and unwitting bystanders—often have an interest in health care outcomes.

(ii) Patients don't know what they need and cannot evaluate the treatment they are getting.

(iii) Health care providers are often paid not by the patients but by private or government health insurance.

(iv) The rules established by these insurers, more than market prices, determine the allocation of resources.

(v) In the light of the foregoing four points, the invisible hand can't work its magic, and so the allocation of resources in the health care market can end up inefficient.

Health care is not the only good or service in the economy that departs from the standard model of supply, demand, and the invisible hand. But health care may be the most important good or service that departs so radically from this benchmark. Examining the special features of this market is a good starting point for understanding why the government plays a large role in the provision of health care and why health policy is often complex.

(a) The Prevalence of Externalities: Market outcomes may be inefficient when there are externalities. An externality arises when a person engages in an activity that influences the well-being of a bystander but neither pays nor receives compensation for that effect. If the impact on the bystander is adverse, it is called a negative externality. If it is beneficial, it is called a positive externality. In the presence of externalities, society's interest in a market outcome extends beyond the well-being of buyers and sellers who participate in the market to include the well-being of bystanders who are affected indirectly. Because buyers and sellers neglect the external effects of their actions when deciding how much to demand or supply, the externality can render the unregulated market outcome inefficient. This general conclusion is crucial for understanding health care, because externalities in the market are so prevalent. These externalities can call for government action to remedy the market failure.

Consider vaccines, for example, if one person vaccinates herself against a disease, she is less likely to catch it. But because she is less likely to catch it, she is less likely to become a carrier and infect other people. Thus, getting vaccinated conveys a positive externality. If getting vaccinated has some cost, either in money, time, or risk of adverse side effects, too few people will choose to get themselves vaccinated because they will likely ignore the positive externalities when weighing the costs and benefits. The government may remedy this problem by subsidizing the development, manufacture, and distribution of vaccines or by requiring vaccination. Another example of an externality in the health care system concerns medical research. When a physician figures out a new way to treat an ailment, that information enters society's pool of medical knowledge. The benefit to other physicians and patients is a positive externality. Without government intervention, there will be too little research. Governments respond to this externality in many ways. Sometimes, the government grants the researcher a patent on the new product, as is the case with new pharmaceutical drugs. The patent gives an incentive for research because the patent holder can profit from a temporary monopoly. The patent is said to internalize the externality. Yet this approach is not perfect because the monopoly price is higher than the marginal cost of production. The high monopoly price reduces the consumption of the patented treatment, leading to inefficiency as measured by the dead weight loss. Moreover, the high price may be hard on lower-income patients. Another approach to dealing with the positive externality from medical research is for the government to subsidize the research. This policy requires taxation to raise the necessary funds, and taxation entails deadweight losses of its own. But if the externalities from the funded research exceed the cost of the research, including the deadweight losses, overall welfare will increase.

(b) The Difficulty of Monitoring Quality: In most markets, consumers know what they want, and after a transaction is completed, they can judge whether they are happy with what they got. Health care is different. When you get sick, you may not know what the best treatment is. You rely on the advice of a physician, who has years of specialized training. And even with hindsight, you cannot judge for yourself whether the treatment the physician offered you was the right one. Sometimes state-of-the-art medicine fails to improve a patient's

health. And given the natural restorative power of the human body, the wrong treatment can sometimes appear to work. The inability of health care consumers to monitor the quality of the product they are buying leads to various regulations. Most importantly, the government requires physicians, dentists, nurses, and other health professionals to have licenses to practice. These licenses are granted only after an individual attends an approved school and passes rigorous tests. Similarly, the food and drug administration (e.g. NAFDAC) oversees the testing and release of new pharmaceutical drugs to make sure they are safe and effective. In addition to government regulation, the medical profession monitors itself by accrediting medical schools, promoting best practices, and establishing professional norms of behaviour. A physician's advice is supposed to be based on the patient's best interest, not on the physician's personal gain. When patients accept the advice, they rely on a degree of trust, which is often fostered by long-term relationships between doctor and patient. Suspicions about the standard economic motive of self-interest and the role of trust in health care relationships may explain the prevalence of non-profit hospitals. In some ways, hospitals are like hotels, but while most hotels are for-profit businesses, most hospitals are run by the government or established as non-profit entities. When consumers cannot judge the quality of the product they are buying, they may be more willing to trust an institution that is not set up primarily to enrich its owners. These public and private regulations of health care have their critics. For example, some economists have argued that there are too many hurdles to opening new medical schools. They suggest that the medical profession acts like a monopoly. By restricting the number of doctors, it drives up doctors' salaries and consumers' health care costs. Other economists have argued that the food and drug administration is too slow in approving new drugs. Some patients who might have benefited from experimental treatments are forced to go without them.

(c) The Insurance Market and Its Imperfections: Spending on health care is unpredictable because people don't know when they are going to get sick or what kind of medical treatments they will need. This uncertainty, and how people respond to it, is a key reason why we have the health institutions. Some of the issues with health insurance include:

(i) The Value of Insurance: Most people are risk averse. That is, they dislike uncertainty. Imagine that you face a choice between a certain income of ₦100, 000 and a 50-50 chance of income of ₦50,000 or ₦150, 000. The two options offer the same average income, but the second is riskier. If you prefer the certain ₦100, 000, you are risk averse. The same behaviour arises from the randomness of health spending. Suppose that some disease affects 2 percent of the population and that everyone is equally likely to be afflicted. Treatment costs ₦30,000. In this case, the expected cost of healthcare is 2 percent of ₦30,000, which is ₦600. If people are risk averse, they prefer to pay ₦600 with certainty over a 2 percent chance of having to pay ₦30,000. Giving people this option is the purpose of insurance. The general feature of insurance contracts is that a person facing a risk pays a fee (called a premium) to an insurance company, which in return agrees to accept all or part of the risk. Health insurance covers the risk of an expensive medical treatment. In our example, a health insurance company can charge a premium of ₦600 (or slightly more to make a profit) in exchange for promising to cover the cost of the ₦30,000 treatments for the 2 percent of its customers who get the disease. Markets for insurance are useful in reducing risk, but two problems impede their ability to do so efficiently.

(ii) Moral Hazard: The first problem that hinders the operation of insurance markets is moral hazard. When people have insurance to cover their spending on health care, they have less incentive to engage in behaviour that will keep that spending to a reasonable level. For example, if patients don't have to pay for each visit to a doctor, they may go too quickly when they experience minor symptoms (a runny nose, an achy finger). Similarly, physicians may be more likely to order tests of dubious value when they know an insurance company is picking up the bill. Health insurance companies try to reduce the problem of moral hazard by finding ways to encourage people to act more responsibly. For instance, rather than picking up the entire cost of a visit to a physician, they may charge patients co-pays of, say, ₦20 per visit to deter patients from making unnecessary visits. Similarly, insurance companies may have

strict rules about the circumstances under which they will cover the cost of certain tests that physicians order.

(iii) Adverse Selection: The second problem that impedes the operation of insurance markets is adverse selection. If customers differ in their relevant attributes (such as whether they have a chronic disease) and those differences are known to customers but not observable by insurers, the mix of people who choose to buy insurance may be expensive to insure. In particular, people with greater hidden health problems are more likely to buy health insurance than are healthy people. As a result, for an insurance company to cover its costs, the price of health insurance must reflect the cost of a sicker-than-average person. Even people with average health may see the high price and decide to go without insurance. As people drop coverage, the insurance market fails to achieve its purpose of eliminating the financial risk from illness. Adverse selection can lead to a phenomenon called the death spiral. Suppose that, because a person's health profile is private information, insurance companies must charge everyone the same price. At first, it might seem to make sense for a company to base the price of insurance on the health characteristics of the average person in the population. But after it does so, the healthiest people may decide that insurance is not worth the cost and drop out of the insured pool. With a sicker group of customers than expected, the company has higher costs and therefore has to raise the price of insurance. The higher price now induces the next healthiest group of people to drop insurance coverage, which drives up the cost and price again. As this process continues, more people drop coverage, the insured pool gets less healthy, and the price keeps rising. In the end, the insurance market may disappear. The problem of adverse selection has been central in the debate over health policy.

(d) Health care as a Right: Normally, when some people don't buy a good or service, perhaps because they think it costs too much given their income, that outcome is not a major problem for society. For example, suppose that a ticket to a cinema becomes expensive and lower-income consumers choose other forms of entertainment. We may argue that good theatre is not enjoyed more widely, but few would argue that this is a great

injustice. Health care is different. When a person gets sick, it seems wrong that a low income would be a reason to deny treatment. Health care, unlike a ticket to a cinema, is perhaps a human right. This judgment goes beyond the scope of economics and is best left to political philosophers, but we should acknowledge this belief as we study the economics of health care. In some ways, health care is like food. Food is necessary to survive, and as a society we try to ensure that everyone has the resources to get the food they need. There is, however, an important distinction between food and health care. Over time, the price of food has risen more slowly than incomes, and so affording an adequate diet has taken up a declining share of the typical household's budget. By contrast, because the cost of state-of-the-art health care has risen rapidly, affording it has required an increasing share of the typical household's budget. Health care being viewed as a right, along with its rising cost, has led to a large role for the government. In many countries, the government runs the health care system, financed mostly by taxes. This system is sometimes called single payer because one entity — the government's health service — pays all the bills. By contrast, in the United States and Nigeria, most people have private health insurance, often through their employers, but the government still has a sizable presence. In the United States, Medicare provides health insurance for those 65 and older; Medicaid provides health insurance for the poor; and the Affordable Care Act regulates the market for private health insurance and gives insurance subsidies to many lower-income households. There is a little doubt that, with health care often viewed as a human right, the government will continue to play a large role in the health care system.

(e) The Rules Governing the Health care Market Place: The importance of health insurance, whether provided by private companies or the government, requires that the market for health care works differently than most other markets in the economy. The typical markets — say, the market for rice — looks like panel (a) of Figure 1.1. The market has buyers and sellers. A seller offers a good or service at a price. A buyer who wants the item simply has to offer the right amount of money. An exchange is made, and soon the seller is counting her profit and the buyer is enjoying his rice.

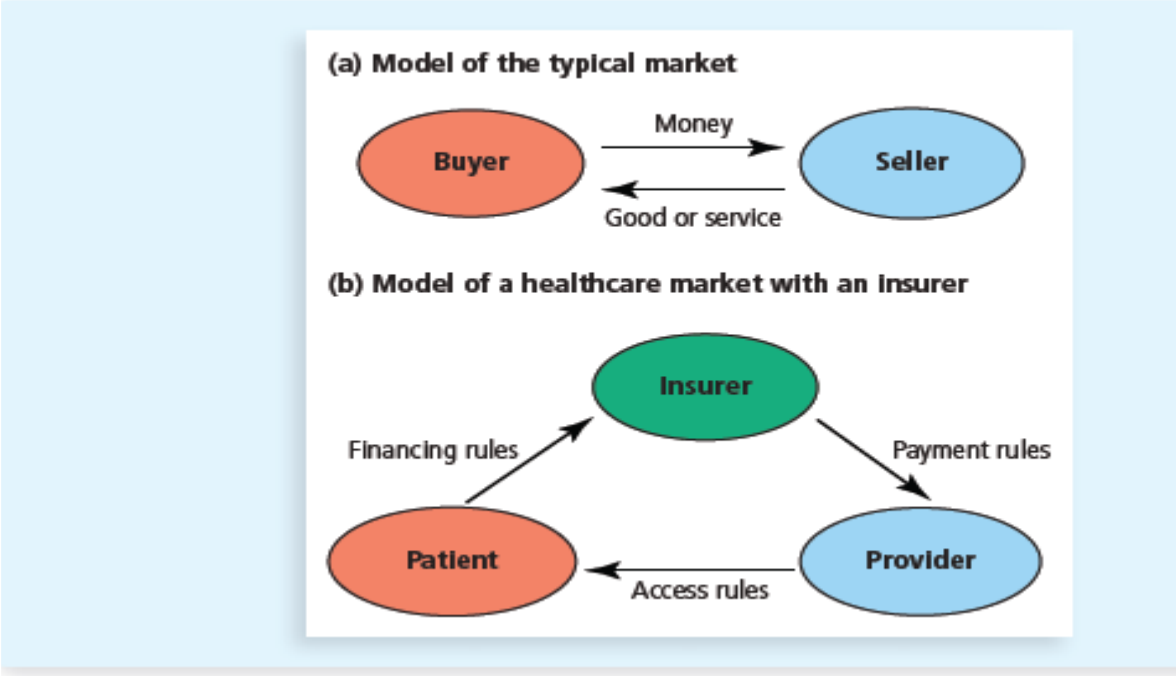


Figure 1.1: Models of Typical Market and Health Care Market

The market for health care looks more like panel (b). The provider (the seller of medical services) is not paid directly by the patient (the buyer). Instead, the patient pays money to an insurer in the form of either a premium (if the insurer is a private company) or taxes (if the insurer is the government). The insurer uses this money to compensate the provider, who in turn provides medical services to the patient. This process requires three sets of rules to guide behaviour. The first set determines the financing—that is, who pays for the insurance and how much they pay. If the insurer is the government, paying for health care becomes part of designing the tax system. If the insurer is a private company, health care is financed by the price that health insurance purchasers pay for their coverage. The price is set in the insurance market, which (like other markets) bases price on costs. In many cases, however, state and federal governments regulate the market for private insurance. For example, they may limit the extent to which companies can charge different prices based on age, sex, and pre-existing conditions. Thus, even when the financing of health care occurs between a patient and a private insurer, it is still shaped by public policy. The second set of rules determines a patient's access to health care. Since insured patients do not pay the marginal

cost of each medical service they consume, there is the possibility of overuse (moral hazard). To mitigate this problem of moral hazard, the insurer (whether the government or a private firm) has rules to limit access to when it makes sense. In other words, these rules ration the use of medical services based on estimated costs and benefits. For example, a patient may be able to get a routine check-up no more than once a year, may have access to only a limited number of doctors, or may need a referral from a general practitioner before making an appointment with a more expensive specialist. Such access rules are necessary because, once people have insurance to pick up the cost, market prices are no longer giving them the right signals about how to allocate scarce resources. The third set of rules determines the payments from insurers to providers. These rules establish both what an insurer will pay for and how much he will pay. Treatment prices influence which treatments providers guide patients toward. Insurers may deem some treatments too expensive, too experimental, or insufficiently valuable to pay for them at all. In such cases, providers will often not offer patients the services. Sometimes, however, providers will offer the services only if the patient pays the full cost of the treatment (as is often true with cosmetic procedures). In this case, the market for health care reverts from panel (b) in Figure 1.1 to the more typical market in panel (a). The rules regarding financing, access, and payment are related, and together they shape the kind of health care system a nation has. For nations with a government-run system, these rules are set by public policy. For nations with more private insurance, such as the United States, these rules are set by insurance companies as they compete for customers, subject to various government regulations.

SELF ASSESSMENT EXERCISE

Discuss the special features of health care market in any economy.

3.3 Measuring Health

Can health status be measured? Intuitively it is clear that a closer analysis of the use of resources for improving health conditions, for society or for single individuals, will depend on how a state of health is measured. It would be very helpful if a numerical measure of health were available, so that "marginal health effect" of each conceivable treatment might be computed as change in

health per monetary value spent in the treatment. As already mentioned, there are considerable difficulties connected with such a measurement. There is no obvious unit of measurement for health, and even the concept of "health" as such is not clear. This should not be a cause of despair, since most of the economic disciplines run into similar difficulties. Even when seemingly exact measures exist, problems show up at a closer analysis (such as e.g. in national accounts: What does the GNP actually measure?). On the other hand, it is clear that the analysis improves with more precise measures of the consequences of economic choices. Therefore, it is important to investigate how far one can get in measuring health.

This measurement problem encompasses all of health economics. At the outset it is easily seen that there can be no measurement of health corresponding to those of the national accounts (where it makes sense to consider differences of two measured values as an expression of the magnitude of the improvement), but one might still hope for constructing a suitable scale and positioning different health states on this scale in such a way that higher scale value corresponds to better health. There is also problem of interpersonal comparisons – is it possible to compare the measures of health of two persons, concluding that one of them has a better state of health than the other – and further on, can we aggregate the health of a whole society and then compare the health state of two different countries? It may be seen from this discussion, that it creates a more detailed argument about the nature of the scales on which health is to be measured (a discussion known from the distinction between cardinal and ordinal utility in consumer theory). Therefore, an overview of the methods for measuring health employed in practice is considered. Since health a priori is something ranging from perfectness to total absence of health (death), a scale for measuring health states can naturally be chosen as the interval of real numbers from 0 to 1. The approach taken is as follows: First of all, some fundamental characteristics of health are isolated; each of them describes certain aspects of health. The degree of fulfillment of the demand for perfect health in each of these aspects is then measured on a scale from 0 to 1 (or from 0 to 100). The difficult part of the measurement is then the weighing together of the scores in each of the health characteristics. For this a panel of individuals are questioned about the trade-offs between different states of health (where health is perfect in all except one of the

aspects) and the average evaluation is the used for weighing the scorings of each of the aspects together to an aggregate health score. A total of eleven characteristics were chosen: ability to move around, ability to hear, ability to talk, sight, ability to work, breathing, incontinency, ability to sleep, ability to eat, intellectual and mental functioning and social activity. For each of these characteristics a numerical value is determined belonging to a precisely described state of imperfect functioning. For example, the ability to move around is specified as follows:

– normal ability to walk, both outdoor and indoor and on stairs,
– normal ability for indoor movement, but outdoor movement and/or movement on stairs with trouble,
– can move around indoor (possibly using equipment, but outdoor and/or on stairs only with help from others,
– can move around only with help from others, also indoor,
– conscious, but bed stricken and unable to move around; can sit in a chair if aided,
– unconscious,
– dead.

The people interviewed will be asked to assign numbers between 0 and 100 to each of the described situations, so that the most desirable state gets the value 100 and the less desirable 0; the remaining states would be evaluated so that if for example, the number 75 is assigned to a state which is 3/4 as desirable as the best one, 33 to a state which is only 1/3 as desirable as the best one, etc. (whether it at all makes sense for the interviewed to desire something "3/4 as much" as something else is another question entirely). With the described characteristics it is possible to move on the aggregation phase, which proceeds more or less in the same way, having the interviewed assign numbers to the most desirable state within each of the 11 characteristics on a scale from 0 to 100. When this has been done, numbers can be assigned to all the described states of health by multiplying the score obtained within the characteristic by the score of the characteristic (now all scores are taken as numbers between 0 and 1) and adding over the 11 characteristics. The method has the advantage of being rather simple and easy to understand (what is not always the case in health status measurement, for example the "standard gamble method" in Quality Adjusted Life Years (QALYs) measurement, which presupposes certain knowledge of probability on the side of the interviewed. The results show a considerable degree of coincidence in the answers of different individuals, which gives some promise that the

measurement results are well founded. Again it must be said that the measurement has no obvious theoretical foundation. If state of health is something to be measured in an objective way – which certainly is not to be excluded and indeed is the basic idea behind the measurements attempted – it would be comforting to have at least some conjecture of the reason why such a shared ranking of health states should exist. Indeed, the economist is accustomed to take the opposite viewpoint, namely that people a priori have very different tastes and desires (and this is indeed what makes trade possible), so that an observation of identical preferences would call for a special explanation. So far it has been the other way round in health state measurement; preferences are for some unexplained reason assumed to be identical among individuals. The method assumes that the individual rankings made for each of the characteristics involved are independent of the state of events in the other characteristics. This assumption is dubious – if you happen to be in the unconscious state described above, you might well be pretty indifferent as to whether you can read a newspaper without glasses or whether you cannot move around without a dog. This is the property of independence which is at stake, and though not always reasonable, it is often assumed in order to have a manageable preference relation in contexts of empirical investigations. There is always a trade-off between theoretical purity and practical applicability, and seen in this light the assumption of independence is quite acceptable. Even stronger assumptions may be accepted if they open up for practical measurement of health status, a field which has so many potential applications.

3.3.1 Health Status Measures

Measurement of health status has been carried through by several researchers over the years, and there is a steadily increasing activity in this field. This is partly explained by the fact that a measure of how patients consider their own situation – self-experienced health – is important in medical research, and in particular it is important to have a method of measurement which is reasonably objective, so that improvement in health conditions may enter the medical documentation of new medicine or new methods of treatment. There is need for documented effects of treatments and the discussion of a suitable choice of “outcome” or “end points” of a medical intervention points to the need for such measurements. In many cases, the directly

observable outcomes relate directly to treatment rather than to the effect on the general health condition of the patients, and this takes us back to health status measurement. The trend has been to include more and more aspects which involve "quality of life"; however, it should be added that even if quality of life is important, certain more general aspects of quality of life should be left out (social status, housing conditions, education), so that what is wanted is what should properly be called "health-related quality of life". The most commonly used health status measures are given in Table 1.1.

Table 1.1: Some Commonly Used Health Status Measures

	QWB	SIP	HIE	NHP	EQOL	SF-36
Aspects:						
Physical function	•	•	•	•	•	•
Social function	•	•	•	•	•	•
Role function	•	•	•	•	•	•
Mental prob.		•	•	•	•	•
Self-experienced health			•	•	•	•
Pain		•	•	•		•
Energy/fatigue	•		•	•		•
Mental condition			•			
Sleep		•		•		
Cognitive functions		•				
Quality of life		•				
Reported change						•
Method:						
Administration	I,T	S,I,T	S,P	S,I	S	S,I,T
No. of questions	107	136	86	38	9	36
Scoring method	SI	P,SS,SI	P	P	SI	P,SS

Signature
QWB = Quality of Well-Being Scale (1973)
SIP = Sickness Impact Profile (1976)
HIE = Health Insurance Experiment Surveys (1979)
NHP = Nottingham Health Profile (1980)
EQOL = European Quality of Life Index (1990)
SF-36 = MOS 36-Item Short-Form Health Survey (1992)
Method of administration: S = Self, I = Interviewer, T = Third party
Scoring: P = Profile, SS = summarized scores, SI = Index

As can be seen from the table, there are several distinct proposals in the literature as to how health status of health related quality of life should be measured, which aspects of health should be included, which methods of observation (administration of questionnaires) to be applied, and how the result of the measurement should be presented. At the general level there seems to be agreement that health has both physical and mental aspects.

SELF ASSESSMENT EXERCISE

Write short notes on Quality Adjusted Life Years (QALYs).

3.4 Future Challenges to Health Care Systems

Actors on markets are under continuous pressure to adjust. Consumers' changes in taste lead to changes in demand, new technologies provide rivals with competitive advantage, and public authorities step in to regulate or even prohibit business. This pressure to adjust is spread by price signals indicating to firms the need to adapt their goods and services to new circumstances. In health care, however, fluctuating market prices for medical services are incompatible with the key principal-agent relationship between the patient and the service provider because they might violate both the participation and the incentive compatibility constraints. One possibility of avoiding fluctuating prices is bargaining over fee schedules, which paves the way for the important role of professional associations and public authorities in health care. The inflexibility of fees and prices is further enhanced by the fact that purchases of health care goods and services such as pharmaceuticals abroad are often legally prohibited. This serves to insulate domestic markets from international shocks but also competition. This departure from allocation through prices, however, has the adverse effect of reducing the system's speed of adjustment. For example, structural adjustment of a fee schedule usually takes years. On the one hand, this sluggishness prevents physicians, dentists, and hospitals from initiating and swiftly concluding contractual agreements with insurers because of temporary advantages, which would be to the detriment of many patients. On the other hand, it causes considerable delays in adjustment to exogenous shocks. The ensuing disequilibria are perceived as 'challenges to health care systems'. Such challenges have emerged in four areas:

(i) The technological challenge: In 1980 the magazine Newsweek presented the following medical innovations: a new piece of equipment of significance comparable to the CT scanner which makes brain-waves visible, revolutionary surgical methods for eliminating shortsightedness and infertility in women, new drugs for jaundice, sexually transmitted diseases and gout, various new cancer treatments, an operation for safe implantation of artificial breast following breast removal in females, new life-saving techniques in child heart surgery, and a new type of electric shock treatment to regenerate muscle and nerve tissue. Almost all of these innovations are product innovations, i.e., they save lives or contribute to an improved quality of life, although at (much) higher cost. "When Christian Barnard transplanted the first

human heart on 3rd December 1967, at that very moment the cost of such a treatment rose from zero to US$110,000". Conversely, process innovations which enable a particular service to be produced at lower cost are rare. There seem to be even fewer organizational innovations in health care, which promise cost savings through a rebounding of production processes resulting in economies of scope. Thus, technological change in medicine threatens to become the driving force of future 'cost explosions'.

(ii) The demographic challenge: At first sight, this is simply to say that more and more people are getting older. Old age is associated with an increased demand for medical and in particular nursing services, and the issue is how to meet this demand. On closer inspection, however, it is more the proximity to death than calendar age that seems to matter. This would mean that the last, expensive year of life would simply occur at old age. It is another demographic change that may turn out to be a more important driver of health care expenditure. In the United States, the number of single person greatly increased, e.g., from 13 to 25 percent within just twenty years. Persons living alone are much less able, in the case of illness, to fall back on support and care from relatives, causing them to demand more health care services.

(iii) The challenge of the 'Sisyphus Syndrome': The success of modern medicine reminds one of Sisyphus, the hero of Greek mythology who was condemned to roll a lump of rock up only to see it slip out of his grasp just before reaching the summit, forcing him to start all over again. By prolonging human life, technological change in medicine may also increase the number of those who make more than average demands on the health care system. Because of their increased political clout, the elderly might exert their influence in public health care expenditures. As a result of this process, the success of medicine may turn into a growing burden on the economy and on society.

(iv) The challenge of international competition: This often neglected challenge to health care has its origin in the increasing economic integration of nations. With labour able to move about freely within the European Union, workers will be attracted not only by attainable income but also by things such as the relative performance of a country's health care system. Physicians and medical staff will also be able to move about more freely, and international

direct investment into private health insurance and hospitals will become more prominent. National health care systems will increasingly be subject to international competition.

The focus of the future challenges to health care systems was on technological and demographic change and on the increasing globalization of health care markets. The main results are:

(i) Innovation in health care must achieve a certain improvement of health status. From the ex-ante point of view of an individual, the benchmark for acceptance is equal for process and product innovation but lower for organizational innovation.

(ii) In the transition from the individual to the social level, the required standards of performance for all three types of innovation in health care are lowered. Insurance-induced moral hazard and medical imperatives at the level of objectives and means constitute reasons for this.

(iii) The ageing of the population risks the financial equilibrium of social health insurance. The adjustment of contributions for safeguarding equilibrium may turn membership into an unprofitable investment for present and future generations of workers.

(iv) Empirical analysis using OECD country data suggests that additional health care expenditure does contribute to higher remaining life expectancy at higher ages and that higher life expectancy can be claimed to bias spending in favour of health care expenditure. The Sisyphus Syndrome is also found to be of some importance in that 68 percent of extra expenditure in the current year carries over to the following year.

(v) There are three hypotheses with regard to the impact of ageing on health care expenditures. According to the status-quo hypothesis, present age profiles can be extrapolated into the future. The expansion-of-morbidity hypothesis predicts increased survival of costly individuals that would have died. The time-to-death hypothesis claims that the expensive years prior to death will simply be shifted to higher ages.

(vi) In social health insurance, capital funding is not necessary to alleviate the financial burden falling on future generations. The same objective can be achieved through age-dependent contributions, which would necessitate capital accumulation in private households, where it is better protected from political pressure than in the hands of a social insurer.

(vii) Economic integration will expose domestic social health insurers and health care providers to international competition, at least within the European Union. In comparison, international migration of physicians and nursing staff represents a less pressing challenge in the future.

SELF ASSESSMENT EXERCISE

Discuss the future challenges to health care systems and the likely results that may originate.

4.0 CONCLUSION

We can conclude that health economics as a scientific discipline started more than seventy years ago, in the sense that the specific treatment of topics related to the economics of the health care sector has become common. Also, a well-known definition of "health" proposed by the World Health Organisation (WHO) was examined. The argument surrounding the measurement of health status and various measures of health status was also discussed. Finally, future challenges of health care system and results emanated from this were as well examined.

5.0 SUMMARY

In this unit, we have discussed vividly the discipline of health economics and challenges therefrom, special characteristics of health care, measurement of health status, as well as future challenges of health care system.

6.0 TUTOR-MARKED ASSIGNMENT

1. There are three hypotheses with regard to the impact of ageing on health care expenditures. Discuss these hypotheses.
2. Discuss the future challenges to health care systems and the likely results that may emanate from these challenges.
3. Discuss the special features of health care market in any economy.

7.0 REFERENCES/FURTHER READING

McGuire, A., J. Henderson, and G.Mooney (1988) The Economics of Health Care, Routledge and Kegan Paul, London, 1988.

Mooney, G.H., (1986) Economics, Medicine and Health Care, Harvester Wheatsheaf, 1986.

Mougeot, M. (1986) Le syst`eme de Sant´ e, Centralisation ou d´ecentralisation?, Economica, Paris, 1986.

UNIT 2: The Four Basic Questions

CONTENTS

1.0 INTRODUCTION

The study of health economics involves the application of various microeconomics tools, such as demand or cost theory to health issues and problems. The goal is to promote a better understanding of the economic aspects of health care problems so that corrective health policies can be designed and proposed. A thorough understanding of microeconomic analysis is essential for conducting sound health economics analysis. The tools of health economics can be applied to a wide range of issues and problems pertaining to health and health care. For example, health economics analysis might be used to investigate why 29 out of every 1,000 babies born in Nigeria died before their first birthday, whereas all but 3 out of 1,000 babies born in Japan live to enjoy their first birthday cake. The tools of health economics can also be used to understand the economic desirability of a contested merger between two large hospitals in a major metropolitan area. The main question is, will the merger of the two hospitals result in lower hospital prices due to overall cost savings or higher prices due to monopoly power?

Health economics is difficult to define in a few words because it encompasses such a broad range of concepts, theories, and topics. The Mosby Medical Encyclopedia (1992, p. 361) defines health economics as follows:

Health economics….. *studies the supply and demand of health care resources and the impact of health care resources on a population.*

Notice that health economics is defined in terms of the determination and allocation of health care resources. This is logical because medical goods cannot exist without them. Health Care resources consist of medical supplies, such as pharmaceutical goods, personnel such as physicians and lab assistants and capital inputs including nursing homes and hospital facilities, diagnostic and therapeutic equipment and other items that provide medical care services. Unfortunately, health care resources, like resources in general, are limited or scarce at a given time and wants are limitless. Thus, trade-off is inevitable and a society, whether it possesses a market-driven or a government-run health care system, must make a number of fundamental but crucial choices. These choices are normally couched upon four basic questions

(i) What combinations of non-medical and medical goods and services should be produced in the macro-economy?
(ii) What particular medical goods and services should be produced in the health economy?
(iii) What specific health care resources should be used to produce the chosen medical goods and services?
(iv) Who should receive the medical goods and services that are produced?

How a particular society chooses to answer these four questions has a profound impact on the operations and performance of its health economy.

The first two questions deal with allocative efficiency: what is the best way to allocate resources to different consumption uses? The first decision concerns what combinations of goods and services to produce in the overall economy. Individuals in a society have unlimited wants regarding non-medical and medical goods and services, yet resources are scarce. As a result, decisions must be made concerning the mix of medical and nonmedical goods and services to provide, and these decision making process has trade-offs. If more people are trained as doctors or nurses, fewer people are available to produce nonmedical goods such as food, clothing, and shelter. Thus, more medical goods and services imply fewer nonmedical goods and services and vice versa, given a fixed amount of resources. The second consumption decision involves the proper mix of medical goods and services to produce in the health economy. This decision also involves trade-offs. For example, if more health care resources, such as nurses and medical equipment, are allocated to the production of maternity care services, fewer resources are available for the production of nursing home care for elderly people. Allocative efficiency in the

overall economy and the health economy is achieved when the best mix of goods is chosen given society's underlying preferences.

The third question – what specific health care resources should be used? – Deals with production efficiency. Resources or inputs can be combined to produce a particular good or service in many different ways. For example, hospital services can be produced in a capital or labour-intensive manner. A large amount of medical equipment relative to the number of patients served reflects a capital-intensive way of producing hospital services, whereas a high nurse-to-patients' ratio indicates a labour-intensive process. Production efficiency implies that society is getting the maximum output from its limited resources because the best mix of inputs has been chosen to produce each good.

2.0 OBJECTIVES

At the end of this unit, students should be able to understand:

(i) Allocative and Production efficiency in health care
(ii) Distributive efficiency in health care
(iii) Implications of the ways society chooses to answer allocative and distributive questions.

3.0 MAIN CONTENT

3.1 Production, Allocative Efficiency and the Production Possibilities Curve

The most straightforward way to illustrate production and allocative efficiency is to use the production possibilities curve (PPC). A PPC is an economic model that depicts the various combinations of any two goods or services that can be produced efficiently given the stock of resources, technology and various institutional arrangements. Figure 1.2 shows a PPC. The quantity of maternity services, M, and nursing homes, N, are shown on the vertical and horizontal axes, respectively (assuming society has already made its choice between medical and nonmedical goods). Points on the bowed-out PPC depict the various combinations of maternity and nursing home care services that can be produced within a healthy economy assuming that the amounts of health care resources and technology are fixed at a given point in time. Every point on the PPC implies production efficiency, since all health care resources are

being fully utilized. Examples are points A, B, C, D and E on the PPC. At each of these points, medical inputs are neither unemployed nor underemployed (for example, a nurse involuntarily working part time rather than full time) and are being used in the most productive manner so that society is getting their maximum use. If a movement along the curve from one point to another occurs, units of one medical service must be forgone to receive more of the other medical service.

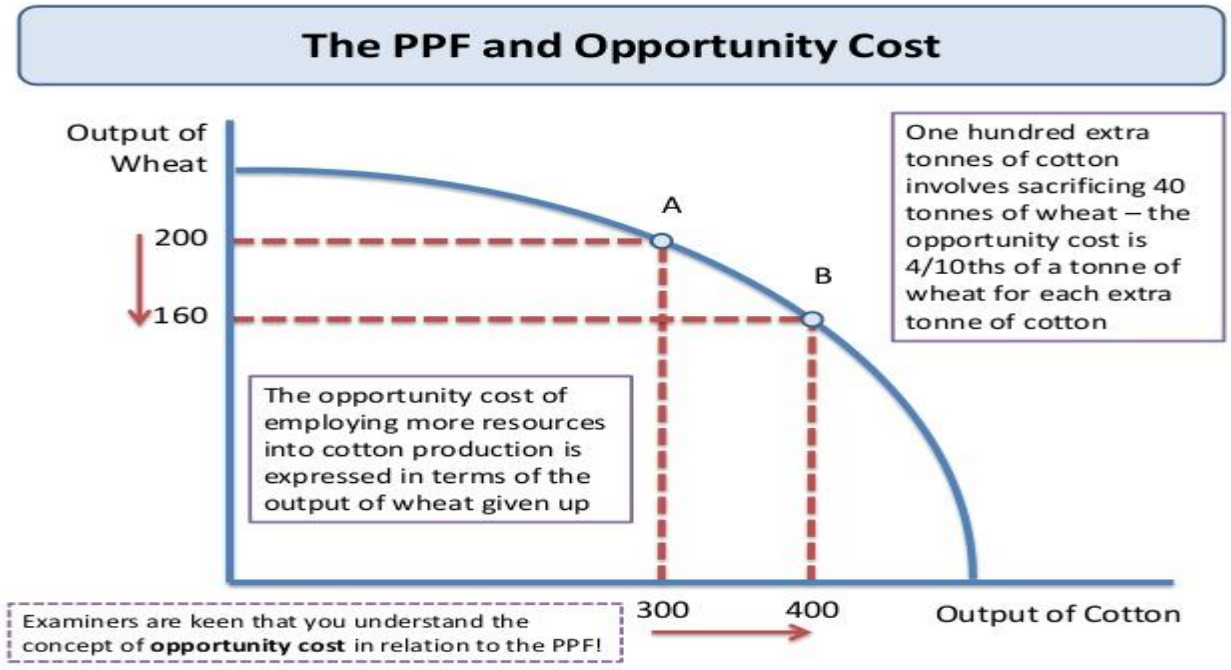


Figure 1.2 Production Possibility Curve

Specifically, assume the health economy is initially operating at point C with M_C units of maternity care services and N_C units of nursing home services. Now suppose health care decision makers decide that society is better off at point D with one more unit of nursing home services, N_D - N_C. The movement from point C to point D implies that M_C – M_D units of maternity services are given up to receive the additional unit of nursing home services. Because medical resources are fully utilized at point C, a movement to point D means that medical input must be reallocated from the maternity care services to the nursing home services market. As a result, the quantity of maternity care services must decline if an additional unit of nursing home services is produced. The forgone units of maternity care services, M_C -M_D, represent the opportunity cost of producing additional unit of nursing home services. Generally, opportunity

cost is the value of the next best alternative that is given up. The bowed-out shape of the PPC implies that opportunity cost is not constant but increases with a movement along the curve. Imperfect substitutability of resources is the reason for the so-called law of increasing opportunity cost. For example, suppose the nursing home services market expands downward along the PPC, to produce more nursing home services, employers must bid resources away from the maternity care services market. Thus, the law of increasing opportunity cost suggests that ever-increasing amounts of one good must be given up to receive successively more equal increments of another good.

If medical inputs are not fully utilized because some inputs are idle or used unproductively, more units of one medical service can be produced without decreasing the amounts of another medical service. An example of underutilization is indicated by point F in the interior of the PPC. At point F, the health care system is producing only M_F units of maternity services and N_F units of nursing home services. Notice that by moving to point B on the PPC, both maternity care services and nursing home services can be increased without decreasing the other. The quantity of both services increases because some resources are idle or underutilized at point F. Health care resources are inefficiently employed at point F. A point outside the current PPF, such as G, is attainable if the stock of health resources increases, a new productivity-enhancing technology is discovered, or various economic, political, or legal arrangements change and improve productive relationship in the economy. If so, the PPC shifts out and passes through a point like G. Production efficiency is attained when the health economy operates at any point on the PPC, since medical inputs are producing the maximum amount of medical services and no unproductive behaviour or involuntary unemployment exists. Allocative efficiency is attained when society chooses the best or most preferred point on the PPC. All points on the PPC are possible candidates for allocative efficiency. The optimal point for allocative efficiency depends on society's underlying preferences for the two medical services.

SELF ASSESSMENT EXERCISE

Discuss production and allocative efficiency in the health economy.

3.2 The Distribution Question

The answer to the fourth question – who should receive the medical goods and services deals with distributive justice or equity. It asks whether the distribution of services is equitable or fair, to everyone involved. In practice, countries around the world have chosen to address this distribution question involving medical care in many different ways. When thinking about the distribution question, it is sometimes useful to consider two theoretical opposite ways of distributing output: the pure market system and a perfect egalitarian system. Goods and services are distributed in a pure market system based on each person's willingness and ability to pay because decisions concerning the four basic questions are answered on a decentralized basis within a system of markets. That is, goods and services are distributed or rationed, to only those people who are both willing and able to purchase them in the market place. Because people face an incentive to earn income to better afford goods and services in a pure market system, they tend to work hard and serve appropriately for present and future consumption. Consequently, productive resources tend to be allocated efficiently in a pure market system. In other words, the incentives associated with the pure market system means that the economy operates on the PPC. In many cases, differences in ability to pay among individuals reflects some have consciously chosen to work harder and save more than others. Unfortunately, differences in ability to pay may also indicate that some people have less income because of some unfortunate circumstances of life such as a mental, physical or social limitation. Regardless of the specific reasons, it follows that people without sufficient incomes face a financial barrier to obtaining goods and services in a pure market system in which price serves as a rationing mechanism. Given income disparities some people may be denied access to needed goods and services. Consequently, the pure market system is viewed as inherently unfair by many when it comes to the distribution of important goods and services such as health care.

In direct contrast, a central committee, such as a federal or subnational unit of government, may answer the distribution question by ensuring that everyone receives an equal share of goods and services. In an egalitarian system of this kind, everyone has access to the same goods and services without regard to income status or willingness to pay. Thus, no one is denied access to

needed goods and services. But an incentive may exist for people to choose to work and save less because the consumption decision is divorced from the distribution of earned income. Because of this inefficient allocation of resources, fewer goods and services may be available for distribution in an egalitarian system. In this case, the economy may operate inside the PPC.

Most countries have adopted a mixed distribution system in practice, with the reliance on central versus market distribution varying by degree across countries. For example, in the United States, many goods and services are distributed by both the market and the government. The temporary assistance for needy families and Medicaid represent some of the many policies adopted by the government to redistribute goods and services in the United States. Some people applaud this programme, whereas others argue that they worsen both efficiency and equity. They argued that efficiency and equity are compromised when those who choose to commit fewer resources to production are rewarded through redistributive programmes and productive individuals are penalized via taxation. The efficiency and equity implications of redistributive policies are constantly debated in various countries. In the context of the health care, the consequences of this debate regarding distribution might determine who lives and who dies.

SELF ASSESSMENT EXERCISE

Discuss different distribution policies and their efficiency and equity implications

3.3 Implications of the Four Basic Questions

Given a scarcity of economic resources, a society generally wishes to produce the best combinations of goods and services by employing least-cost methods of production. Trade-offs is inevitable. As the PPC shows, some amount of good or service must be given up if the production of one good or service increases. As a result, each society must make hard choices concerning consumption and production activities because scarcity exists. Choices may involve sensitive trade-offs, for example, between the young and the old, between prevention and treatment, or between men (prostate cancer) and women (breast cancer). In addition, some individuals lack financial access to necessary goods and services such as food, housing, and medical care. Because achieving equity is a desirable goal, a society usually seeks some redistribution of income. The redistribution normally involves taxation. However, a tax on

labour or capital income tends to create disincentives for employing resources in their most efficient manner. Inefficient production suggests that fewer goods and services are available in the society (production inside PPC). Thus, a trade-off exists between equity and efficiency goals, and, consequently, hard choices must be made between the two objectives. The design of a nation's health care system normally reflects the way the society has chosen to balance efficiency and equity concerns.

SELF ASSESSMENT EXERCISE

The design of a nation's health care system normally reflects the way the society has chosen to balance a trade-off between efficiency and equity. Discuss.

3.4 Economic Models and Analysis

The production possibility curve is an example of economic model. Models are abstractions of reality and are used in economics to simplify a very complex world. Economic models can be stated in descriptive form (verbal), graphical, or mathematical form. Usually, an economic model like the PPC describes a hypothesized relation between two or more variables. For example, suppose the hypothesis is that health care expenditures, *E,* are directly *(as opposed to inversely)* related to consumer income, *Y*. It simply means that expenditures on health care services tend to rise when consumer income increases. Mathematically, a health care expenditure function can be stated in general form as

$$E = f(Y) \text{ ---------- } (1.1)$$

Equation (1.1) implies that health care spending is a function of consumer income; in particular, health care expenditures are expected to rise with income. An assumption underlying economic models is that all factors, other than the variables of interest, remain unchanged. For example, the hypothesis that health care expenditures are directly related to income assumes that all other likely determinants of health care spending, such as prices, tastes, and preferences, remain constant. Economists normally qualify their hypothesis with the Latin phrase *ceteris paribus*, meaning "all things held constant". By holding other things constant, we can isolate and describe the pure relation between any two variables. The expenditure function in equation (1.1)

is expressed in general mathematical form, but a hypothesis or model is often stated in a specific form. For example, the following equation represents a linear expenditure function for health care services:

$$E = a + bY \text{ ---------- (1.2)}$$

where *a* and *b* are the fixed parameters of the model. This equation simply states that health care expenditures are directly related to consumer income in a linear (rather than nonlinear) form. Mathematically, the parameter *a* reflects the amount of health care expenditures when income is zero, whereas, b is the slope of the expenditure function. The slope measures the change in health care expenditures that results from a one-unit change inn income, or $\Delta E/\Delta Y$. For example, assume the parameter a equals ₦1, 000 per year and b equals one-tenth or 0.1. The resulting health care expenditure function is thus:

$$E = 1{,}000 + 0.1Y \text{ ---------- (1.3)}$$

Equation (1.3) implies that health care expenditures rise with income. The slope parameter of 0.1 suggests that each ₦1, 000 increase in consumer income raises health care spending by ₦100. Note that the expenditure function clearly represents the hypothesis concerning the direct relation between income and health care spending.

3.4.1 Positive and Normative Analysis

Health economists perform two types of analysis. Positive analysis uses economic theory and empirical analysis to make statements or predictions concerning economic behaviours. It seeks to answer the question "What is?" or "What happened?" For example, we might investigate the exact relation between income and health care spending. Because positive analysis provides explanations or predictions, it tends to be free of personal values. Normative analysis, on the other hand, deals with the appropriateness or desirability of an economic outcome or policy. It seeks to answer the question "What ought to be?" or "Which is better?" For example, an analyst might conclude that household with income less than ₦30, 000 per year should be subsidized by the government because they are unable to maintain a proper level of health care spending.

Naturally, this implies that the analyst is making a value judgement statement. Because opinions vary widely concerning the desirability of any given economic outcome and the role the government should play in achieving outcomes, it is easy to see why normative statement generally spark more controversy than positive ones. For instance, when 518 health economists were asked whether the Canadian health care system is superior to the U.S. system, there was much disagreement. Fifty-two percent of the economists agreed, and 39 percent disagreed with the statement. The remaining 10 percent had no opinion or lacked the information needed to respond to the question. The following sets of positive and normative economic statements should give a better understanding of the difference between the two. Notice that the positive statements deal with what is or what will be, whereas the normative statement concern what is better or what ought to be.

Positive: According to Becker and Murphy (1988), a 10 percent increase in the price of cigarettes leads to a 6 percent reduction in the number of cigarettes consumed.

Normative: The government should increase the tax on cigarettes to prevent people from smoking.

Positive: National health care expenditures per capita are higher in the United States than Canada

Normative: To control health care expenditures, the United States should adopt a national health insurance programme similar to Canada's.

SELF ASSESSMENT EXERCISE

Discuss positive and normative questions as it applies to health care sector.

4.0 CONCLUSION

Scarcity of economic resources makes choices necessary. This necessitates choosing between efficiency and equity and answering four important economic questions in a healthy economy. Health economics employed economic tools of analysis in answering some of these questions. This unit concludes that the way society chooses to answer these questions determines who get healthy and who dies.

5.0 Summary

This unit discussed the four basic economic questions in any healthy economy. It looks at how these questions are answer and the implications of these questions on allocative, productive and distribution efficiency. It also looks at the economic tools of analysis and positive and normative questions in health economics.

6.0 Tutor-Marked Assignment

Health policy experts have often argued that the health industry is different from other industries. The "differentness" argument says that health care is too important to be treated as a commodity and that the health industry does not function like a market in any meaningful way. One of the most often cited reasons for the latter – that health care markets are not anything like normally functioning markets – is the prevalence of high levels of information asymmetry in most health care transactions. Critics of this viewpoint, on the other hand, submit that health care is really not different from other industries, and that information asymmetry alone explains the very view of the unique features of health care industry. Although health care is essential to health, so also are food and shelter, yet those industries are not considered "different" and perhaps more important, and do not exhibit the high levels of organizational heterogeneity, laws, and entry barriers common to the health care industry. In other words, they are essential to health but allowed to function more or less as free markets.

Questions for Discussion

1. List three to five points in favour and against the differentness argument.
2. What are the merits and demerits to thinking of health care as an industry like many other industries?

7.0 References/Further Reading

Donaldson Cam and Karen Gerard (1993) Economics of Health Care Financing: The Visible Hand. Macmillan Press Ltd. London.

Folland S., A. Goodman & M. Stano (2010) The Economics of Health & Health Care, Sixth Edition, Prentice Hall, New Jersey.

Jacobs, P. (1991) The Economics of Health and Medical Care Maryland: Aspen Pub Inc. Jack,

Williams (1964) Principles of Health Economics for Developing Countries. WBI Development Studies. The World Bank, Washington D. C.

Jones Andrew (2007) Applied Econometrics for Health Economists: A Practical Guide, 2nd Edition OHE

Phelps Charles E. (1992) Health Economics, New York: Harper Collins Pub Inc.

Santerre E. & S.P. Neun (1996) Health Economics: Theories, Insights & Industry Studies, Irwin, Chicago.

Zweifel P., F. Breyer & M. Kifmann (2009) Health Economics, Second Edition, Springer Verlag Heidelberg.

Unit 3: The Main Features of the Health Care Service and Its Relation with Economic Development

CONTENTS

1.0 INTRODUCTION

Health economics can be seen as the application of economic theories, tools and concepts of economics as a discipline to the topics of health and health care. Since health economics is concerned with issues relating to the allocation of scarce resources to improve health, this includes both resource allocation within the economy to the health sector and within the health care system to different activities and individuals. In Nigeria and other countries of the world, the need for health care is increasing due to rapid population growth and changes in disease pattern. Related with this, health care costs are expected to be rapidly increasing. Apart from explosion of costs, inequity, misallocation and inefficiency are believed to be serious challenges to the health care system. These problems put a considerable strain on the limited health care resources. There are also very visible signs of change in the health care market. Attention is shifting from the "passive" funding and administration of systems, in which physicians identify and provide appropriate care, to concerns about the resource costs of care and the health outcomes achieved from providing care. The principles of health economics consider supply and demand issues and how the two might interact given that the standard market solution generally fails due to problems such as adverse selection, moral hazard, asymmetric information and supplier induced demand. This unit considers features of the health care, distinctive

characteristics of health care services from other commodities, the political economy of health care and health and economic development.

2.0 OBJECTIVES

At the end of this unit, the students should be able to:

1. Understand health as one of the social sectors with economic implication.
2. Understand the specific nature of the health care service in implementing economic principles and techniques.

   Be able to know the implications of economic development to the health care services.
3. Understand the effect of some economic factors on health status of society.
4. Identify the ways through which improvement of the health system can create conducive conditions for economic development and vice versa.

3.0 MAIN CONTENT

3.1 General Features of the Health Care

There are different understandings of health – each with different implications for the roles of government. It is important to recognize, first, the difference between health and 'health care'. The term health refers to a state either of an individual or of a community. A number of factors including 'health care' may influence this state of health. However, other factors that affect health are poverty, level of education, food intake, access to clean water and sanitary and housing conditions. The narrowest concept of health sees it as a measure of the state of the physical body organs. An individual is unhealthy if there is a malfunctioning of part of the body. A broader, but related, definition sees health just in terms of the mechanics of the different bodily organs, but in the ability of the body as a whole to function.

In contrast, the WHO definition of health as "a state of physical, mental and social well-being and not merely the absence of disease or infirmity" indicates a clear shift away from earlier narrow organic or functionally-based definition of health to a more holistic view, it sees the health of an individual or community as being concerned not only with physical (and mental) status, but also with social and economic relationships.

Individual conceptualization of health will affect the type of intervention and planning that is possible. The narrowest definitions are closely associated with a medical model of health in which the role of health services is seen as paramount in restoring the functioning of the unhealthy body. Wider primary health care concepts suggest that broader interventions, including community empowerment and anti-poverty measures, are necessary to promote health. Three perspectives can be used to distinguish health on the importance of health and on the possible roles of the state in promoting it:

(a) Health as a right: This is viewed by some as a right similar to justice or political freedom. Indeed, the WHO constitution states that '… the enjoyment of the highest attainable standard of health is one of the fundamental rights of every human being without distinction of race, religion, political belief, economic or social condition'. Although it is difficult to believe that equal health status is attainable in the same way that equal political freedom may be, health is seen as so fundamental that constraints to its full attainment must be minimized. In part, this involves ensuring access to health care. The government is seen as having a responsibility to ensure this, comparable with its role in ensuring equal justice. According to such a view, a government will be particularly concerned with issues of equity in health and health care.

(b) Health as consumable good: Health is also seen as an important individual objective that is not comparable with justice, but rather with material aspects of life. Such a view often refers to health as consumption good. The government here has no special responsibilities in the promotion of health, but leaves decisions as to its comparative importance to individual consumers. The role of the state under such a view might be limited to ensuring that the health care provided is of adequate quality (such as ensuring professional standards in the same way that it would monitor the quality of any good or service, such as food).

(c) Health as an investment: A third view of health is that it is important, but largely it affects the productive ability of the workforce. Illness may affect overall production, either through absenteeism or by lowering productivity through its debilitating effects.

SELF ASSESSMENT EXERCISE

Discuss the three main perspectives of distinguishing health on the importance of health and on the possible roles of the state in promoting it:

3.2 Distinctive Characteristics of the Health Care Services from other Commodities

Why not leave health care to the market? Most people believe that you cannot buy and sell health care like other goods and services. They believe that health care is different. This is what is sometimes called a "common-sense" approach to the issue. Economists approach the same question differently. For many people, the word market conjures up a picture of a town square with lots of small stallholders selling everything, from fruit and vegetables to meat and fish. For economists, the term has a much wider meaning:

(i) It is used to describe any process of exchange between buyers and sellers.

(ii) Formally, a market could be defined as any set of arrangements that allows buyers and sellers to communicate and thus arrange exchange of goods, services or resources.

(iii) A free market is where such exchange occurs without interference from the government.

(iv) Information is a vital ingredient for any market. Both buyers and sellers need to have access to sufficient information to allow them to make rational decisions.

In theory, markets produce goods and services in the right quantities and at the lowest possible cost. This is why markets are so powerful. Nevertheless, in the real world markets do not always work in the way theory predicts. It is possible for a free market to produce a Pareto inefficient result - i.e. the market fails. A Pareto inefficient situation can occur as a result of the following:

(a) Imperfect information: We get goods at the lowest possible cost if the market is able to transmit all the information about benefits and costs between producers and consumers. If this information is less than perfect, then the market will fail. Think about buying a CD. You know what a CD is, and you will have a good idea of the kind of music on the disc. Therefore, you are able to relate your benefit to the price of the CD. If we look at the market for CDs, people will go on buying CDs until the extra satisfaction from the last CD is exactly equivalent to the price

of the CD. However, health care is rather different from CDs. We face very acute information problems, which make rational purchasing decisions difficult. For instance, most people do not know the best way to treat a stomach ulcer so they would find it difficult to buy such treatment. This analysis also assumes that the only people receiving benefit or satisfaction from the CDs are the people buying them. In other words, the price of a CD accurately conveys the level of satisfaction received. This ignores the possibility of externalities or 'spillovers'. Think about someone hearing your CD and enjoying it - they are also receiving satisfaction from the disc, but the market is unable to provide any information about the benefits they are receiving unless they specifically share the cost of buying the CD. Whenever externalities occur, the market fails. Many economists believe that there are strong externality effects related to health care. For example caring for a sick person can impose financial costs on that person's family.

(b) Perfect Competition: An efficient free market requires producers to be operating under conditions of perfect competition. This requires a stringent set of conditions - perfect information, many buyers and sellers, a uniform product and freedom of entry and exit - which ensure that firms are price takers, producing for the lowest possible cost in the long run and only earning normal profits. If producers do not operate in this way and, in particular, if they have a significant power to influence price or the total quantity being produced, then the market will fail. Doctors and other suppliers of health care often have this power.

(C) Problems of Risk and Uncertainty: If we are going to buy health care in a free market, then we have to have enough money to pay for it. Nevertheless, health care is expensive and we cannot predict when we are going to be ill. What makes this worse is that postponing buying health care is often risky. So, we face the problems of risk and uncertainty. The market response to this problem is to develop an insurance market to remove the uncertainty and risk from health care spending. We pay an agreed amount of money per year whether we need health care or not. Then, when we need care, the insurer pays the bills, however large they are. So, a free market in health care requires an effective health care insurance market. Unfortunately, the health care insurance market itself is often not efficient. Moral hazard and adverse selection both cause significant market failure. Unequal information, moral hazard and adverse selection explain why

a free market in health insurance is unlikely to be efficient. However, health care markets face even more fundamental information problems.

(d) Consumers as Satisfaction Maximisers: Are consumers' rational satisfaction maximisers? Market theory assumes that consumers know what is best for themselves - that is they can make choices which will maximize their total satisfaction. If this assumption is wrong, then markets will not automatically produce efficient results. The satisfaction gained depends on the quantity and mix of goods and services chosen. The theory assumes that consumers get more satisfaction from more goods and services, but that the increase in satisfaction from consuming another unit - the marginal utility - diminishes as consumption rises. "By choosing a particular bundle of goods, people show that they prefer it to all others; thus, it is best for them. In addition, if all people are in their best position, then society - which is simply the aggregation of all people - is in its best position. Therefore, allowing people to choose in the marketplace results in the best of all possible economic worlds". Thomas Rice in the Economics of Health Reconsidered suggests a range of reasons why this view of consumer behaviour could be mistaken. These include:

(i) The idea that consumer utility depends on the bundle of goods and services consumed. If this were true then people in rich developed economies ought to be happier than people in poor economies. But, research by Easterlin (1974) argued that utility depended on relative consumption - so rich people were happier than poor people in all societies. This means that if you consume more, that could reduce my utility because I am now relatively worse off.

(ii) Traditional theory ignores the issue of how tastes are determined. Evidence from social psychology suggests that tastes are determined by people's past and present environments. So for instance, if you are in a peer group which smokes then you are likely to develop a 'taste' for smoking, which will remain, even after you have left the peer group. If this is true, then it is not clear that satisfying tastes will actually make people better off. In fact, "If one believes that tastes are determined in such a way, it becomes clear that a society might be better off pursuing some goods and services that are not demanded mostly by the public. This is because people may not know what alternatives are available that will make them better off".

(iii) Are consumers rational? What do economists mean by the concept of rationality? In a narrow sense, they mean that people will behave consistently - so if they prefer A to B and B to C then, they will prefer A to C. More widely, they mean that people will behave in a reasonable manner. If consumers are not rational in this sense, then they will not necessarily make decisions, which maximize their welfare. Social psychology suggests that people are often not rational in this sense - instead they exhibit what is called cognitive dissonance. In other words, they simultaneously hold two ideas that are psychologically inconsistent and use various forms of self-justification and rationalization to overcome the tension.

(e) Imperfect Competition: The free market models predict large numbers of buyers and sellers - all of whom have no power individually to influence the market price. However, a significant proportion of health care is delivered by hospitals and these hospitals can often exercise monopoly power within the health care market in the local area.

(f) Externalities: The economist defines external effects as involving positive and negative results for others that are the consequences of one's own actions. Externalities or spillover effects provide another source of market failure. Again the problem is related to information. This time the market price does not accurately contain all the information about the benefits and costs of the market transaction.

SELF ASSESSMENT EXERCISE

In theory, markets produce the goods and services we want in the right quantities and at the lowest possible cost but in the real world markets do not always work in the way theory predicts. Articulate reasons for this.

3.3 The Political Economy of Health Care

A normative statement was usually based on the efficiency criteria of welfare economics. This raised the issue of whether a Pareto-optimal design of a health care system might ever be achieved. This section raises the question of what determines the actual (rather than any desired) institutional structure of a health care system. This type of question is the topic of 'Political Economy', also known as 'Public Choice'. With regard to health policy and regulation, the following agents can be distinguished.

(i) Citizens: In a direct democracy, citizens may challenge a law that has been passed by a popular referendum. They may also force the legislature to deal with an issue through a popular initiative. In a purely representative democracy, voters have a mere indirect influence by voting for candidates for political office or parties who promise to pursue a certain policy. Even in a dictatorship, citizens are not without influence because at least some of them must be won over to keep public administration and the economy functioning. The more closely the health policy adopted by a dictatorial government matches the preferences of the citizenry, the less costly it is for it to maintain its power.

(ii) Politicians: In a democracy, politicians need to obtain votes. Promising to organize the provision of health care services (or at least the availability of health insurance) may be a selling proposition. Meanwhile, younger voters may think that these public programmes place a heavy financial burden on them while benefitting mainly the elderly.

(iii) Executive member of government: Gaining or maintaining executive power may calls for a great deal of financial support, which comes from large companies engaged in the health care sector (insurers, pharmaceutical companies) or professional associations (of physicians, nurses and hospitals). In general, the 'supply side' tends to prevail in health policy at the governmental level.

(iv) International organizations: The World Health Organization (WHO) has had considerable success in influencing national health policy by emphasizing the risks posed by epidemics. Increasingly, decisions affecting health policy and regulation are made by the World Trade Organization (WTO) and in particular the European Union (EU). In both instances, the fact that traded commodities may have an impact on health while some health goods are tradable provides a justification for intervention.

The focus here is on the viewpoint of citizens and voters. Their interests are decisive in countries with good governance because the other levels must take them into account in order to ensure their political survival

SELF ASSESSMENT EXERCISE

Write short note on the political-economy of health care

3.4 Health and Economic Development

Development is the concern of all developing countries. The health planner, manager, etc., is equally charged with that concern and must be knowledgeable about what development implies and the role health should play in the development of a given country. The following questions are of paramount importance for the health worker in a developing country: what is development? How does it differ from economic growth? How can development be measured? What role does health play in development? What role should the health worker play in facilitating development? This subsection provides some insights to these questions.

3.4.1 The Meaning of Economic Development

The modern view of development perceives it as both a physical reality and state of mind in which society has, through combination of social, economic and institutional processes, secured the means for obtaining a better life. The definition of "a better life" may vary from one society to another. Development in all societies must consist of at least the following three objectives:

(i) To increase the availability, distribution and accessibility of life-sustaining goods such as food, shelter, health, security and protection to all members of society;

(ii) To raise standards of living, including higher incomes, the provision of more jobs, better education and better health, and more attention to cultural and humanistic values so as to enhance not only material well-being, but also to generate greater individual community and national esteem.

(iii) To expand the range of economic and social opportunities and services to individuals and communities by freeing them from servitude and dependence on other people and communities and from ignorance and human misery.

Development and Economic growth were used interchangeably for a long time. Although the two are related, they are, however, different. Economic growth can be defined as an increase in a country's productive capacity, identifiable by a sustained rise in real national income over a period of years. The differences between growth and development can be outlined as follows:

(i) Development encompasses the total well-being of the individual, a community or a nation, while economic growth is concerned with the increase in per capita earnings of the people making up the nation.

(ii) Economic growth is one characteristic of development. It is possible for a country to experience economic growth without becoming developed. A country, for example, may acquire a great wealth from its mineral deposits, but have a low level of health services. This is due to the fact that the wealth goes into the hands of a very small minority who might squander it on luxury goods instead of establishing a viable infrastructure.

(iii) Development is concerned with the total person, his economic, social, political, physiological, psychic and environmental requirements. If one of these is not fully catered for, development has not been achieved.

3.4.2 Measurement of Economic Development

The measurement of development has presented social scientists with a problem of finding the suitable tools and techniques to do so and of interpreting the results of such measurements. Several suggestions have been presented for measuring development. One line of research has suggested the use of social indicators. The purpose of these is to measure the well-being of the population by examining factors such as health and nutritional status, level of education, housing conditions and so forth. However, it is easier to calculate GNP, per capita incomes and growth rates. As a result, in most reports these variables are used as indicators of development.

In addition to a rise in per-capita income, economic development implies fundamental changes in the structure of the economy characterized by:

(i) Rising share of industry, along with the failing share of agriculture in GNP and increasing percentage of people who live in cities rather than the rural areas or villages.

(ii) Changes in consumption patterns as people no longer spend all their income on necessities, but move on to consume durables and to leisure-time products and services.

(iii) Meeting the needs of the present without compromising the ability of future generations to meet their own needs (sustainability)

(iv) Participation by the citizens of the country in the process as well as the benefit, while economic development and modern economic growth involve much more than a rise in per capita income, there can be no development without economic growth.

3.4.3 Health Implications of Economic Development

The associations between health and national development are complex. The interaction is a two-way phenomenon with health being both influenced by and influencing economic development. Improved health has been considered solely a result of economic growth, a part of the product of growth rather than one of its causes. Some development experts have maintained that health should have low priority in development funding and have tried to justify their opinions with comments such as "only a rich nation can afford the programmes to assure its population's health", or "a poor nation cannot afford improved health". The concern of development planners is emphasized by the fact that during the demographic transition, lower death rates are often associated with sustained high birth rates which results in rapid population growth. While the supply of labour may increase as a result of improved health and reduced death rates, there may be no corresponding gain in per capita output. Thus, if economic growth is too slow to absorb the additions to the labour force associated with expanded health programmes, greater unemployment may result. Thus, improved health in poor societies can be postulated to produce larger populations, greater poverty and ultimately deterioration in health.

However, other development planners and economists are more optimistic regarding the impact of health and nutrition programmes on economic growth. There are three different ways by which improved health programmes can accelerate development.

- Improved health may increase productivity or efficiency of the labour force leading to greater output and reduced cost per unit of output.
- Better health conditions may serve to open new regions of a country for settlement and subsequent development.
- Attitudinal changes towards entrepreneurship may be linked to health and nutrition programmes. This linkage may stimulate entrepreneurship in poor countries.

It has been apparent that where conditions are worst, relatively simple and low cost health programmes can produce dramatic reductions of disability of the labour force. In these situations major increments in productivity are readily apparent. For instance, in the Philippines at one time a survey of major enterprises indicated a daily absenteeism rate of 35 percent, attributed largely to malaria. After initiation of an anti-malaria programme the rate of absenteeism was reduced to 2-4 percent and nearly one-fourth fewer labourers were required for any given task. Although one could argue that economic growth has to accelerate the eradication of poverty many economists felt that its impact occurred too slowly. In other words, many do not believe in an immediate trickle-down effect of economic growth. Subsequently, a more direct method of poverty reduction, namely the basic needs approach, was advocated. Its aim was the direct fulfillment of basic needs such as health, clothing, sanitation, shelter, nutrition and education.

3.4.4 Major Determinants of Poor Health

The main determinants of poor health which have direct or indirect interdependence with economic development include population growth, malnutrition, sanitary conditions and inadequate shelter and education. There remains a debate on the relation between health status improvements and economic growth. It is argued that health status improvements are attained at the expense of fixed capital entailing a smaller economic growth. That is, the investment funds that could have been used for the growth of the economy at large are to be used for investments in the health service sector which has in part a consumption character. Some argue, however, that investment in basic needs, such as in the health service sector, are investments in the health service sector which have in part a consumption character. Some argue, however, that investment in basic needs, such as in the health service sector, are investments in human capital which in turn is growth promoting. Although some tend to conclude that there is a positive relationship between health and economic development, this does not prove that improvement of the health service sector is a sufficient condition for economic development. It should be noted that, a better health status does not guarantee a faster economic growth. Therefore, the following conclusions may be drawn from the discussions of the relations between health and development:

(i) Development is not a simple process. It is a complex intermingling of economic, social, environmental, physiological, psychic, cultural and political factors.

(ii) The measurement of development is not an easy task. Economics provides certain tools which can be brought to bear on crucial areas of choice where decisions are required. Further research is required in this area so as to develop tools and techniques for evaluation in those areas that are not readily quantifiable.

SELF ASSESSMENT EXERCISE

Briefly discuss the relation between health and economic development.

4.0 CONCLUSION

In this unit it is clear that development is linked not just to the improvement of economic indicators or the attainment of basic needs, but with wider aspirations such as high health status, and with social well-being and change. The development process embraces not only the productive sectors of the economy, but also the social sectors.

5.0 SUMMARY

In this unit, we have discussed extensively the general characteristics of health care, health and economic development and the major determinants of poor health. Measurement of economic development and the health implication of economic development were also considered.

6.0 TUTOR-MARKED ASSIGNMENT

1. Outline the major determinants of poor health in a developing country.
2. Improved health can be considered as a precondition for economic development – how?
3. Discuss the policy implications of the WHO definition of health as a concept.

7.0 REFERENCES

Donaldson Cam and Karen Gerard (1993) Economics of Health Care Financing: The Visible Hand. Macmillan Press Ltd. London.

Folland S., A. Goodman & M. Stano (2010) The Economics of Health & Health Care, Sixth Edition, Prentice Hall, New Jersey.

Jones Andrew (2007) Applied Econometrics for Health Economists: A Practical Guide, 2nd Edition OHE

Santerre E. & S.P. Neun (1996) Health Economics: Theories, Insights & Industry Studies, Irwin, Chicago.

UNIT 4: Health and Medical Care: An Economic Perspective

CONTENTS

1.0 Introduction

Health is defined as "a state of physical mental and social well-being and the absence of disease or other abnormal condition." Economists take a radically different approach. They view health as a durable good, or a type of capital, that provides services. The flow of services produced from the stock of health "capital" is consumed continuously over an individual's lifetime (see Grossman, 1972a, 1972b). Each person is assumed to be endowed with a given stock of health at the beginning of a period, such as a year. Over the period, the stock of health depreciates with age and may be augmented by investments in medical services. Death occurs when an individual's stock of health falls below a critical minimum level.

Naturally, the initial stock of health, along with the rate of depreciation, varies from individual to individual and depends on many factors, some of which are uncontrollable. For example, a person has no control over the initial stock of health allowed at birth, and a child with a congenital heart problem begins life with a below-average stock of health. However, we learn later that medical services may compensate for many deficiencies, at least to some degree. The rate at which health depreciates also depends on many factors, such as the individual's age, physical makeup, lifestyle, environmental factors, and the amount of medical care consumed. For example, the rate at which health depreciates in a person diagnosed with high blood pressure is likely to depend on the amount of medical care consumed. (is this person under a doctor's care), environmental factors (does he have a stressful occupation?), and lifestyle (does the

person smoke or have a weight problem?). All these factors interact to determine the person's stock of health at any point in time, along with the pace at which it depreciates.

2.0 Objectives

After studying this unit, you should understand:

- utility analysis in health care services
- conceptual definition of medical care from economic perspective
- analysis of production of good health

3.0 Main Content

3.1 Why Good Health? Utility Analysis

Health, like any other durable goods, generates a flow of services. These services yield satisfaction, or what economists call **utility.** Your television set is another example of a durable good that generates a flow of services. It is the many hours of programming, or viewing services, that your television provides that yield utility, not the set itself. As a good, health is desired for consumption and investment purpose. From a consumption perspective, an individual desire to remain healthy because she receives utility from an overall improvement in the quality of life. In simple terms, a healthy person feels greet and thus is in a better position to enjoy life. The investment element concerns the relation between health and time. If you are in a positive state of health, you allocate less time to sickness and pursue other activities, such as leisure. Economists look at education from the same perspective. Much as a person invests in education to enhance the potential to command a higher wage, a person invests in health to increase the likelihood of having healthier days to work and generate income.

The investment element of health can be used to explain some of the lifestyle choices people make. A person who puts a high value on future events is more inclined to pursue a healthy lifestyle to increase the likelihood of enjoying healthier days than a person who put a low value on future events. A preference for the future explains why a middle-aged adult with high cholesterol orders a salad with dressing on the side instead of a steak served with a baked potato smothered in sour cream. In this situation, the utilities generated by increasing the likelihood of

having more healthy days in the future outweigh the utility received from consuming the steak dinner. In contrast, a person who puts a much lower value on future events and prefers immediate gratification may elect to order the steak dinner and ignore the potential ill effects of high cholesterol and fatty foods.

Naturally, each individual chooses to consume that combination of goods and services, including the services produced from the stock of health that provides the most utility. The isolated relation between an individual's stock of health and utility is captured in figure 1.3, where the quantity of health, H, is measured on the horizontal axis and the level of utility, U, is represented on the vertical axis. The positive slope of the curve indicates that an increase in a person's stock of health directly enhances total utility. The shape of the curve is particularly important because it illustrates the fundamental economic principle of the law of diminishing marginal utility. This law states that each successive incremental improvement in health generates smaller and smaller additions to total utility. in other words, utility increases at a decreasing rate with respect to health.

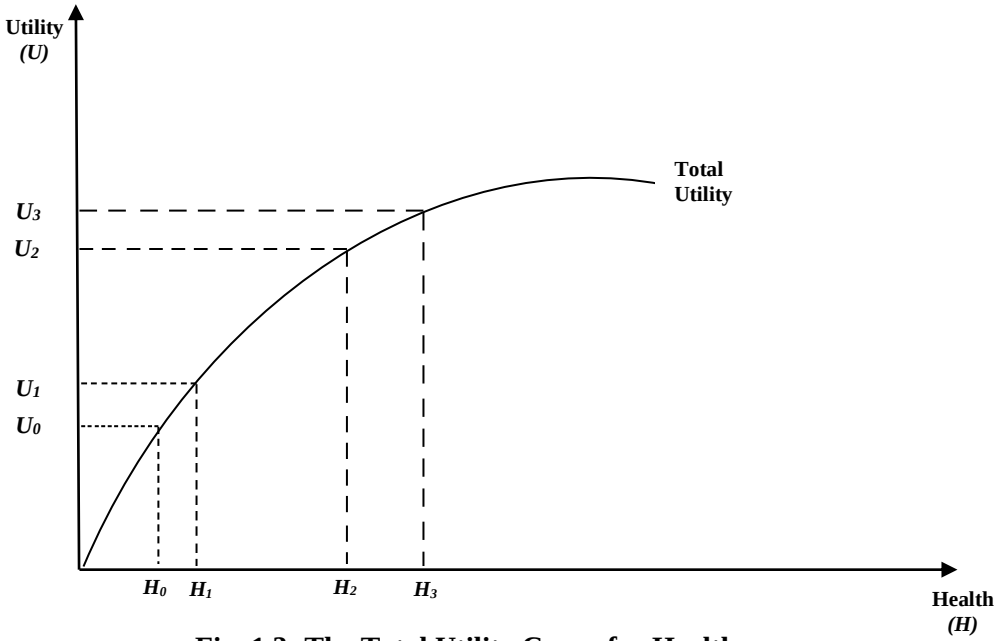


Fig. 1.3: The Total Utility Curve for Health

For example, in figure 1.3 an increase in health from H_0 to H_1 causes utility to increase from U_0 to U_1 while an equal increase in health from H_0 to H_1 generates a much smaller increase in utility, from U_0 to U_1. In the second case, the increase in utility is less when the stock of health is greater due to the law of diminishing marginal utility. The implication is that a person values a marginal improvement in health more when sick (i.e., when having a lower level of health) than when healthy. This does not mean that every individual derives the same level of utility from a given stock of health. It is possible for two more people to receive a different amount of utility from the same stock of health. The law of diminishing marginal utility requires only that the addition to total utility decreases with successive increases in health for a given individual.

Another way to illustrate the law of diminishing marginal utility is to focus on the marginal utility associated with each unit of health. Marginal utility equals the addition to total utility generated by each successive unit of health in mathematical terms,

$$MU_H = >U/>H. \qquad (1.4)$$

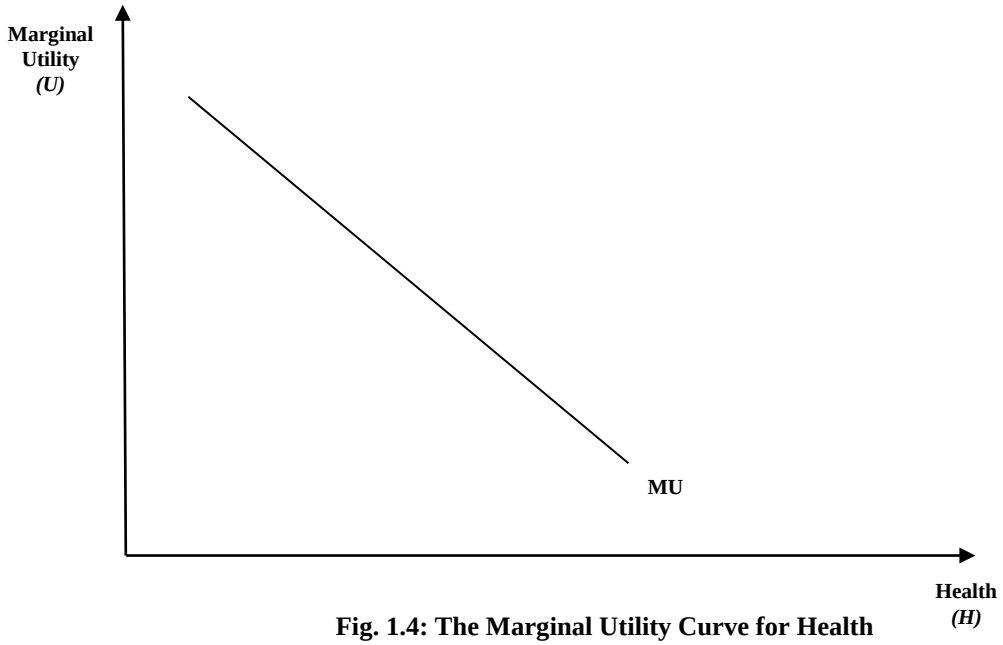


Fig. 1.4: The Marginal Utility Curve for Health

Where MU_H equals the marginal utility of the last unit of health consumed and > represents the change in utility of health in figure 1.4. Equation 1.4 represents the slope of a tangent line at

each point on the total utility curve. The bowed shape of the total utility curve implies that the slope of the tangent line falls as we move along the curve, or that MU_H falls as health increases. Figure 1.4 captures the relation between marginal utility and the stock of health. The downward slope of the curve indicates that the law of diminishing marginal utility holds because each new unit of health generates less additional utility than the previous one.

SELF ASSESSMENT EXERCISE
Briefly discuss the concept of utility as it applies to health care.

3.2 What is Medical Care?

Medical care is composed of myriad of goods and services that maintain, improve, or restore a person's physical or mental well-being. For example, a young adult might have shoulder surgery to repair a torn rotator cuff so that he can return to work; an elderly woman may have hip replacement surgery so that she can walk without pain, or a parent may bring a child to the hygienist for an annual cleaning of his teeth to prevent future dental problems. Prescription drugs, wheelchairs, and dentures are examples of medical goods, while surgeries, annual physical exams, and visits to physical therapists are examples of medical services. Because of the heterogeneous nature of nature care, units of medical care are difficult to measure precisely. Units of medical care are also hard to quantify because most represent services rather than tangible products. As a service, medical care exhibits four characteristics that distinguish it from a good: intangibility, inseparability, inventory, and inconsistency. The first characteristic, intangibility means that a medical service is incapable of being assessed by the five senses. Unlike a new car or a new CD, the consumer cannot see, taste, feel, or hear a medical service.

Inseparability means that the production and consumption of a medical service take place simultaneously. For example, when you visit your dentist for a checkup, you are consuming dental services at the exact time the dentist is producing them. In addition, a patient often acts as both producer and consumer. Without the patient's active participation, the medical product is likely to be poorly produced.

Inventory is directly related to inseparability. Because the production and consumption of a medical service occur simultaneously, health care providers are unable to stockpile or maintain an inventory of medical services. For example, a dentist cannot maintain an inventory of dental checkup to meet demand during peak period. Finally, inconsistency means that the composition and medical services consumed vary widely across medical events. Although everyone visits a physician at some time or another, not every visit to a physician is for the same reason. One person may go for a routine physical, while another may go because he needs heart bypass surgery. The composition of medical care provided or the intensity at which it is consumed can greatly differ among individuals and at different points in time.

The quality of medical care is also difficult to measure. Quality differences are reflected in the structure process, and/or outcome of a medical care. Structural quality is reflected in the physical and human resources of the medical care provider, such as the facilities (level of amenities), medical equipment (type and age), personnel (training and experience), and administration (organization structure). Process quality reflects the specific actions health care providers take on behalf of patients in delivering and following through with care. Process quality might include access (waiting time), data collection (background history and testing), and communication with the patient, diagnosis and treatment (type and appropriateness).

Outcome quality refers to the impact of care on the patient's health and welfare as measured by patient satisfaction, work time lost to disability, or post care mortality rate. Because it is extremely difficult to keep all three aspects of quality constant for every medical event, the quality of medical services, unlike that of physical goods, is likely to be inconsistent.

Medical care services are difficult to quantify. In most cases, researchers measure medical care in terms of availability or use. If medical care is measured on an availability basis, such measures as the number of physicians or hospital beds available per 1,000 people are used. If medical care is measured in terms of use, the analysis employs data indicating how often a medical service is delivered. For example, the quantity of office visits or surgeries per capita is often used to represent the amount of physician services rendered, whereas the number of

inpatient days is frequently used to measure the amount of hospital or nursing home services consumed.

SELF ASSESSMENT EXERCISE
Discuss the four characteristics that distinguish medical goods from other goods.

3.3 The Production of Good Health

Health economists take the view that the creation and maintenance of health involves a production process. Much as a firm uses various inputs, such as capital and labour, to manufacture a product, an individual uses medical inputs and other factors, such as healthy lifestyle, to produce health. The relation between medical inputs and output can be captured in what economists call a production function. A health production function indicates the maximum amount of health that an individual can generate from a specific set of inputs in a given period of time. In mathematical terms it shows how the level of output (in this case, health) depends on the quantities of various inputs, such as medical care. A generalized short-run health production function for an individual takes the following form:

Health = H (medical care, technology, profile, lifestyle, socio-economic environment) (1.5)

Where health reflects the level of health at a point in time; medical care equals the quantity of medical care consumed; technology refers to the state of medical technology at a given point in time; profile captures the individual's mental and physical profile as of a point in time; lifestyle represents a set of lifestyle variables, such as diet and exercise; socioeconomic status reflects the joint effect of social and economic factors, such as education, income and poverty; and environment stands for a variety of environmental factors, including air and water quality. To focus on the relation between health and medical care, we assume initially that all other factors in the health production function remain constant. Figure 1.5 depicts this relation, where q is a hypothetical measure of health care, holding technology constant, and H represents the level of health. The intercept term represents the individual's level of health when zero medical care is consumed. As drawn, the total product curve implies that an individual's level of health is positively related to the amount of medical care consumed.

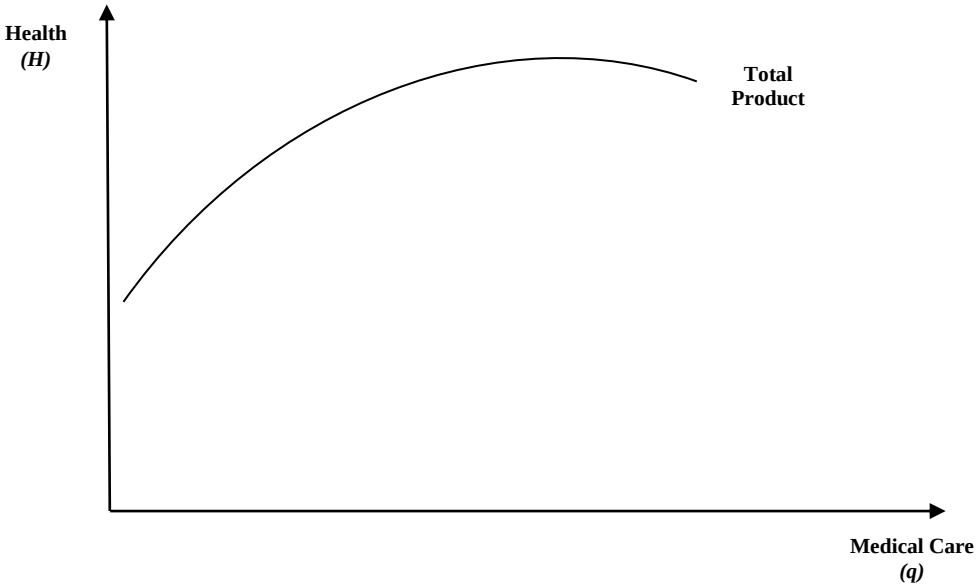


Fig. 1.5: The Total Product Curve for Medical Care

The shape of the curve is very similar to that in figure 1.4 and reflects the law of diminishing marginal productivity. This law implies that health increases at a decreasing rate with respect to additional amounts of medical care, holding other inputs constant. For example, suppose an individual makes an initial visit and several follow-up visits to a physician's office for a specific illness or treatment over a given period of time. It is very likely that the first few visits have a more beneficial impact on the individual's stock of health than the later visit. Thus, each successive visit generates a smaller improvement in health than the previous one.

$$MPq = >H/>q \qquad (1.6)$$

Where MPq equals the marginal product of the last unit of medical care services consumed. The law of diminishing marginal productivity holds that the marginal product of medical care diminishes as the individual acquires more medical care. A graph of this relationship appears as a negatively sloped curve in figure 1.6. The other variables in the health production function can also be incorporated into the analysis. In general terms, a change in any one of the other variables in the production function alters the position of the TP curve. The TP curve may shift in some instances and/or rotate in others. In the latter case, the curve rotates because the marginal productivity of medical care has changed in response to the change in the other factors.

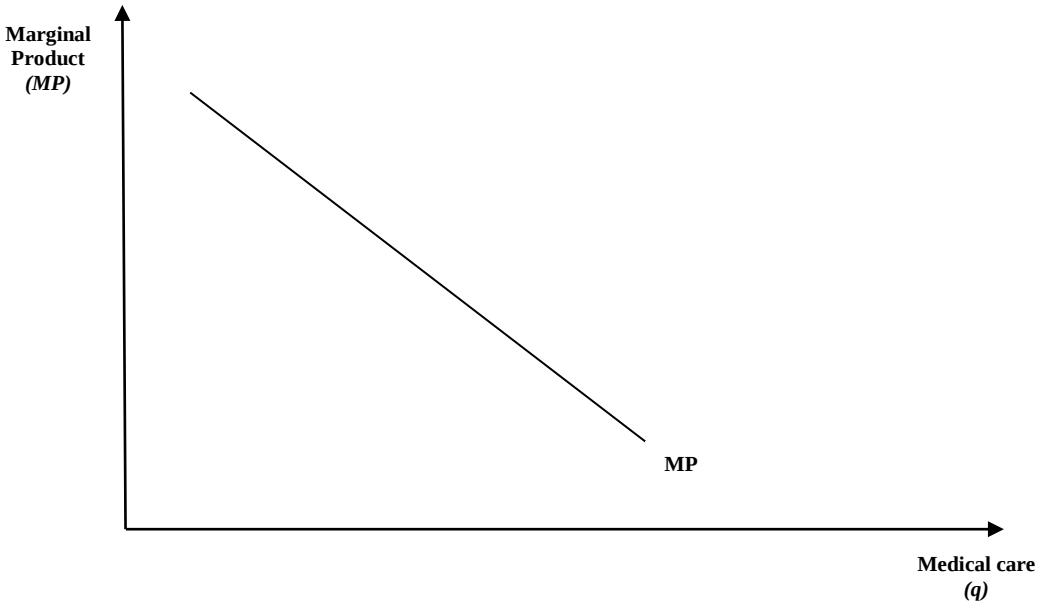


Fig. 1.6: The Marginal Product Curve for Medical Care

New medical technologies have affected all aspects of the production of medical care. In the broadest of terms, examples of new technologies include the development of sophisticated medical devices, the introduction of new drugs, the application of innovative medical and surgical procedures, and most recently, the use of computer-supported information systems, just to name a few. Technological change can result in treatment expansion, treatment substitution, or some elements of both. Treatment expansion occurs when more patients are treated by a new medical intuition which occurs when the new technology substitutes for or replaces an older one.

In the context of our health production model, the development and application of a new medical technology causes the total product curve to pivot, and rotate upward because the marginal productivity of each unit of medical care consumed increases, as illustrated in figure 1.7. Notice that the total product curve rotates upward from TP_0 to TP_1 and each unit of medical care consumed now generates a greater amount of health. The movement from point A to point B in figure 1.7 illustrates the case in which the improvement in medical technology brings about an increase in the amount of medical care consumed from q_0 to q_1 along with an improvement in health from H_0 to H_1. This movement represents the treatment expansion resulting from the new

medical technology. Movement represents the treatment expansion resulting from the new medical technology. Movement from point A to C illustrates the situation in which the new technology has no impact on health but results in less consumption of medical care from q0 to q2. In this case, the new technology is cost saving, all things being equal. It should be noted that the increase in the marginal product of medical care brought about by the medical technology also causes the marginal product curve to shift to the right.

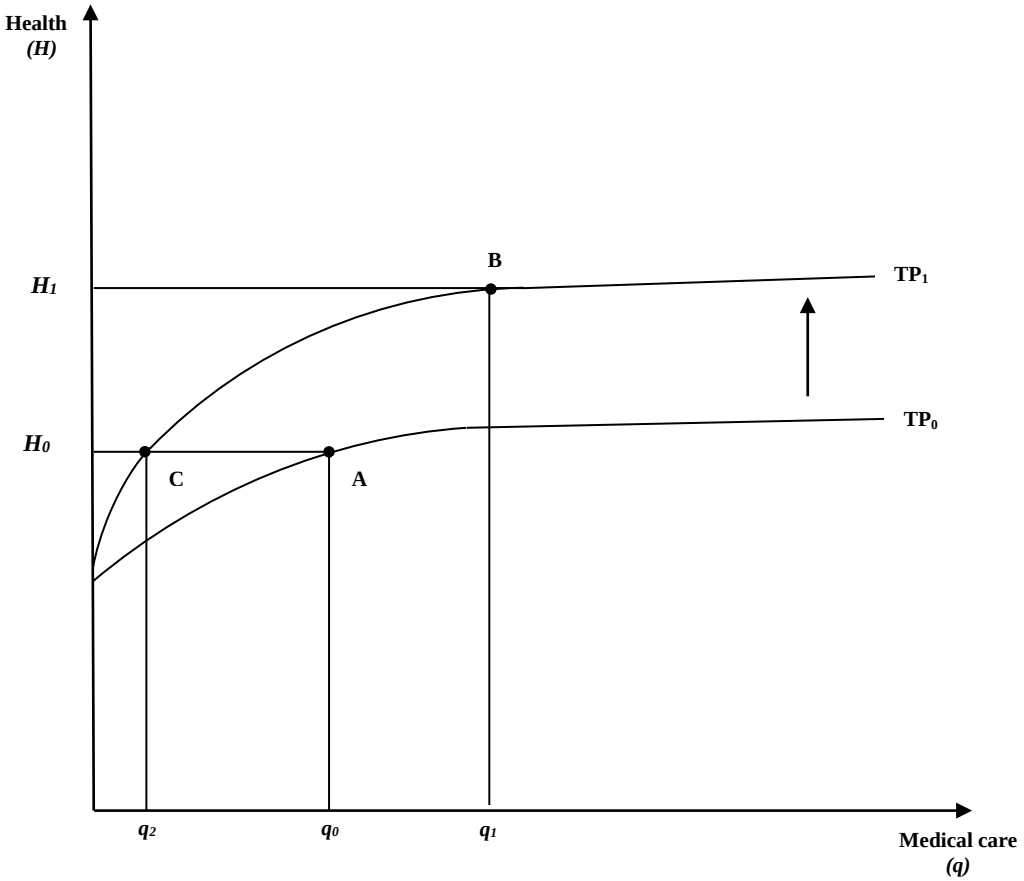


Fig. 1.7: The Effect of Technological Change on the Total Product Curve for Medical Care

SELF ASSESSMENT EXERCISE

The increase in marginal product of medical care brought about by the medical technology causes the marginal product curve to shift to the right. Discuss.

4.0 Conclusion

Regardless of how you define it, health is a vague concept that defies precise measurement. In terms of measurement, heath depends as much on the quality of life which has become an

increasingly important issue in recent years due to the life-sustaining capabilities of today's medical technology. Because the quality of life is a relative concept that is open to wide interpretation, researchers have wrestled with developing an instrument that accurately measures health. Health economists often use the inverse of mortality or morbidity rates as a measure of heath.

5.0 Summary

This unit looked at the utility concept in health care and conceptual definition of medical care in economics. It also looks at the issue of production of health and assumed that an individual is bestowed with a stock of health when born and death occurs when this stock of health reaches a particular critical stage.

6.0 Tutor-Marked Assignment

Evidence supporting the notion that exercise is good for the brain as well as the body continues to surface. The science behind the evidence illustrates that exercise can boost a person's ability to process data, reduce depression and anxiety, and reduce illness that can affect a person's mental functioning, such as Alzheimer's. A 2005 study of 884,715 fifth-, seventh-, and ninth-graders published in the journal of exercise physiology shows that students in the best physical shape have the best test scores. But exercise cannot replace intellectual exertion; it can only strengthen it. Evidence also points to the long-term benefits of exercise and indicates that exercise can make the brain act younger. The results of a study at the University of Illinois show that when seniors exercise, it produces patterns of brain activity usually seen in 20-year-olds.

Questions

1. Health can be thought of as a personal asset. We invest in health by consuming medical care and other health-related inputs, such as exercise. Discuss how the results of these studies are likely to impact the role of exercise in the "production" of health.
2. We know two things for sure in public health: smoking is bad and exercise is good. But we have had trouble stopping the bad behavior and encouraging the good. Smoking is addictive, which adds to the complexities of figuring out how to stop it. Why has it been so difficult to encourage people to exercise?

7.0 References/Further Reading

Folland S., A. Goodman & M. Stano (2010) The Economics of Health & Health Care, Sixth Edition, Prentice Hall, New Jersey.

Jacobs, P. (1991) The Economics of Health and Medical Care Maryland: Aspen Pub Inc. Jack, Williams (1964) Principles of Health Economics for Developing Countries. WBI Development Studies. The World Bank, Washington D. C.

Santerre E. & S.P. Neun (1996) Health Economics: Theories, Insights & Industry Studies, Irwin,

Chicago.
Zweifel P., F. Breyer & M. Kifmann (2009) Health Economics, Second Edition, Springer Verlag Heidelberg.

MODULE TWO: THE DEMAND AND SUPPLY OF MEDICAL CARE

Unit 1: The Demand for Medical Care
Unit 2: The Supply of Physician Services and other Medical Services
Unit 3: Medical Care Production and Costs
Unit 4: Hospital Services and Efficiency

UNIT 1: The Demand for Medical Care

CONTENTS

1.0 Introduction

The stock of health can be treated as a durable good that generates utility and is subject to the law of diminishing marginal utility. This means that each incremental improvement in health generates successively smaller additions to total utility. Medical services are an input in the production of health because a person consumes medical care services for the purpose of maintaining, restoring, or improving health. However, the law of diminishing marginal productivity causes the marginal improvement to health brought by each additional unit of medical care consumed to decrease.

From this discussion, it follows that medical care indirectly provides utility. Specifically, medical care helps to produce health which in turn generates utility. Consequently, utility can be

specified as a function of the quantity of medical care. The shape of the total utility curve indicates that utility increases at a decreasing rate with respect to medical care, or that medical care services are subject to diminishing marginal utility. Marginal utility decreases because each successive unit of medical care generates a smaller improvement in health than the previous unit (due to the law of diminishing marginal productivity) and each increase in health, in turn, generates a smaller increase in utility (due to the law of diminishing marginal utility). This unit deals with the analysis of demand for medical care using the traditional demand analysis.

2.0 Objectives

After studying this unit, you should be able to:

(i) Understand the utility-maximising rule

(ii) Understand the demand for medical care services.

(iii) Understand elasticity of demand for medical care services

3.0 MAIN CONTENT

3.1 The Utility-Maximizing Rule

Given market prices at a point in time, consumers must decide which combination of goods and services, including medical care, to purchase with their fixed income. According to microeconomic theory, each consumer chooses the bundle of goods and services that maximizes utility. Consumer utility is maximized when the marginal utility gained from the last naira spent on each product is equal across goods and services purchased. This condition is known as the utility maximizing rule, and it basically states that total utility reaches its peak when the consumer receives the maximum "bang for the buck" in terms of the marginal utility per naira of income from each and every good. In mathematical terms, the rule states that utility is maximized when

$$MU_x/P_x = MU_y/P_y \qquad (2.1)$$

Where: MU_x represents the marginal utility received from the last unit of medical care purchased, x, and MU_y equal the marginal utility derived from the last unit of all other goods, y.

The latter good is often referred to as a composite good in economics. To illustrate why the utility maximizing rule must hold, suppose that

$$MU_x/P_x > MU_y/P_y \qquad (2.2)$$

In this case, the last naira spent on medical care generates more additional utility than the last naira spent on all other goods. The consumer can increase total utility by reallocating expenditures and purchasing more units of medical care and fewer units of all other goods. As the consumer purchases more medical services at the expense of all other good (remember that the consumer's income and composite good's price are fixed) the marginal utility of medical care falls and the marginal utility of other goods increases. This in turn, causes the value of MU_x/P_x to fall and the value of MU_y/P_y to increase. The consumer purchases additional medical services until the equality in equation (2.1) again holds, or the last naira spent on each product generates the same additional satisfaction. At this point, total utility is maximized and any further changes in spending patterns will negatively affect total utility. The relationship between utility and medical care is shown in Figure 2.1 below.

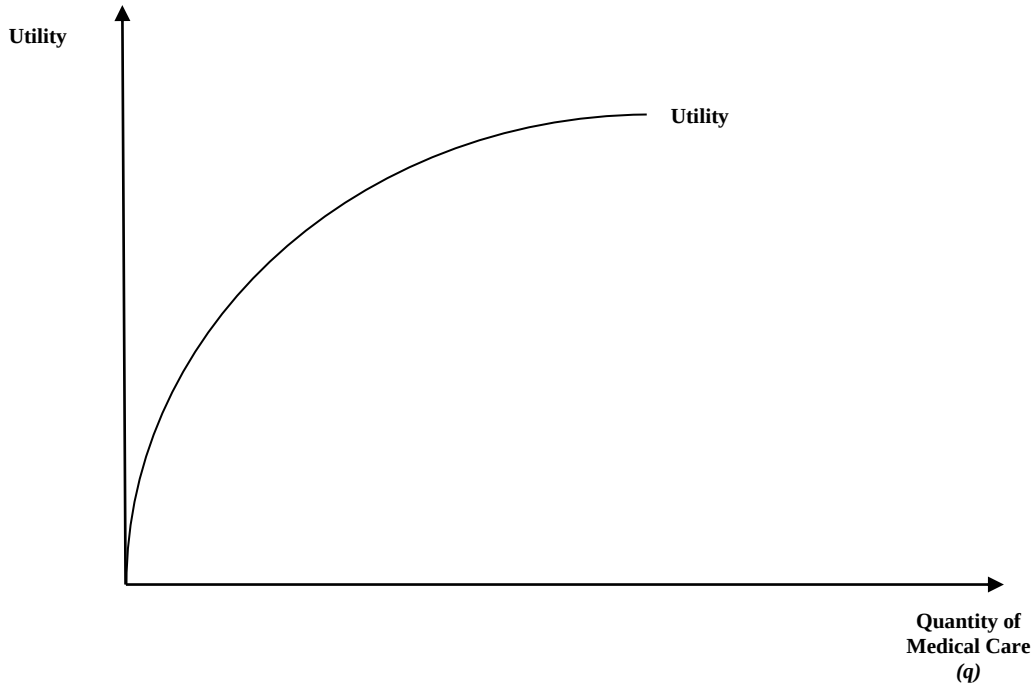


Fig. 2.1: The Relationship between Utility and Medical Care

3.1.1 The Law of Demand

The equilibrium condition specified in equation (2.1) can be used to trace the demand curve for a particular medical service, such as physician services. For simplicity, assume the prices of all other goods and income remain constant and initially the consumer is purchasing the optimal mix of physician services and all other goods. Now assume the price of physician services increases. In this case, MU_x/P_x is less than MU_y/P_y (where MU_x and P_x represent the marginal utility and price of physician services respectively). Consequently, the consumer receives more satisfaction per naira from consuming all other goods. In reaction to the price increase, the consumer purchases fewer units of physician services and more units of all other goods. This reallocation continues until MU_x/P_x increases and MU_y/P_y decreases and the equilibrium condition of equation (2.1) is again in force such that the naira spent on each good generates an equal amount of utility. Thus, an inverse relation exists between the price and the quantity demanded of physician services.

If the price of physician services continually changes, we can determine a number of points representing the relation between the price and the quantity demanded of physician services. Using this information, we can draw the demand curve as shown in figure 2.2, where the horizontal axis indicates the amount of physician services consumed (as measured by the number of visit) and the vertical axis equals the price of physician services. The curve is downward sloping and reflects the inverse relation between the price and the quantity demanded of physician services, ceteris paribus. For example, if the price of physician services equals P_0 the consumer is willing and able to purchase q_0. Notice that if the price falls to P_1, the consumer purchases q_1 amount of physician services. In this case, price represents the physician out-of-pocket expense the consumer incurs when purchasing medical services from a physician. As such, it equals the amount the consumer must pay after the impact of third-party payments has been taken into account. Naturally, if the visit to the physician is not covered by a third party, the actual price of the visit equals the out-of-pocket expense.

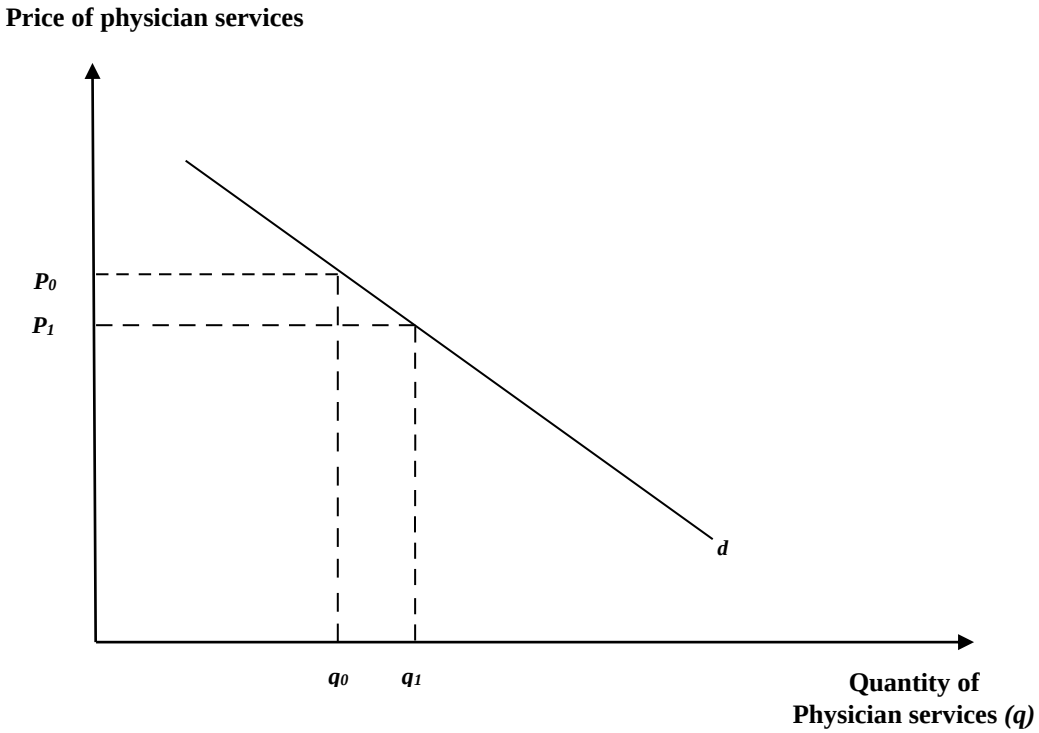


Fig. 2.2: The Individual Demand Curve for Physician Services

The substitution and income effects associated with a price change offer another theoretical justification of the inverse relationship between price and quantity demanded. Both of these effects predict that a higher price will lead to a smaller quantity demanded and, conversely, a lower price will result in a greater quantity demanded. According to the substitution effect, a decrease in the price of physician services causes the consumer to substitute away from the relatively higher-priced medical goods, such as hospital outpatient services, and purchasing more physician services. That is, lower-priced services are substituted for higher-priced ones. As a result, the quantity demanded of physician services increases as price decreases. According to the income effect, a lower price also increases the real purchasing power of the consumer. Because medical care is assumed to be a normal good (that is, the quantity demanded of medical services increases with income), the quantity demanded of physician services increase with the rise in purchasing power. That also generates an inverse relation between price and quantity demanded because as price falls, real income increases and quantity demanded rises. Taken

together, the substitution and income effects indicate that the quantity demanded of physician services decreases as price increases.

In summary, figure 2.2 captures the inverse relationship between the price the consumer pays for medical care (e.g. physician services) and the quantity demanded. The curve represents the amount of medical care the consumer is willing and able to purchase at every price. Utility analysis, or the income and substitution effects, can be used to generate this relationship. This inverse relationship is sometimes referred to as the law of demand. It is important to note that the demand for medical care is a derived demand, because it depends on the demand for good health. A visit to a dentist illustrates this point. An individual receives no utility directly from having a cavity filled. Rather, utility is generated from an improvement in dental health. Other economic and non-economic variables also influence the demand for health care. Unlike price, which causes a movement along the demand curve, other factors influence quantity demanded by altering the position of the demand curve.

3.1.2 Other Economic Demand-Side Factors

Income is another economic variable that affects the demand for medical services. Because medical care is generally assumed to be a normal good, any increase in income, which represents an increase in purchasing power, should cause the demand for medical services to rise. Figure 2.3 illustrate what happens to the demand for physician services when income increases. The increase in income causes the demand curve to shift to the right, from d_0 to d_1, because at each price the consumer is willing and able to purchase more physician services. Similarly, for each quantity of medical services, the consumer is willing to pay a higher price. This is attributable to the fact that at least some portion of the increase in income is spent on physician services. Equally, a decrease in income causes the demand curve to shift to the left.

The demand for a specific type of medical service is also likely to depend on the prices of other goods, particularly other types of medical services. If two or more goods are jointly used for consumption purpose, economists say that they are complements in consumption: because the goods are consumed together, an increase in the price of one good inversely influences the

demand for the other. For example, the demand for eyewear (i.e. glasses) and the services of an optometrist are likely to be complementary.

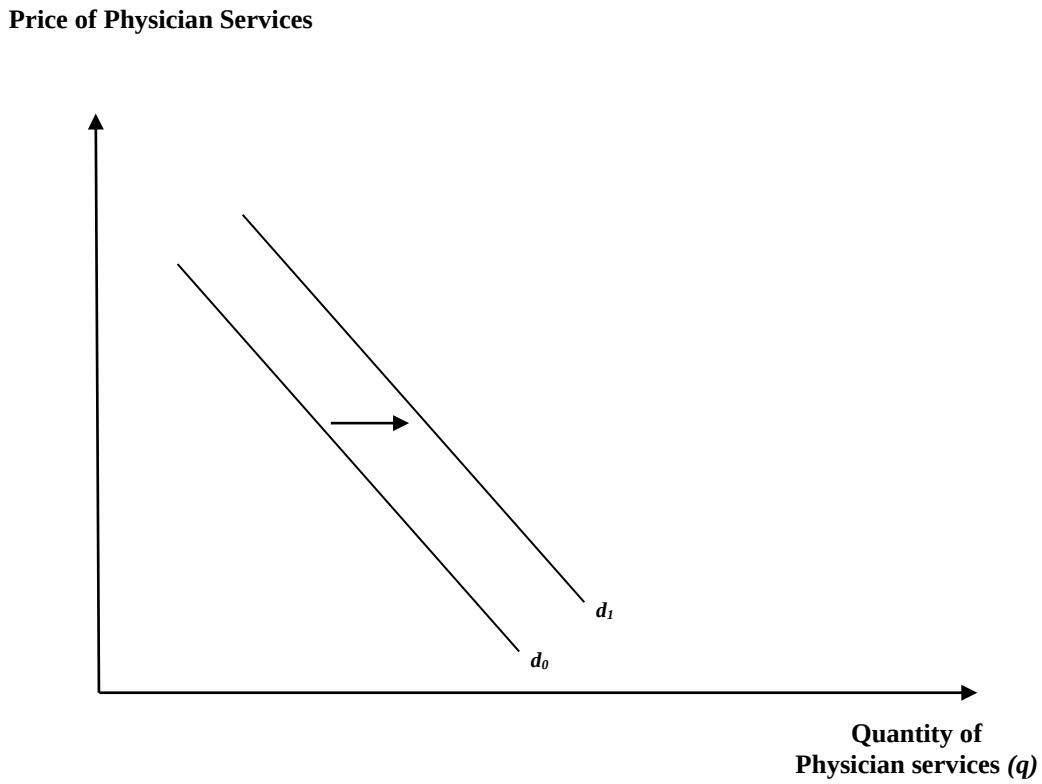


Fig. 2.3: Shift in the Individual Demand Curve for Physician Services

Normally, an individual has an eye examination before purchasing eyewear. If these two goods are complements in consumption, the demand for optometric services should increase in response to a drop in the price of eyewear. As a result, the demand curve for optometric services shifts to the right. Another example of a complementary relation exists between obstetric and pediatric services. An increase in the price of pediatric services should inversely influence the demand for obstetric services. If, for example, a woman postpones pregnancy because of the high cost of pediatric services, her demand for obstetric services also falls. The demand curve for obstetric services shifts to the left.

It is also possible for two or more goods to satisfy the same wants or provide the same characteristics. If that is the case, economists say that these goods are substitutes in

consumption. The demand for one good is directly related to a change in the price of a substitute good. For example, suppose physician services and hospital outpatient services are substitutes in consumption. As the price of outpatient services increases, the consumer is likely to change consumption patterns and purchase more physician services because the price of a visit to the doctor is cheaper in relative terms. That causes the demand curve for physician services to shift to the right. Generic and brand-name drugs provide another example of two substitute goods. The demand for brand-name drugs should decrease with a decline in the price of generic drugs. If so, the demand curve for brand-name drugs shifts to the left. Finally, eyeglasses and contact lenses are likely to be substitute in consumption.

Time costs also influence the quantity demanded of medical services. Time costs include the monetary cost of travel, such as bus fare or gasoline, plus the opportunity cost time. The opportunity cost of an individual's time represents the naira value of the activities the person forgoes when acquiring medical services. For example, if a plumber who earns ₦50 an hour takes two hours off from work to visit a dentist, the opportunity cost of the time equals ₦100. The implication is that the opportunity cost of time is directly related to a person's wage rate. Given time costs, it is not surprising that children and elderly people often fill doctors' waiting rooms. Time costs can accrue while traveling to and from a medical provider, waiting to see the provider, and experiencing delays in securing an appointment. In other words, travel costs increase the farther an individual has to travel to see a physician, the longer the wait at the doctor's office, and the longer the delay in getting an appointment. It stands to reason that the demand for medical care falls as time costs increase (i.e. as the demand curve shift to the left).

3.1.3 Non-Economic Determinants of the Demand for Medical Care

Four general non-economic factors influence the demand services: taste and preferences, physical and mental profile, state of health, and quality of care. Taste and preference include personal characteristics such as marital status, education, and lifestyle, which might affect how people value their healthy time (i.e. their marginal utility of health) or might lead to a greater preference for certain types of medical services. Marital status is likely to impact the demand for health care in the market primarily through its effect on the production of health care in the

home. A married individual may demand less medical care, particularly hospital care, because of the availability of a spouse to care for him, such as when recuperating from an illness.

The impact of education on the demand for medical care is difficult to predict. On the one hand, a consumer with additional education may be more willing to seek medical care to slow down the rate of health depreciation because that consumer may have a better understanding of the potential impact of medical care on health. As an example, an individual with a high level of education may be more inclined to visit a dentist for periodic examination. Thus, we should observe a direct relation between educational attainment and demand. On the other hand, an individual with a high level of education may make more efficient use of home-produced health care services to slow down the rate of health depreciation and, as a result, demand fewer medical care services. For example, such an individual may be more likely to understand the value of preventive medicine (such as proper diet and exercise). In addition, the individual may be more likely to recognize the early warning signs of illness and be more apt to visit a health care when symptoms first occur. As a result, health care problems are addressed early when treatment has a greater probability of success and is less costly. That means that we should observe an inverse relation between the level of education and demand for medical care, particularly acute care. Finally, lifestyle variables, such as whether the individual smokes cigarettes or drinks alcohol in excessive amounts, affect health status and consequently the amount of health care demanded. For example, a person may try to compensate for the detrimental health impact of smoking by consuming more health care services. That translates into an increased demand for medical care.
The profile variable considers the impact of such factors as gender, race/ethnicity and age on the demand for medical services. For example, females generally demand more health care services than males primarily because of childbearing. In addition, certain diseases, such as cardiovascular disease, osteoporosis, immunologic diseases (such as thyroid disease and rheumatoid arthritis), mental disorders, and Alzheimer's disease, are more prevalent in women than men. Age also pays a vital role in determining the demand for medical care. As stated earlier, as an individual age, the overall stock of health depreciates more rapidly. To compensate for this loss in health, the demand for medical care is likely to increase with age; at least beyond

the middle years (the demand curve shifts to the right). Thus, we should observe a direct relation between age and the demand for medical care.

State of health controls for the fact that sick people demand more medical services, everything else held constant. As you might expect, health status and the demand for health care are also likely to be directly related to the severity of illness. For example, a person who is born with a medical problem is likely to have a much higher than average demand for medical care. In economics jargon, an individual who is endowed with less health is likely to demand more medical care in an attempt to augment the overall stock of health. Finally, quality of care is also likely to impact the demand for medical care. Because quality cannot be measured directly, it is usually assumed to be positively related to the amount and types of inputs used to produce medical care. Feldstein (1967, pp. 158-62) defined the quality of care as "a catch-all term to denote the general level of amenities to patients as well as additional expenditures on professional staff and equipment". For example, a consumer may feel that larger hospitals provide better-quality care than smaller ones because they have more specialists on staff along with more sophisticated equipment. Or, that same individual may think that physicians who have graduated from prestigious medical schools provide a higher quality of care than those who have not. It matters little whether the difference in the quality of care provided is real or illusory. What matters is that the consumer perceives that differences in quality actually exist. As Feldstein's definition indicates, quality can depend on things that have little to do with the actual production of effective medical care. For example, the consumer may prefer a physician who has a pleasant office with a comfortable waiting room along with courteous nurses. Thus, any increase in the quality of care provided is likely to increase consumer's demand for medical care regardless of whether it affects the actual production of health care.

We must also distinguish between a movement along the demand curve and a shift of the curve. A change in the price of medical services generates a change in the quantity demanded, and this is represented by a movement along the demand curve. If any of the other factor changes, such as income or time costs, the demand curves for medical services shifts. This shift is referred to

as a change in demand. Thus, a change in the quantity demanded is illustrated by a movement along the demand curve, while a change in demand is illustrated by a shift of the curve.

In summary, the variable we expect to influence an individual's demand for medical care in economic theory indicates that the demand equation should look something like the following:

Quantity demanded = *f* (out-of-pocket price, income, time costs, prices substitutes and complements, taste and preferences, profile, state of health, and quality of care) (2.3)

Equation (2.3) states that the quantity demanded of medical services depends on the general factors listed. Note that a change in the first factor results in a movement along a given demand curve, whereas an adjustment in the other factors produce a shift of the demand curve. A rightward shift indicates a greater demand and a leftward shift reveals a lower demand.

SELF ASSESSMENT EXERCISE

The demand for medical care is a derived demand, discuss.

3.2 The Market Demand for Medical Care

The market demand for medical care, such as physician services, equals the total demand by all consumers in a given market. In graphical terms, we can construct the market demand curve for medical care services by horizontally summing the individual demand curves. This curve represents the amount of medical services that the entire market is willing and able to purchase at every given price. For example, if the average price of a visit to a doctor is ₦50 and at this price consumer A is willing to see a physician three times over the course of a year while consumer B is willing to make four visits, the total, or market demand for physician services is seven visits per year at ₦50 per visit. The market demand curve is downward sloping for the same reasons the individual demand curves are downward sloping. In addition, the factors that shift the individual demand curves also shift the overall market demand curve, provided the changes take place on a market wide basis. The market demand curve also shifts if the overall number of consumers in the market increases or decreases. For example, the demand for medical

care in a particular community may increase if an influx of new residents occurs. This causes the market demand curve to shift to the right.

The development of a market curve allows us to distinguish between the intensive and extensive margins. The intensive margin refers to how much more or less of a product consumers buy when its price changes. The extensive margin captures how many more or fewer people buy a product when its price changes. Obviously, this is an important distinction to make for a product like medical care. Many medical purchases such as surgeries happen only once for a particular individual. As another example, an individual can have a particular tooth pulled only once. This is also a one-shot purchase that either happens or does not happen. If the price of tooth extraction falls, however, we may still observe an inverse relationship between the price and number of teeth extracted. That is, at the extensive margin, more consumers elect to purchase this one-time form of dental services as price falls. Thus, quantity demanded may increase with a reduction in price because of changes that occur at the intensive and extensive margins.

SELF ASSESSMENT EXERCISE
Provide a graphical illustration of the market demand for physician services.

3.3 The Fuzzy Demand Curve

We have assumed so far that the market demand curve for medical care is a well-defined line, implying a precise relation between price and quantity demanded. In reality this is usually not the case, and we need to refer to the derivation of the demand curve for medical care to see why. Recall that the demand for medical care is a derived demand and depends on the demand for health and the extent to which medical care influences the production of health. The relation between medical care and health, however, is far from exact. That is because there is a considerable lack of medical knowledge concerning the efficacy of certain types of medical interventions. As a result, health care providers disagree about the treatment of some certain types of medical problems, and the demand for medical service becomes fuzzy. For example, there is debate among physicians concerning when surgery is necessary for elderly males with prostate cancer. Also, in some instances consumers may lack the information or medical knowledge they need to make informed choices. Consequently, consumers tend to rely heavily

on the advice of their physicians when making such decisions as when a particular medical test or surgery is necessary. The implication is that physicians rather than consumers choose medical services, which makes the demand curve fuzzier. Further complicating matters is the inability to measure medical care produced during a one-hour therapy session with a psychiatrist.

All these factors combined make it extremely difficult to accurately describe the relation between the price and the quantity demanded of medical care. In other words, the relation between price and quantity demanded is rather fuzzy. A more accurate depiction of the relation between price and quantity may not be a well-defined line but a gray band similar to the one depicted in figure 2.4.

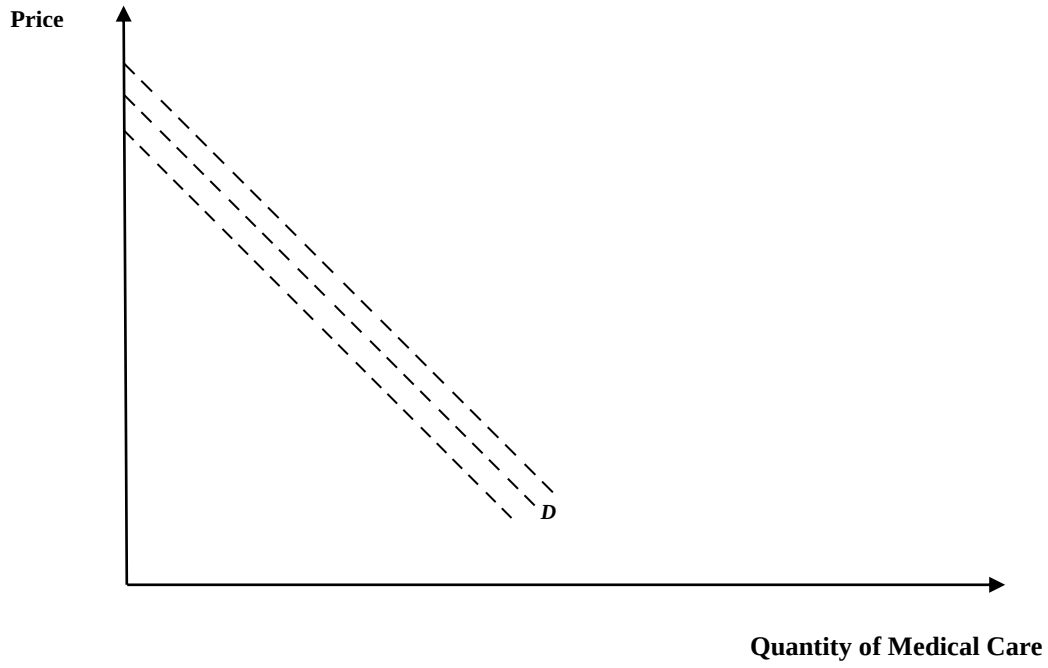


Fig. 2.4: The Fuzzy Demand Curve for Medical Care

There are implications associated with the fuzzy demand curve. First, for a given price, we may observe some variation in the quantity or types of medical services rendered. Researchers have documented variations in physician practice styles across geographical areas. Secondly, for a given quantity or type of medical service, we are likely to witness price differences. Feldstein (1988) reported a substantial variation in physician fees in the same geographical area.

However, the existence of the band is unlikely to detract from the inverse relation between the price and the quantity demanded for medical care.

SELF ASSESSMENT EXERCISE
Discuss factors that make it difficult to accurately describe the relation between the price and the quantity demanded of medical care.

3.4. Elasticity

Economic theory gives us insight into the factors that influence the demand for medical care along with the direction of their influence. For example, we know that if the price of physician services increases by 15 percent, the quantity demanded falls. But by how much does it fall? Is there any way to determine whether the decrease is substantial or negligible? The answer is yes; with the help of measure economists call elasticity. Elasticity measures the responsiveness of quantity demanded to a change in an independent factor.

3.4.1 Own-Price Elasticity of Demand

The most common elasticity is the own-price elasticity of demand. This measure estimates the extent to which consumers change their consumption of a good or service when its own price changes. The formula for elasticity is:

$$E_D = \%\Delta Q_D / \%\Delta P \qquad (2.4)$$

Where: E_D denotes the price elasticity of demand, $\Delta\%\ Q_D$ represents the percentage change in quantity demanded, and $\%\Delta P$ is the percentage change in price. From the formula, E_D is a simple ratio that equals the percentage change in quantity demanded divided by the percentage change in price. Because elasticity is specified as a ratio of two percentage changes, it is scale free. This makes it much easier to compare elasticity across different goods. For example, we can compare the price elasticity of demand for physician services with that for nursing home care and not have to concern ourselves with the fact that the demand for physician service is usually measured in terms of the number of visits while the demand for nursing home care is measured in terms of the number of inpatient days.

The value of E_D is negative and reflects the inverse relationship between price and quantity demanded. In economics, the normal practice is to take the absolute value of the price elasticity of demand measure, or $/E_D/$, and eliminate the minus sign. If the price elasticity of demand is greater than 1 in absolute terms ($/E_D/>1$), the demand for the product is referred to as price elastic. In arithmetic terms, $/E_D/ > 1$ if the absolute value of the percentage change in price is smaller than the absolute value of the change in the quantity demanded, or $/\%\Delta P/ < /\%\Delta Q_D/$. For example, if the price elasticity of demand for dental services equals 1.2, this means the quantity consumed falls by 12 percent if the price of dental care increases by 10 percent, ceteris paribus.

The price elasticity of demand is referred to as inelastic if $/E_D/ < 1$ but greater than zero. In this case, $/\%\Delta P/ < /\%\Delta Q_D/$, or the percentage change in price is greater than the percentage in quantity demanded in absolute value terms. For example, if the elasticity of demand for physician services equals 0.6, a 10 percent decrease in price leads to a 6 percent increase in quantity demanded. If $/E_D/$ equal to 1 because $/\%\Delta P/$ equals $/\%\Delta Q_D/$, the price elasticity of demand is unit elastic. This implies that a 10 percent decrease in the price of the product leads to a 10 percent increase in the quantity demanded.

A demand curve that is vertical is said to be perfectly inelastic because no change occurs in the quantity demanded when the price changes. In mathematical terms, E_D equals zero because $\%\Delta Q_D$ equals zero. At the other extreme, if the demand curve is horizontal, it is referred to as being perfectly elastic and $/E_D/$ equals infinity (∞). Any change in price leads to an infinite change in the quantity demanded. It stands to reason that the more elastic the demand for the product, the greater the response of quantity to a given change in price. Compare the effects of a 10 percent decrease in price of two goods-one with a price elasticity of -0.1 and another with a price elasticity of -2.6. In the first case, the quantity demanded increases by only 1 percent, while in the second case, it increases by 26 percent. We can also use the elasticity of demand to make inferences regarding the slope of the demand curve. Generally, the more elastic the demand for a product, the flatter the demand curve at any given price. This also means that the curve is relatively steep at any given point for an inelastic demand. Consider the two linear

demand curves that intersect at point P_o, Q_0 in figure 2.5. If the price of the product increases to P_1, the quantity demanded decreases to Q_a of the flat curve.

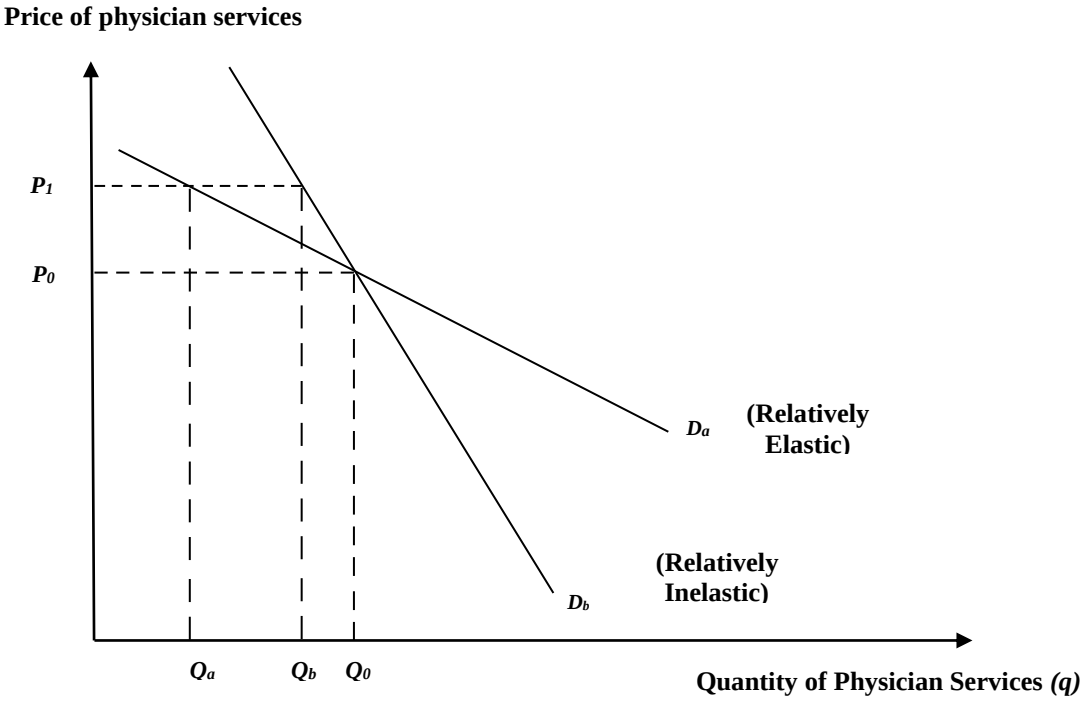


Fig. 2.5: The Elasticity of Demand and the Slope of the Demand Curve

The own-price elasticity of demand varies across products, and economists point to several factors that determine its value. Among the factors often mentioned are the portion of the consumer's budget allocated to the good, the amount of time involved in the purchasing decision, the extent to which the good is a necessity and the availability of substitutes. As the portion of a consumer's budget allocated to a good increase, the consumer may become more sensitive to price change. Demand therefore becomes more elastic. An increase in the decision-making time frame may also make demand more elastic. If the consumer has more time to make informed choices, he or she is likely to react more strongly to price changes. Because the consumer typically pays a small portion of the cost of medical services because of insurance, and because medical services are sometimes of an urgent nature, these two considerations suggest that in many cases, the demand for medical services is inelastic with respect to price.

If a good is a necessity, such as a basic foodstuff, the own-price elasticity should be relatively inelastic. The product is purchased with little regard for price because it is needed. Basic phone service might be considered another example of a necessity. Because our society depends so heavily on the phone as a form of communication, it is difficult to imagine a household functioning effectively without one. Naturally, basic health care falls into the same category. If an individual needs a particular media service, such as an operation or a drug, and if not having it greatly affects the quality of life, we can expect that person's demand to be inelastic with respect to price. In addition, when a person needs a particular medical service in a life-or-death situation, demand is likely to be perfectly inelastic because the medical service must be purchased regardless of price if the person has sufficient income.

Given that many medical services are necessities; we expect the overall demand for medical services to be somewhat inelastic. But this does not mean the amount of health care demanded does not react to change in price. Rather, it means the amount of change in price generates a small percentage change in the quantity demanded of medical services. For some types of medical care, however, demand may be more elastic. Elective medical care such as cosmetic surgery may fall into this category, because in most instances it is considered a luxury rather than a necessity. As a result, price may play an important role in the decision to have the surgery. To a lesser degree, dentist services and eyewear might fall into this category. In fact, any medical service that can be postponed is likely to display some degree of price elasticity.

The availability of substitutes is another determinant of price elasticity. As mentioned earlier, various types of medical services may serve as substitutes for one another. The larger the number of substitutes, the greater the opportunity to do some comparison shopping. As a result, the quantity demanded of any medical service is likely to be more sensitive to price changes when alternative means of acquiring medical care are available. The own-price elasticity of demand for any given product should be directly related to the number of substitutes available. Stated another way, demand should become more price elastic as the number of substitute expands. One implication is that the demand for an individual medical care provider is likely to

be more elastic than the market demand for medical care. One more point to note concerning elasticity is that the own-price elasticity of demand can be used to predict what happens to total health expenditures if price increases or decreases. Total revenues (or total expenditures, from the consumer's perspective) equal price times quantity. In mathematical notation,

$$TR = P*Q_D \qquad (2.5)$$

Where: TR represents total revenue. Demand theory tells us that as the price of a product increase, the quantity demanded decreases, or that P and Q_D move in opposite directions. Whether total revenue increases or decreases when the price changes is dictated by the relative rates at which both variables change, and the elasticity of demand. Consider an increase in the price of physician service where demand is inelastic. This means that $/\%\Delta Q_D/ < /\%\Delta P/$, or that the percentage increase in price is larger than the percentage decrease in quantity demanded in absolute value terms. In terms of equation 2.5, P increases faster than Q_D falls. This means total revenue must increase with a higher price. If demand happens to be elastic, the opposite occurs: the quantity demanded falls faster than the price increases, and as a result, total revenue decreases. No change occurs in total revenue when demand is unit elastic because the increase in price is matched by the same percentage decrease in quantity demanded.

3.4.2 Other Types of Elasticity

The concept of elasticity can be used to measure the sensitivity of quantity demanded to other demand side factors as well. The income elasticity of demand represents the percentage change in quantity demanded divided by the percentage change in income, or $E_Y = \%\Delta Q_D/\%\Delta Y$, where $\%\Delta Y$ equals the percentage change in income. It quantifies the extent to which the demand for a product changes when real income changes. If E_Y is positive, the good is referred to as a normal good because any increase in income leads to an increase in quantity demanded. For example, if E_Y equals 0.78, this means a 10 percent increase in income causes the quantity consumed to increase by 7.8 percent. An inferior good is one for which E_Y is negative and an increase in income leads to a decrease in the amount consumed. For most types of medical care, the income elasticity of demand should be larger than zero.

The cross-price elasticity (E_C) measures the extent to which the demand for a product changes when the price of another good is changed. In mathematical terms, $E_C = \%\Delta Q_X / \% \Delta P_Z$, where the numerator represents the percentage change in the demand for good X and the denominator equals the percentage change in the price of good Z. if E_C is negative, we can infer that the two goods are complements in consumption. The cross-prices elasticity between the demand for optometric services and price of eyewear should be negative. If the price of eyewear increases, the demand for optometric service should drop. Two goods are substitutes in consumption when the cross-price elasticity is positive. For example, the cross-price elasticity of the demand for physician services with respect to the price of outpatient service may turn out to be positive. If E_C equals zero, the demand for the product is independent of the price of the other product.

SELF ASSESSMENT EXERCISE

Discuss the implications of a price decrease on total revenue when demand is elastic, inelastic, and unit elastic.

4.0 Conclusion

The demand for medical care obeys the law of demand. From the utility theory, the stock of health is a durable good that generates utility and is subject to the law of diminishing marginal utility. This means that each incremental improvement in health generates successively smaller additions to total utility. Medical services are also an input in the production of health because a person consumes medical care services for the purpose of maintaining, restoring, or improving health. But, the law of diminishing marginal productivity causes the marginal improvement to health brought by each additional unit of medical care consumed to decrease. Hence, medical care provides utility.

5.0 Summary

This unit looked at the demand analysis of medical care. It discussed different factors that affect medical care demand and elasticity of demand for medical care. The unit also looks at the implication of elasticity of demand of medical care on total revenue. The unit concluded that medical care demand analysis can be analysed using the traditional demand analysis in economics because medical care generates utility.

6.0 Tutor-Marked Assignment

1. Moral hazard in health care usually refers to the additional consumption of medical care by the insured relative to the uninsured. That is, as the price falls to zero, people consume more services. Should all of the additional "insurance-related" consumption waste? Why or why not?
2. To what extent can the strategy- excepting certain kinds of primary care from deductibles and copayment-be adopted for other kinds of health care? Are there "optimal" copayments for all medical procedures and services, one that can discriminate between welfare-enhancing versus welfare-reducing moral hazard?

7.0 References/Further Reading

Culyer J.A. & J.P. Newhouse (2000) Eds, Handbook of Health Economics: Vols 1A & 1B, Elsivier, North-Holland.

Folland S., A. Goodman & M. Stano (2010) The Economics of Health & Health Care, Sixth Edition, Prentice Hall, New Jersey.

Santerre E. & S.P. Neun (1996) Health Economics: Theories, Insights & Industry Studies, Irwin, Chicago.

Zweifel P., F. Breyer & M. Kifmann (2009) Health Economics, Second Edition, Springer Verlag Heidelberg.

Unit 2: The Supply of Physician Services and other Medical Services

CONTENTS

1.0 INTRODUCTION

An implicit assumption for the study of the determinants of demand for medical services and the uncertainty surrounding health needs was that individuals had well-defined and well-informed preferences for health and health care, and that they made their consumption and risk-reduction decisions rationally. But this assumption, which is basic to most economic analysis, must be relaxed to some degree in the analysis of the supply of medical services, because it is clear that as well as providing operational services (that is, injections, surgery, and the like), one of the primary roles of a medical care worker is the provision of information that affects the demand for services. The implications of this connection for the level and quality of care can be significant and lead us to question the efficiency of a market-determined allocation of resources in the health sector. This topic considered the motivation and behaviour of physicians and other health sector workers, the institutions within which they operate, and the resource allocation outcomes that alternative financial structures and other incentives might yield. Of particular concern is the efficiency of health service production, the appropriateness of the services, and the allocation of physician and other labour to rural and urban areas. The analysis is essentially one of the equilibrium between supply and demand. However, the special position of medical care providers in relation to their patients means that demand may be in some sense a function of the behaviour of suppliers. The interest is particularly in the extent to which physicians can induce consumers to purchase more medical care than they would if they were fully informed

about its effects, and the impact such strategies might have on market responses to supply shocks and the effectiveness of government price interventions.

2.0 OBJECTIVES

After studying this unit, you should be able to:

(i) Understand the analysis of inputs into the production of health

(ii) Understand the short-run production analysis of health care services.

(iii) Understand elasticity of the long-run production analysis of health care services.

3.0 MAIN CONTENTS

3.1 Inputs into the Production of Health Care

3.1.1 Physicians

A defining characteristic of the demand side of the health care market is the lack of information that individuals have about the cause, nature, and treatment of disease. Unlike the demand for food, clothing, and other standard consumption commodities, individuals often have poorly defined preferences over health care services. Given this essential feature of demand, physicians of all kinds and specialties play two distinct roles as health care providers. First, they provide information and advice to patients on the nature of their condition; the likely impacts of particular treatments, both positive and negative, and their recommended course of action. In addition to these services, physicians engage in the physical delivery services, including surgery, administering of injections, writing of drug prescriptions, and so forth. If individuals were fully aware of the effects of various treatments in improving their health, they would not require the first kind of service (the provision of information) and physicians would be just like bakers and barbers. In these situations, economists find it useful to think of the service provider as the "agent" of the consumer. Of particular reliance is that the agent-the physician or other health care provider in our case-has more extensive information about the consequences and costs of his or her actions than does the patient (who is known in the literature as the "principal").

Two important points regarding this description of roles of physicians should be noted. Firstly, it is possible, at least in principle, to imagine some physicians providing just advisory services,

and other physicians providing the executive or operational services, in which case the dual roles would be separated to some degree. This could have important effects on the pattern and cost of services delivered. Secondly, while they are usually better informed than their patients, physicians are unlikely to have full information about the consequences of their actions.

3.1.2 Other Medical Personnel

Physicians make up only a fraction of the personnel resources used in the health care sector. Also included are nurses, administrators, clerks, receptionists, traditional healers, and general staff. Some of the labour services provided by such individuals are substitutes for each other and for the work of physicians. For example, trained nurses can administer injections and oral drugs, monitor patients, and so forth. Receptionists could probably manage the taking of a temperature, but their productivity in doing so would likely be much less than that of a nurse, because they may make mistakes. Of course, it is not the absolute productivity in performing a task that should determine the allocation of tasks among individuals, but the relative productivity, or comparative advantage. Thus, even though the physician may be a better typist than the receptionist, it is economically efficient for the receptionist to type up prescriptions and let the physician concentrate on diagnosing and treating patients.

One important substitute for physician services is lower level medical services (provided by nurses or other clinicians, for example) delivered at the appropriate time. It may be possible to substitute the use of a doctor's labour at a time when a patient has developed a severe illness for the use of a clinician's time at a much earlier stage in the patient's life, for example. This is essentially an argument about the possibility of preventive care reducing the need for future curative care. Notice that the desirability of this substitution depends on the relative costs of the two types of care, the cost imposed on the patient (in ill health and the like) and other social costs. Most analysts argue that the returns from such a substitution are substantial in sub-Saharan Africa, but some studies in more advanced economies suggest that intensive preventive care, such as screening for certain types of cancer, may or may not be desirable.

3.1.3 Non-labour Inputs

Medical supplies, particularly drugs, instruments, and capital equipment are essential inputs into the production of health services. There may be some degree of substitutability between labour and non-labour inputs, such as consultations with other physicians that might reduce the quantity of drugs required for a given patient, or the use of additional secretariat services in place of office equipment. But many drugs do not have close non-drug substitutes and represent complements to physician care rather than substitute. Absence of these inputs can constrain the productive capacity of the medical services. It is of concern in some developing countries that the quantity of inputs is sometimes very difficult to detect. This is particularly the case with drugs, which often have a non-descript appearance in tablet form. If labeling can be changed at low cost, the scope for fraud is wide and high-quality drugs may be unavailable, except in the black market. This can have two effects: first, treatment may not be provided, and second, treatments with poor-quality drugs may lead to serious unintended effects. In addition, drugs are most effective when prescribed in the correct dosages and in conjunction with suitable complementary treatments and actions (such as abstinence from drinking alcohol).

While capital goods are important inputs in hospitals, they tend to make up a much smaller share of cost in lower-level outlets, such as clinics. As well as the financial costs involved in procurement of the equipment, resources must be spent to keep them in good working order. These costs are essentially depreciation costs and are incurred with all machines that do not last forever. The problem is that the rate of depreciation can be a function of the uses of the equipment, and inappropriate use can increase the rate at which the stock of physical capital becomes ineffective. This, coupled with the possibility that local workers might be untrained in the maintenance of the machines, can impose additional real-resource costs on health care service organization. Other non-labour inputs include:

(i) Drugs

Despite the scope for substitution among inputs, modern drug therapies play an essential role in the treatment of many diseases, and they are second only to personnel costs. Their share of recurrent costs in African economies represents complementary inputs to physicians' visits in health production as is evidenced by large positive correlations between the supply of drugs and

the demand for health facility visits. By far the greatest component of the cost of production of most drugs is incurred at the research and development stage, and the marginal costs of production are often close to zero, except when transportation and storage costs are significant. This characteristic of the production process has at least two important implications. First, to give drug companies incentives to incur the large investment costs required, new drugs can be patented, allowing the producer to exercise monopoly power. These patents typically last for a limited amount of time, and when they expire, the drugs can be produced (at low marginal cost) and sold by other companies.

Competition then forces prices down, and the drugs become much less profitable. At this stage, drugs with expired patents become known as "generic" drugs. They may have similar properties to other patented (and more expensive) drugs in their effectiveness in combating disease, but they are usually much cheaper.

The existence of products that appear to be different, but are actually similar, means that information problems at the consumer level, which are significant in the market for drugs, might be exacerbated. While one might expect that in such situations uniformed individuals could choose cheap but inappropriate drugs, it is also possible that they will infer that there are additional benefit associated with the more expensive varieties.

Some accounts suggest that the preference of consumers for specialty drugs over their generic equivalents can be extremely costly.

(ii) Hospitals

Hospitals combine a large number of inputs and treat a wide variety of conditions, ranging from the mundane to the exotic. To be able to make sensible policy decisions regarding the allocation of resources to and within hospital, it is necessary to have a model (explicit or otherwise) of how these institutions function and means of explicitly measuring their performance. The standard theory of the firm is not used to us here. On the one hand, the goals and decision making processes of hospitals are not necessarily well described by the neoclassical model, because describing the organization of a hospital in economic terms can be difficult when orthodox notions of ownership and control are ill-defined, and the objectives of those in control are only

vaguely characterized. On the other hand, measuring a hospital's performance-in productivity terms, for example, is notoriously difficult, given the myriad services they offer.

3.1.4 Hospital Motives and Behaviour

A striking feature of many non-government hospitals in some countries is that they deem themselves "not for profit" organization. Exactly what such status means for the behaviour of decision makers within these hospitals is not always clear. One would hope, for example that such a status did not mean that resources were wasted, so that any potential profit would be removed by inefficiencies. This hardly seems a reasonable goal for any organization to promote, nor for a government to foster. The best way to understand the nature of a not-for-profit hospital is, first, to consider its alternative-a profit-maximizing hospital. As with any other productive firm, the profits from such an enterprise measure the economic surplus from production (the value of outputs sold over inputs used) that accrues to the owners of the firm-the shareholders. Under the assumption that the owners of the hospital control to hire, the suppliers from whom to purchase inputs, the potential patients to target in advertising campaigns, the surplus represents a return to their capital investment and entrepreneurial action. The profits are typically distributed to the shareholders on a regular basis, such as at the end of each year.

Now suppose that for some reason a particular group of individuals involved in the activities of the hospital happen to be the owners. For example, the resident physicians and attending specialists might be the sole shareholders, or the administrators could be in this position. Alternatively, the patients who are served by the hospital might own it. In each case, the distribution of profits could be effected through an explicit financial transfer. A similar transfer could be implemented by altering the prices at which the owning group trades with the hospital. For example, physicians' salaries might be increased by an amount that compensates them both for their labour services as physicians and for the capital (if any) they have invested in the hospital plus their entrepreneurial services. Alternatively, the transfer could be made in kind, with physician acquiring various perks (such as membership in a hospital-funded social club) that do not show up as earned income. In either case, the hospital would show no accounting profit, but the entire surplus would accrue to the physicians.

3.1.5 Measuring the Performance of Hospitals

When estimating total output of a hospital, we need to aggregate the goods that the hospital produces-the treatments of different types of illnesses and conditions-in a meaningful fashion. The usual way to do this is by weighting the outputs of each good by its average or, preferably, marginal cost. With this one-dimensional measure of output, we can ask questions about the existence of increasing, decreasing, or constant returns to scale. Hospitals are usually assumed to have increasing returns over some range, because they require certain capital equipment and building. As Phelps (1992) has pointed out, however, one necessary input into a hospital's production process is the patient, and if patients must travel a long distance to reach the hospital, total marginal costs may start to increase at some level. This may particularly be important in developing countries with poor transportation infrastructure. In addition, the uni-dimensional output measure allows the estimation of improvements over time in productivity, while accounting for possible changes in the mix of patient conditions.

A substantial impediment to accurate cost and productivity estimation arises from the difficulties of measuring the quality of health care. Health status upon discharged patients is a very continuously from zero to one, so that the number of discharged patients is a very poor measure of the actual amount of health improvement generated by the hospital. A second dimension of quality is associated with the costs borne by consumers in using the hospital facilities. Transport costs were mentioned above, but these costs also include waiting times, the pleasantness of the surroundings, the friendliness or compassion of medical staff, and the like (the last two are negative costs to the consumer). If these costs are not included in the estimation of hospital costs, measures of productivity will be biased. In addition to admission rates, other practical measures of the output of hospitals include the number of bed-days, costs of inputs used, and other input measures. The implicit assumption behind all of these measures is that there is a stable and monotonically increasing relationship between inputs and output, and if the first increases, so does the second. But increases in input levels coupled with inefficient use or poor management can easily lead to less than proportional output increases, or even to reductions in

output. Thus, a hospital that increases the length of stay of patients may be providing more comprehensive treatment, and thus better-quality health care, in which case an inference of higher output would be valid. But longer hospital stays may also be the result of poor organization and staffing, and patients may spend most of their time waiting for physicians and supplies to turn up.

SELF ASSESSMENT EXERCISE
Discuss various inputs into the production of health care

3.2 Incentives and the Allocation of Resources

Given the technical relationships between inputs and outputs described in general terms in the preceding section, what patterns of resources allocation are we likely to see in the health care sector? Such resource allocations can be effected either through direct administrative procedures (for example, by governments) or as a consequence of decisions made by private individuals in the delivery and management of health care services. They rely on the choices and behaviour of physicians and other providers of labour and physical inputs within institutional settings to implement their preferred resource allocations. This section examines the incentive of agents on the supply side of the health sector, and the implications of different financial and other structures in the determination of resource use.

It is useful to identify two margins on which the allocation of resources may be inefficient. First, incentives may induce agents on the supply side to produce too much or too little care of a particular kind, and second, a given level of care or output may be produced inefficiently-with the wrong mix of input, for example, deviation from efficiency on the first of these margins represents allocative inefficiencies, while those on the second represent X-inefficiencies. At a conceptual level, it is not always easy to separate these two notions of efficiency. Given a production function, allocative efficiency requires that the right quantity be produced, as defined in some way on the basis of the estimated social benefits and costs. However, if production is not undertaken efficiently, the “right” level of output may change, and allocative efficiency would be defined in some constrained second-best fashion.

3.2.1 Physicians' Objectives and Behaviour

It is common in economics to describe resource allocations as the outcome of optimizing behaviour of economic agents who have explicit objectives and capacities and who face given financial (and other) reward structures. Within this general framework, physicians can be thought of as both supplying orthodox labour services and acting as entrepreneurs, primarily because of their privileged possession of information about the health production process and the needs of their patients. As the suppliers of labour, their behaviour can be analyzed in a similar fashion to that of other workers as resulting from a tradeoff between consumption of goods purchased in the market and leisure. As medical entrepreneurs, the objectives that determine their behaviour must be expanded to include such dimensions as effort on the job, prestige and reputation, the well-being of their patients (both generally and specifically with regard to their health), and ethical considerations.

3.2.2 Investing in Human Capital - the Decision to become a Doctor.

Faced with market - or government - determined compensation (that is, wage rates), individuals make decisions on whether to become doctors, the type of specialty to undertake, and their hours of work. Like any investment decision, the decision to undertake a typically long-term commitment to study must be made with the opportunity costs in mind. These include forgone current income and any direct financial costs. Of potentially more significance, however, is that, unlike many other investment choices, investment in human capital may be difficult to finance through capital markets. That is, when ownership of individuals is not permitted, a person who invests in human capital through education has no associated physical capital to offer a lender as collateral. Faced with such capital market imperfections, only those with existing assets (that is the well-off) will be able to attain costly education.

SELF ASSESSMENT EXERCISE

Discuss factors most considered when investing in human capital.

3.3 Labour Supply

Having received their training, physicians must make decisions about the amount of work they intend to do and where they will do it. The hour's decision is captured conceptually through the standard income-leisure tradeoff as a function of net wages. As net wages increase, the price of leisure effectively increases, and the substitution effect induces individuals to supply more labour. In opposition to this, higher wages increase each individual's wealth, and the income effect induces the individual to consume more leisure (reduce the labour supply). It is assumed that the first effect dominates, at least over the relevant range, and that the labour supply curve slopes up. Thus, one way to encourage doctors to supply more services is to pay them more.

The location decision presents policymakers with one of the more challenging dilemmas in health care provision. Because of the greater availability of both private and public goods and services in urban area, compensation payments to physicians must usually be correspondingly greater in rural areas to persuade them to locate outside major cities. Although money prices of some goods are likely to be greater in the cities, the effective prices of other may well be high in rural areas where the particular item is unavailable. In addition, if incomes are lower in rural areas, the supportable prices that physicians can charge will be lower, so some form of public subsidy is likely to be necessary. In addition to the rural-urban decision, physicians might have the option of locating outside their home country, moving to neighbouring countries where their incomes could be higher. This incentive can work in the opposite direction, with doctors trained in western countries sacrificing large salaries to work in developing countries with international agencies. In the presence of significant net out-migration of state-trained physicians, the government may examine the net social return to publicly funded medical education.

3.3.1 Moonlighting

Moonlighting is the practice of workers (not necessarily physicians) taking on a second job in addition to their primary employment. Typically, the second job will be in the informal sector of the economy, characterized by a lack of formal administrative structure (Labour laws, unions, and so on) and incomplete tax coverage. There are positive and normative issues that may be of concern to a policy analyst with respect to moonlighting. On the one hand, it is useful to

understand the forces that lead to multiple job holdings; on the other hand, it is important to know the welfare implications of such activities. Should they be controlled, and if so, how? To understand why physicians might engage in both formal and informal sector labour supply, we can begin by describing a scenario in which such multiple job holdings would not be desirable to the individual. Suppose that wages in both sectors are fixed, that the workers can work as much as they wish at the going wage (that is, the labour market is competitive), and that the disutility of work is the same in both sectors. In this case, the optimal labour supply choice is to work in the sector with the highest net wage, and moonlighting would not be observed. However, the conditions imposed in this example suggest situations in which multiple job holdings would be desirable.

First, if the worker cannot choose hours of work freely, then he may face a constraint in one job that makes the second job attractive. For example, suppose the formal sector wage is higher than that in the informal sector, but that no overtime work is permitted (or at least remunerated) in the formal sector, the worker will optimally work as much as possible in the higher-paid job but may choose to work additional hours in the lower-paid informal sector. Second, the worker may be concerned about the stability of his income. If wages in the informal sector are higher but more volatile than those in the formal sector, he may wish to provide himself with insurance by working in both sectors. Note that full insurance could be obtained by working only in the low-paid formal sector, but that this may not be optimal if average wages in the informal sector are high enough or if the individual is not too risk-averse.

A third reason to have jobs in both sectors is that there might be private economies of scope in labour supply. That is, working in one sector may increase the individual's productivity in the other sector. The most likely cases are when formal sector employment increases informal sector returns. For example, some kind of training may be provided in the formal sector job that increases both formal and informal sector productivity. Alternatively, attributes of the formal sector job could be used as inputs in the informal sector, such as office equipment and stationery, professional status and reputation, and access to patient (who are treated outside of

office hours). Some of these attributes, such as status, human capital, and reputation, have a public good aspect, and their use in the informal sector represents a social economy of scope. Others, however, such as use of formal sector inputs, permit the worker to achieve private economies of scope without any social equivalent. That is, the second activities tend to reduce productivity in the formal sector, but they may not reduce the individual's formal sector wage.

The issue of private and social returns to moonlighting suggests that as well as identifying the determinants of this aspect of labour supply, it is important to understand its normative implications. Should all forms of moonlighting be discouraged, and should this be achieved by increasing formal sector wages or increasing penalties on informal sector activities? Should formal sector employment arrangements be made more flexible - either by allowing more discretion in decisions about formal sector labour supply or by allowing outside work - in order to improve the efficiency of labour allocation?

3.3.2 Medical Care Suppliers in the Market

A recurring question in health economics regards the ability of physicians to induce consumers to purchase more medical care than they otherwise would. If such forces exist, it is natural to hope that restricting the supply of doctors might reduce consumption. However, economies usually expect reductions in supply to increase prices. To fully understand these possibilities, it is necessary to examine the interaction of supply and demand.

SELF ASSESSMENT EXERCISE

What are the factors considered by physicians in determining the amount of work to supply.

3.4 Interaction of Demand and Supply-Standard Analysis

So far we have examined the nature of the production processes of medical care and the incentives suppliers face in determining the type and quantity of care provided and the inputs used. In particularly, the influence of prices and other financial variables on supplier behavior was addressed. Abstracting from issues regarding the choice of inputs (that is, production techniques), it seems reasonable to suggest, as we did, that higher prices for physicians' services will lead to an increase in the quantity that physicians desire to supply. Although there will be

offsetting income and substitution effects, we usually assume that the labor supply curve S_0 (measuring, for example, hours worked) is upward-sloping. In a free, competitive market, wage rate is then determined by the intersection of the supply curve and the demand curve D_0, as in figure 2.6. While the relation between the price received by physician and the quantity they desire to provide, as described by the supply curve, is likely to be upward-sloping. We usually expect the relationship between the quantity actually supplied and the equilibrium price to be negative for a given demand curve. For example, suppose that the number of physicians in the marketplace increases because the government increases the number of places available at a public medical school or relaxes immigration controls for foreign doctors). There is no effect on the individual supply curves of physician who were initially active in the market. However, the total supply curve is now the horizontal sum of a larger number of individual supply curve because of the new entrants, and it shift out of S_1, as shown in figure 2.6, the equilibrium price of health services has fallen to P_1.

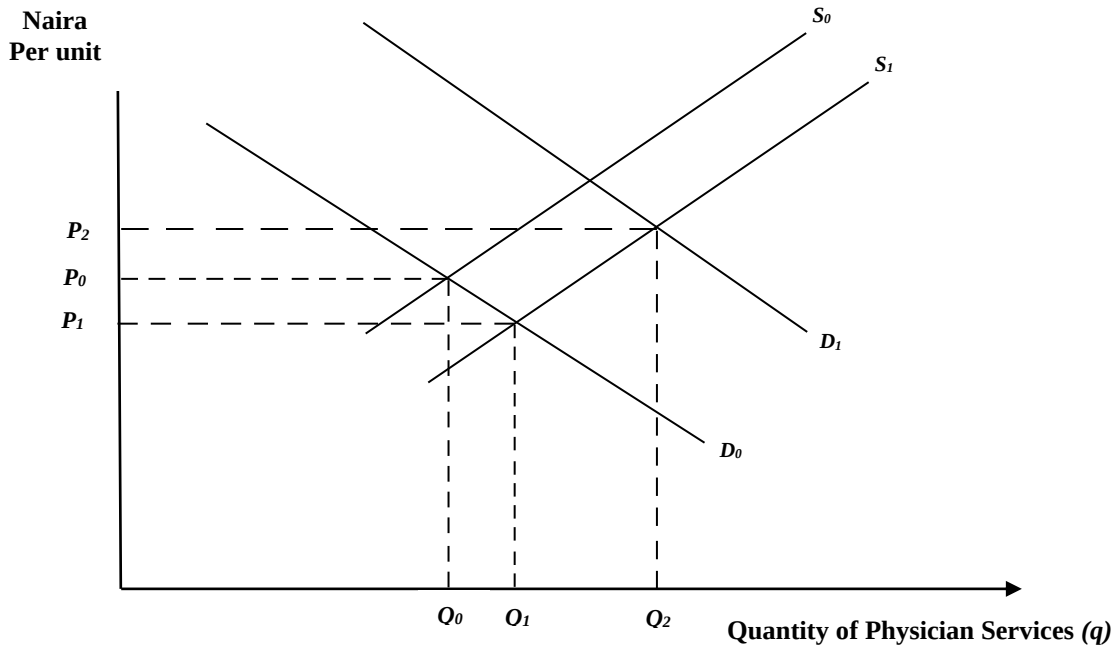


Fig. 2.6: Demand – Supply Analysis

3.4.1 Supplier-Induced Demand

The market analysis above assumed that the demand curve did not change when the supply curve shifted. This is a very orthodox assumption in microeconomic theory, the demand curve being the derived relationship between price and desired consumption, taking preferences (including health status), prices, and incomes as fixed. The behaviour of suppliers, in particular, influences the equilibrium price and quantity consumed but not the demand curve. In the case of health care, consumers rely on physicians for at least the execution of care. The information provided can relate to the effects of certain treatments, their likelihood of success, likely side effects and other risks associated with them, and the effect of having no treatment. Given this information, individuals make informed choices about the type and quantity of care they desire.

The trouble is, of course, that individuals have little opportunity to verify the information provided, and when the providers of the information stand to gain from providing misleading information, it is unlikely that correct information will be forthcoming. Thus, if the provider of information is the same person as the provider of medical services, individuals may be induced to consume more than they would if they had access to purely objective information. The position of the demand curve might thus be affected by the supplier of services.

This possibility does not answer our question about the ways in which a positive relationship between equilibrium prices and supply may arise. Indeed, if physicians can influence demand, why do they not push the demand curve out indefinitely and earn higher incomes? There must be some force that restrains suppliers from acting in such a fashion. One possibility is that physicians feel guilt from effectively “fooling” patients into having more treatment than is necessary, and this disutility acts as a constraint on the extent of induced demand. An alternative, and perhaps more natural, constraint on the ability of physicians to induce demand is competition. Over-servicing imposes some costs on individuals (even if they are fully insured against the financial costs of care), and if one physician is found to continually over-service, that individual will lose patients to other physicians. In a perfectly competitive market, this mechanism would restrain the ability of physicians to over-service completely, and they would

provide correct information. It is generally acknowledged, however, that the market for physician services is better described as monopolistically competitive because of switching costs. Thus, each supplier exercises a degree of monopoly power over his or her own patients, who incur additional costs if they switch to other doctors. These costs include the time it takes to find a new doctor with whom the patient feels comfortable, the uncertainty about the quality of a new doctor, and the additional visit, if any, that are required for the new doctor to establish the patient's medical history and condition. Given such local monopoly power, physician will be able to exercise a degree of demand inducement. The observed positive relationship between equilibrium price and quantity can then be rationalized by assuming that when faced with an increase in the number of physician, and thus an outward shift in the supply curve, each physician increases the amount of demand inducement he or she exercises. The effect is to shift the demand curve out to D_1, figure 2.7 at the new equilibrium-the intersection of D_1 and S_1-the price is P_2 and the total quantity is Q_2.

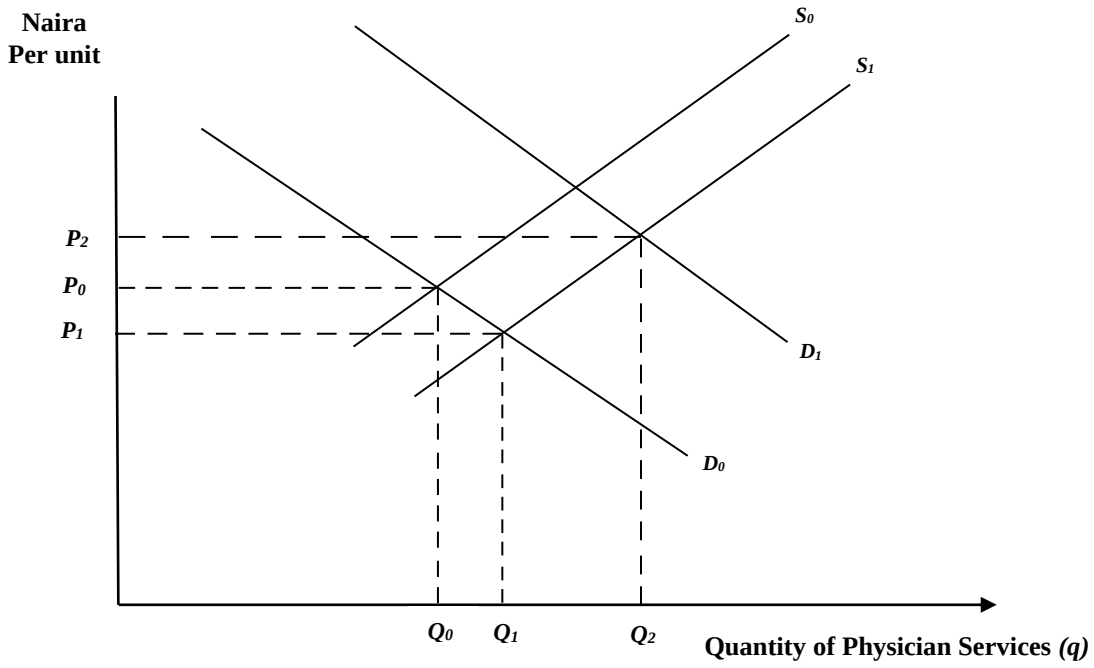


Fig. 2.7: The Supplier-Induced Demand Model

The second mechanism that might increase demand inducement following an increase in supply relates to the extent of the loss a physician experiences when a patient leaves the practice because of over-servicing. The initial increase in supply leads to a fall in unit price. Under our assumption of relatively inelastic demand, this will lead to an increase in the demand by each patient, but to a reduction in the amount of the expenditures. Therefore, the revenue that the physician earns from each patient falls, and the opportunity cost of losing a patient because of over-servicing also falls. That is, the "price" of over-servicing falls, so we expect more of it. Such a partial equilibrium argument only holds if incomes and other variables are fixed, and we know that the physician's income has fallen because of the reduction in the number of patients, as well as the reduction in revenue per patient. However, the price or substitution effect points in the direction of increased inducement following an outward shift in the supply curve. From the equilibrium discussion above, it should not be a surprise that health economists are divided about the existence of, and underlying mechanisms behind induced demand. There are many studies that attempt to document the existence of the phenomenon, with mixed success.

SELF ASSESSMENT EXERCISE

Discuss the concept of supplier-induced demand and state the factors responsible for it.

4.0 Conclusion

Physicians play two distinct roles as health care providers. They provide information and advice to patients on the nature of their condition and the likely impacts of particular treatments and engage in the physical delivery services, including surgery, administering of injections and writing of drug prescriptions. Nurses, administrators, clerks, receptionists, traditional healers, and general staff are also medical personnel. Some of the labour services provided by these individuals are substitutes for each other and for the work of physicians. Medical supplies, particularly drugs, instruments, and capital equipment are essential inputs into the production of health services. When estimating total output of a hospital, we need to aggregate the goods that the hospital produces and the technical relationships between inputs and outputs guide us on the patterns of resource allocation to see in the health care sector. Such resource allocations can be

effected either through direct administrative procedures or as a consequence of decisions made by private institution in the delivery and management of health care services.

5.0 Summary

This unit looked at inputs into the production of health services, incentives and the allocation of resources and labour supply. It also considered moonlighting as the practice of workers (not necessarily physicians) taking on a second job in addition to their primary employment. The unit also looked at the hospitals performance measurement, hospitals medical care suppliers in the market, the interaction of demand and supply-standard analysis and supplier-induced demand.

6.0 Tutor-Marked Assignment

1. Discuss the various inputs into the production of health care.
2. Given the technical relationships between inputs and outputs, what patterns of resource allocation are we likely to see in the health care sector?
3. Discuss the concept of supplier-induced demand and state the factors responsible for it.

7.0 References/Further Reading

Culyer, J.A., & Newhouse, J.P. (2000) Eds, Handbook of Health Economics: Vols 1A & 1B, Elsivier, North-Holland.

Donaldson Cam and Karen Gerard (1993) Economics of Health Care Financing: The Visible Hand. Macmillan Press Ltd. London.

Folland S., A. Goodman & M. Stano (2010) The Economics of Health & Health Care, Sixth Edition, Prentice Hall, New Jersey.

Jacobs, P. (1991) The Economics of Health and Medical Care Maryland: Aspen Pub Inc. Jack, Williams (1964) Principles of Health Economics for Developing Countries. WBI Development Studies. The World Bank, Washington D. C.

Jones Andrew (2007) Applied Econometrics for Health Economists: A Practical Guide, 2nd Edition OHE

Phelps Charles E. (1992) Health Economics, New York: Harper Collins Pub Inc.

Santerre E. & S.P. Neun (1996) Health Economics: Theories, Insights & Industry Studies, Irwin, Chicago.

Zweifel P., F. Breyer & M. Kifmann (2009) Health Economics, Second Edition, Springer Verlag Heidelberg.

Unit 3: Medical Care Production and Costs

CONTENTS

1.0 INTRODUCTION

All medical firms, including hospitals, physician, nursing homes, and pharmaceutical companies, earn revenues from producing and selling some type of medical output. Production activities occur regardless of the form of ownership (that is, for-profit, public, or not-for-profit). Because these activities take place in a world of scarce resources, microeconomics can provide valuable insights into the operation and planning processes of medical firms. This unit examines the medical care production and costs.

2.0 OBJECTIVES

It is expected that the following should be internalized at the end of this unit:

(i) The short-run production and costs analysis

(ii) The long-run production and costs analysis

(iii) Technical and economic efficient concepts

(iv) Marginal and average concepts analysis in decision making in health care.

3.0 MAIN CONTENT

3.1 The Short-run Production Function of the Representative Medical Firm

The focus of this topic is the various economics principles that guide the short-run production process of a hypothetical medical firm. Five major assumptions were made as the foundation of the discussions here. First, we assume that the medical firm produces a single output of medical

services, q. Second, we initially assume only two medical inputs exist: nurse-hours, n, and a composite capital good, k. We can think of the composite capital good as an amalgamation of all types of capital, including any medical equipment and the physical space in the medical establishment. Third, since the short run is defined as a period of time over which the level of at least one input is fixed, we assume the quantity of capital is fixed at some amount. This assumption makes intuitive sense, because it is usually difficult to change the stock of capital than the number of nurse-hours in the short-run. Fourth, we assume the medical firm faces an incentive to produce as efficiently as possible. Finally, we assume the medical firm possesses perfect information regarding the demands for its product. Recall that a production function identifies how various inputs can be combined and transformed into a final output. Thus, the short-run production function for medical services can be mathematically generalized as

$$q = f(n, k) \qquad (2.6)$$

The short-run production function for medical services in equation (2.6) indicates that the level of medical services is a function of a variable nurse input and fixed capital input. The production function identifies the different ways nurse-hours and capital can be combined to produce various levels of medical services. The production function allows for the possibility that each level of output may be produced by several different combinations of the nurse and fixed capital inputs. Each combination is assumed to be **technically efficient,** since it results in the maximum amount of output that is feasible given the state of technology. However, both technical and economic considerations determine a unique least-cost or **economically efficient** method of production. We begin our analysis by examining how the level of medical services, q, relates to a greater quantity of the variable nurse input, n, given that the capital input, k, is assumed to be fixed. Various microeconomic principles and concepts relating to production theory are used to determine the precise relation between the employment of the variable input and the level of total output. A microeconomic principle from production theory is the law of diminishing marginal productivity. This is about production behaviour and states that total output at first

increases at an increasing rate, but after some point increases at a decreasing rate, with respect to a greater quantity of a various input, holding all other inputs constant.

Figure (2.8) applies the law of diminishing productivity by showing a graphical relation between the quantity of medical services on the vertical axis and the number of nurse-hours on the horizontal axis.

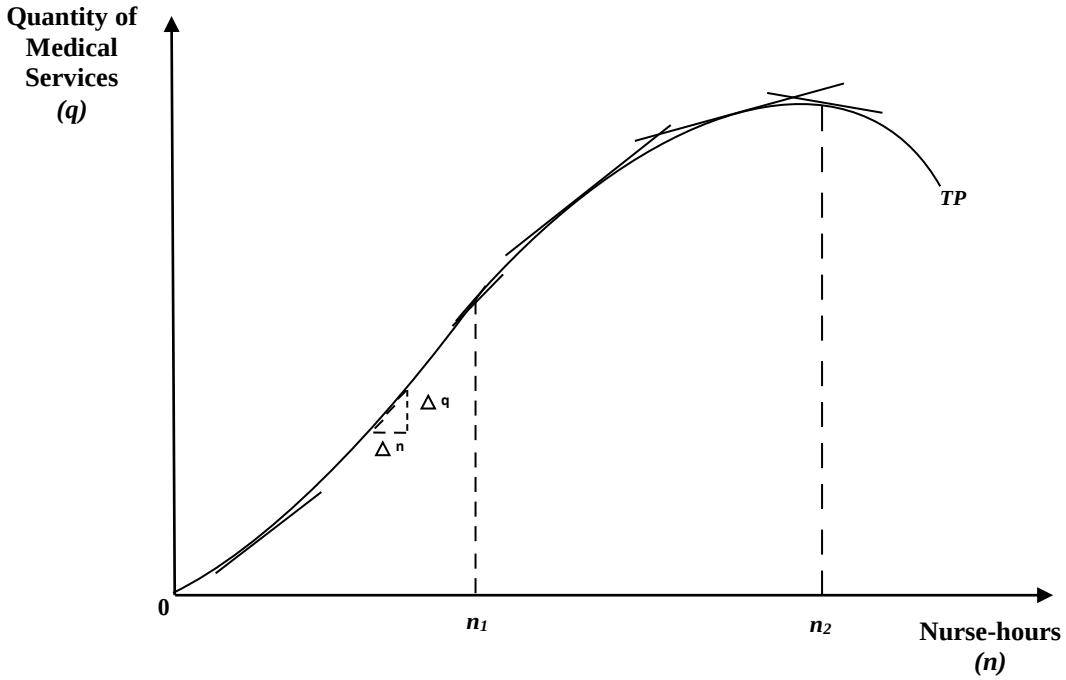


Fig. 2.8: The Total Product Curve

The curve is referred to as the total product curve, TP, because it depicts the total output produced by different levels of the variable input, holding all other input constant. Notice that the quantity of services first increases at an increasing rate over the range from n_1 to n_2, from 0 to n_1. The rate of increase is identified by the slope of the curve at each point. From the graph, the slope of the total product curve increases in value as the tangent lines become steeper over this range of nurse-hours. Beyond point n_1, however, further increases in nurse-hours cause medical services to increase at a decreasing rate. That is the point at which diminishing productivity sets in. Notice that the slope of the total product curve gets smaller as output

increases in the range from n_1 to n_2, (as indicated by the flatter tangent lines). At n_2, the slope of the total product curve is zero, as reflected in the horizontal tangent line. Finally, beyond n_2, there is possibility that too many nurse-hours will lead to a reduction in the quantity of medical services. The slope of the total product curve is negative beyond n_2. In terms of the production decision at the firm level, we have not yet accounted for the specific reasoning underlying the law of diminishing marginal productivity. Economists point to the fixed short-run inputs as the beats for diminishing productivity. For example, when nurse-hours are increased at first, there is initially a considerable amount of capital, the fixed input, with which to produce medical services. The large quantity of capital enables increasingly greater amounts of medical services to be generated from the employment of additional nurses. In addition, a combined effect may dominate initially. The combined effect means that nurses, working cooperatively as a team, are able to produce more output collectively because of labour specialization.

At some point, however, the fixed capital becomes limited relative to the variable input (for example, too little medical equipment and not enough medical space), and additional nurse-hours generate successively fewer incremental units of medical services. In the extreme, as more nurses are crowded into a medical establishment of a fixed size, the quantity of services may actually begin to decline as congestion sets in and creates unwanted production problem. In general, any physical constraint in production, such as the fixed size of the facility or a limited amount of medical equipment can cause diminishing productivity to set in at some point.

SELF ASSESSMENT EXERCISE

Discuss the concepts of diminishing productivity in medical care services.

3.2 Marginal and Average Products

Marginal product (MP) and average product (AP) curves can also be employed to illustrate the fundamental characteristics associated with the production process. In general, the marginal product is the change in total output associated with a one-unit change in the variable input. In terms of our example, the marginal product or quantity of medical services associated with an additional nurse-hour, MP_n, can be stated as follows:

$$MP_n = \Delta q/\Delta n \quad \ldots\ldots\ldots \quad (2.7)$$

The magnitude of the marginal product of a nurse-hour reveals the additional quantity of medical services produced by each additional nurse-hour. It is a measure of the marginal contribution of a nurse-hour in the production of medical services. In figure (2.8), the slope of the total product curve at every point represents the marginal product of a nurse-hour, since it measures the rise (vertical distance) over the run (horizontal distance), or $\Delta q/\Delta n$. Consequently, we can determine the MP of an additional nurse-hour by examining the slope of the total product curve at each level of nurse-hour. Figure (2.9) illustrates the marginal product of a nurse-hour.

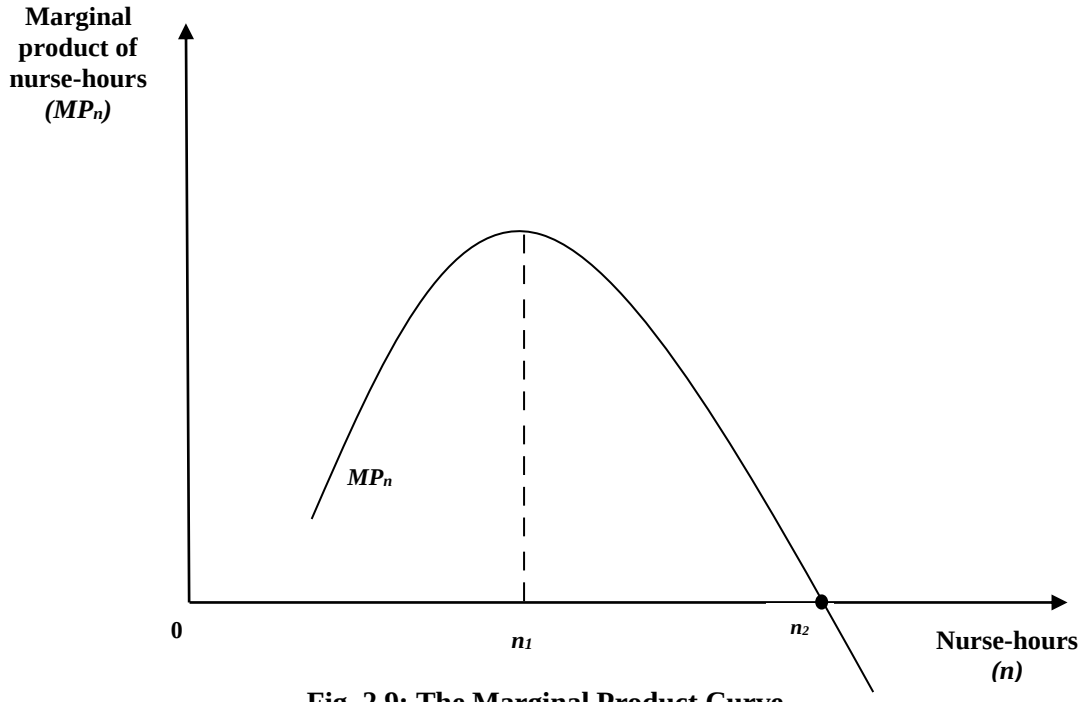


Fig. 2.9: The Marginal Product Curve

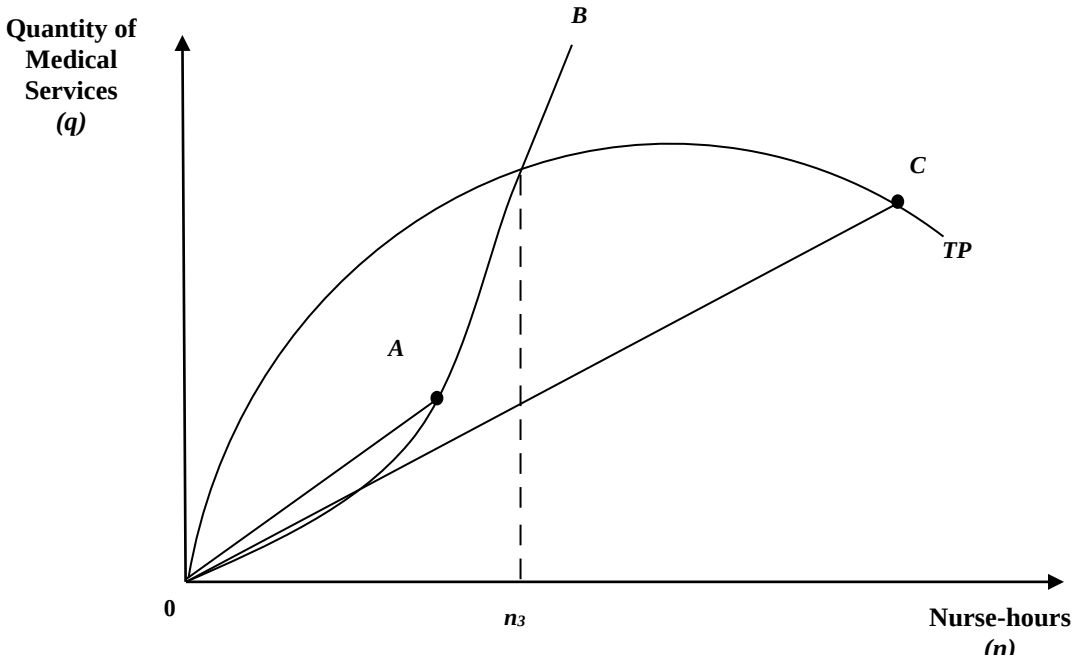
Initially, MP_n, is positive and increases over the range from 0 to n_1 due to increasing marginal productivity. In the range from n_1 to n_2, the MP is positive but decreasing, because diminishing marginal productivity has set in. At n_2, the marginal product of a nurse-hour is zero and becomes negative thereafter. The MP curve suggests that each additional nurse-hour cannot be expected to generate the same marginal contribution to total output as the previous one. The law of

diminishing marginal productivity dictates that in the short run, a level of output is eventually reached where an incremental increase in the number of nurse-hours leads to successively fewer additions to total output (because some other inputs are fixed).

In addition to MP_n, the AP of a nurse-hour can provide insight into the production process. In general, the average product equals the total quantity of output divided by the level of the variable input. In terms of the present example, the average product of a nurse-hour, AP_n, is calculated by dividing the total quantity of medical services by the total number of nurse-hour:

$$AP_n = q/n \text{ ……. } (2.8)$$

The AP of a nurse-hour measures the average quantity of medical services produced within an hour. For example, suppose total medical services are measure by the number of daily patient-hours at a medical facility. Also, suppose 200 nurse-hours are employed to serve 400 daily Patient-hours. The average product of a nurse-hour equals 400/200 or 2 patients per hour. We can also drive the AP of a nurse-hour from the total product curve, as shown in figure (2.10a)



a.

Fig. 2.10a: Deriving the Average Product Curve from the Total Product Curve

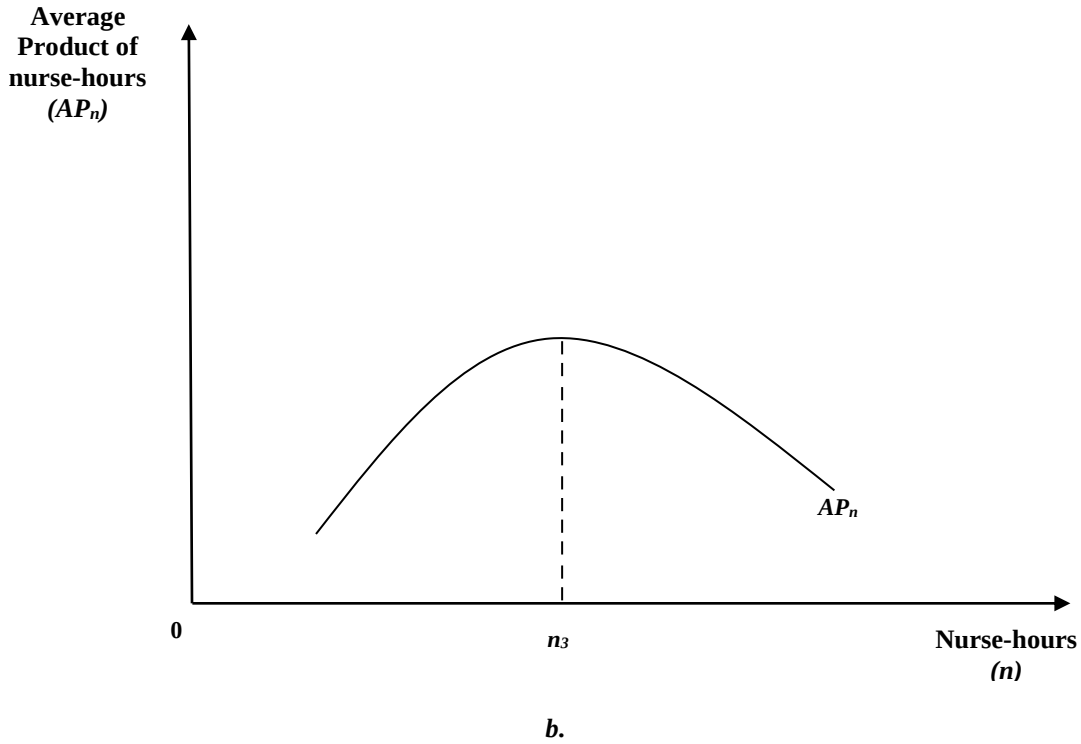


Fig. 2.10b: Deriving the Average Product Curve from the Total Product Curve

To derive AP_n, a ray from the origin is extended to each point on the total product curve. The slope of the ray measures AP_n, for given level of nurse-hours, since it equals the rise over the run, or q/n. In figure (2.10a) three rays, labeled OA, OB, and OC, emanate from the origin to the total product curve. The slope of ray OA is flatter than that of OB and therefore is of a lower magnitude. In fact, as it is drawn, ray OB has a greater slope than any other ray emanating from the origin. At this level of nurse-hours, the AP is maximized. The slope of ray OC is flatter and of a lower magnitude than that of OB. The implication is that AP initially increases over the range from 0 to n_3, reaches a maximum at n_3, and then decreases, as in figure (2.10b). It is the law of diminishing marginal productivity that accounts for the shape of AP_n.

In figure (2.11), the marginal and average product curves are superimposed to illustrate how they are related. Some characteristics of the relation between these two curves are worth mentioning. First, the marginal product curve cuts the average product curve at its maximum

point. In fact, it is a common mathematical principle that the marginal equals the average when the average is at its extreme value. Second, MP_n lies above AP_n, whenever AP_n, is increasing.

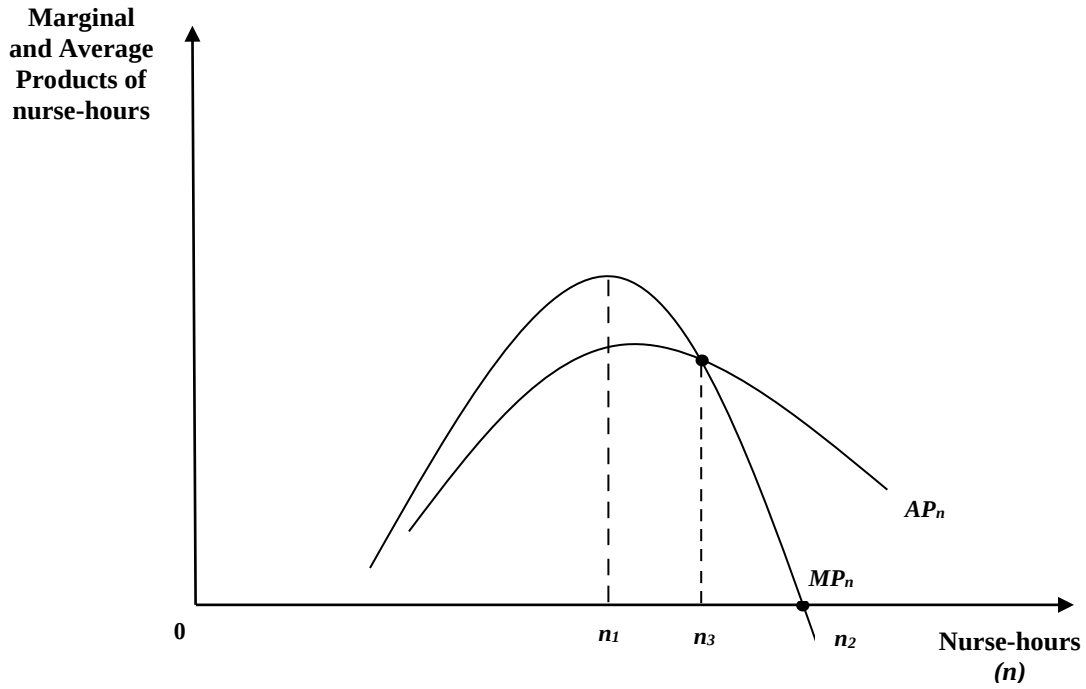


Fig. 2.11: Relationship between the Marginal and Average Product Curves

Third, MP_n, lies below AP_n, whenever AP_n, is declining. The relation between the marginal and average product curves can be examined in terms of our example of nurse-hours and the production of medical services. For this discussion, the marginal product curve is the amount of medical services generated hourly by the next nurse hired while average product curve is the average quantity of medical services generated by the existing team of nurses within an hour – i.e. team average. Looking at figure (2.11), notice that the next nurse hired always generates more services per hour than the team average up to point n_3. Consequently, up to this point, each additional nurse helps pull up the term's average level of output. Beyond n_3, however, the incremental nurse hired generates less service per hour than the team average, as a result; the team average falls. It is important to realize that any increase or decrease in the marginal product has nothing to do with the individual talents of each additional nurse employed. Rather, it involves the law of diminishing marginal productivity. At some point in the production process,

the incremental nurse becomes less productive due to the constraint imposed by the fixed input. The marginal productivity, in turn, influences the average productivity of the team of nurses.

At first glance, it seems logical to assume that a medical firm desires to produce at a point like n_1 or n_2, in figure (2.11). After all, they represent the points at which either the marginal or the average product is maximized. In most cases, however, a medical firm finds it more desirable to achieve some financial target, such as a maximum or break-even level of point. Hence, we need more information about the revenue and cost structures the medical firm faces before we can determine the desired level of production.

3.2.1 Elasticity of Input Substitution

So far, we have assumed only one variable input. Realistically, the medical firm operates with more than one variable input in the short run. Thus, there may be some possibilities for substitution between any two variable inputs. For instance, licensed practical nurses often substitute for registered nurses in the production of patient services, and physician assistants substitute for physicians in the production of ambulatory services. The actual degree of substitutability between any two inputs depends on technical and legal considerations. For example, physician assistances are prohibited by law from prescribing medicines in most states. In addition, licensed practical nurses normally lack the technical knowledge needed to perform all the duties of registered nurses. In general terms, the elasticity of substitution between any two inputs equals the percentage change in the input ratio divided by the percentage change in the ratio of the inputs' marginal productivities, holding constant the level of output, or

$$\sigma = \Delta(I_1/I_2)/I_1/I_2 \div \Delta\,(MP_2/MP_1)/MP_2/MP_1 \ldots\ldots\ldots \quad (2.9)$$

I_i (I = 1, 2) stands for the quantity employed of each input. The ratio of marginal productivities, MP_2,/MP_1, referred to as the marginal rate of technical substitution, illustrates the rate at which one input substitutes the other in the production process, at the margin. For example, suppose the marginal product of a registered nurse-hour is four patients and the marginal product of a licensed practical nurse-hour is two patients. It follows that two-licensed practical nurse-hours

are needed to substitute completely one registered nurse-hour. Theoretically, σ takes on values between 0 and +∞ and identifies the percentage change in the input ratio that results from a 1 percent change in the marginal rate of technical substitution. The magnitude of σ identifies the degree of substitution between the two inputs. For example, if σ = 0, the variable inputs cannot be substituted in production. In contrast, when σ = ∞, the two variable inputs are perfect substitutes in production. In practice, it is more common for σ to take on values between these two extremes, implying that limited substitution possibilities exist.

SELF ASSESSMENT EXERCISE

Using Marginal product (MP) and average product (AP), illustrate the characteristics associated with the production process.

3.3 Short-run Cost Theory of the Representative Medical Firm

Before we begin our discussion of the medical firm's cost curves, we need to address the difference between the ways economists and accountants refer to costs. In particular, accountants consider only the explicit costs of doing business when determining the accounting profits of a medical firm. Explicit costs are easily quantified because a recent market transaction is available to provide an accurate measure of cost. Wage payments to the hourly medical staff, electric utility bills and medical supply expenses are all examples of explicit costs medical firms incur because disbursement records can be consulted to determine the magnitudes of these expenditures. Economists, unlike accountants, consider both the explicit and implicit costs of production. Implicit costs reflect the opportunity costs of using any resources the medical firm owns. For example, a general practitioner (GP) may own the physical assets (such as the clinic and medical equipment) used in producing physician services. In this case, a recent market transaction is unavailable to determine the cost of using these assets. Yet an opportunity cost is incurred when using them because the physical assets could have been rented out for an alternative use. For example, the clinic could be remodeled and rented as a beauty salon, and the medical equipment could be rented out to another physician. Thus, the forgone rental payments reflect the opportunity cost of using the physical assets owned by the GP.

Consequently, when determining the economic (than accounting) profits of a firm, economists consider the total cost of doing business including both the explicit and implicit costs. Economists believe it is important to determine whether sufficient revenues are available to cover the cost of using all inputs, including those rented and owned. For instance, if the rental return on the physical assets is greater than the return on use, the GP might do better by renting out the assets rather than retaining them for use.

3.3.1 The Short-Run Cost Curves of the Representative Medical Firm

Cost theory is based on the production theory of the medical firm and relates the quantity of output to the cost of production. As such, it identifies how (total and marginal) costs respond to changes in output. If we continue to assume the two inputs of nurse-hours, n, and capital, k, the short-run total cost, STC, of producing a given level of medical output, q, can be written as

$$STC(q) = w * n + r * k \quad \ldots\ldots\ldots\ldots \quad (2.10)$$

Where: w and r represent the hourly wage for a nurse and the rental or opportunity cost of capital, respectively. Input prices are assumed to be fixed, which means the single medical firm can purchase these inputs without affecting their market prices. This is a valid assumption as long as the firm is a small buyer of input relative to the number of buyers in the marketplace. Equation (2.10) implies that the short-run total costs of production are dependent on the quantities and prices of outputs employed. The wage rate times the number of nurse-hours equals the total wage bill and represent the total variable costs of production. Variable costs respond to changes in the level of output. The product of the rental price and the quantity of capital represent the total fixed costs of production. Obviously, this cost component does not respond to changes in output, since the quantity of capital is fixed in the short run. The total product curve not only identifies the quantity of medical output produced by a particular number of nurse-hours but also shows, reciprocally, the numbers of nurse-hours necessary to produce a given level of medical output. With this information, we can determine the short-run total cost of producing different levels of medical output by following a three-step procedure. First, we identify, through the production function, the necessary number of nurse-hours, n, for each level

of medical output. Second, we multiply the quantity of nurse-hours by the hourly wage, w, to determine the short-run total variable costs, STVC, of production, or w *n. Third, we add the short-run total fixed costs, STFC, or r x k, to STVC for each level of medical output, we can derive a short-run total cost curve like the one in figure (2.12).

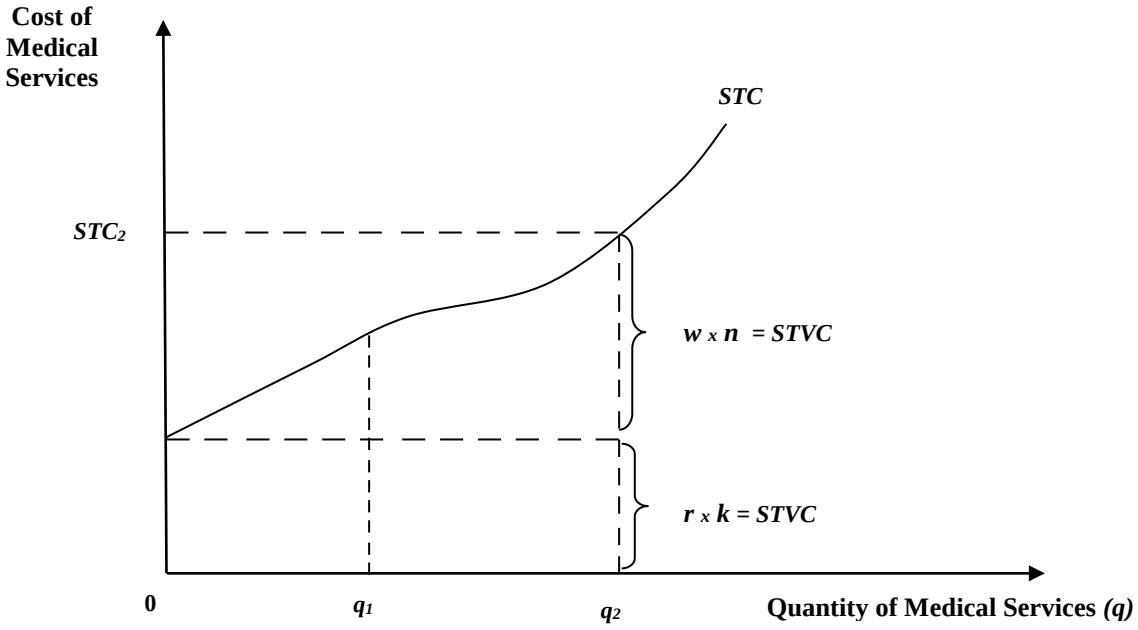


Fig. 2.12: The Short-Run Total Cost Curve

Notice the reciprocal relation between the short-run total cost function in figure (2.12) and the short-run total product curve in figure (2.8). For example, when total product is increasing at an increasing rate up to point n_1 in figure (2.8), short-run total costs are increasing at a decreasing rate up to point q_1, in figure (2.12). This is because the increasing productivity in this range causes the total costs of production to rise slowly. Output increases at a decreasing rate immediately beyond point n_1, in figure (2.8) (as shown by the slope of the total product curve), and as a result, short-run total costs increase at an increasing rate beyond q_1 in figure (2.12). Also notice that total costs increase solely because additional nurses are employed as output expands. Figure (2.12) also shows how short-run total cost can be decomposed into its variable and fixed components for the level of output q_2.

In practice, distinguishing between fixed and variable costs can be challenging. Recall that variable costs change proportionately, whereas fixed costs do not change, in response to any adjustment in the quantity of output produced. Fixed costs occur in the short-run, during the operating period, when the levels of some inputs are fixed. In contrast, all inputs are variable during the long run planning period, when, for instance, future budgets are being designed. The physical size of a production facility is often treated as a fixed input because a significant amount of time is needed to construct a larger building. Hourly workers are treated as a variable input because they can be promptly hired or laid off, depending on the desired adjustment in output. Thus, time plays a crucial role in determining the fixity of inputs and costs. It follows that long-term contracts, which although provide offsetting benefits, impose more fixed costs on a firm's budget. Given that a majority of costs were fixed in the short run, a reduction in hospital services would have a very little impact on costs in the short run.

3.3.2 Short-Run Per-Unit Costs of Production

We can look at the reciprocal relation between production and costs by focusing on the short-run marginal and average variable costs of production. The short-run marginal costs, SMC, of production are equal to the change in total costs associated with a one-unit change in output, or

$$SMC = \Delta STC/\Delta q \quad \ldots\ldots\ldots\ldots\ldots \quad (2.11)$$

In terms of equations (2.10) and (2.11), the short-run MC of production looks like the following:

$$SMC = \Delta(w*n + r*k)/\Delta q \quad \ldots\ldots\ldots \quad (2.12)$$

Because the wage rate and short-run fixed cost are constant with respect to output, equation (2.12) can be rewritten in the following manner:

$$SMC = w^* (\Delta n/\Delta q) = w^* [1/MP_n] = w/MP_n \ldots\ldots\ldots\ldots \quad (2.13)$$

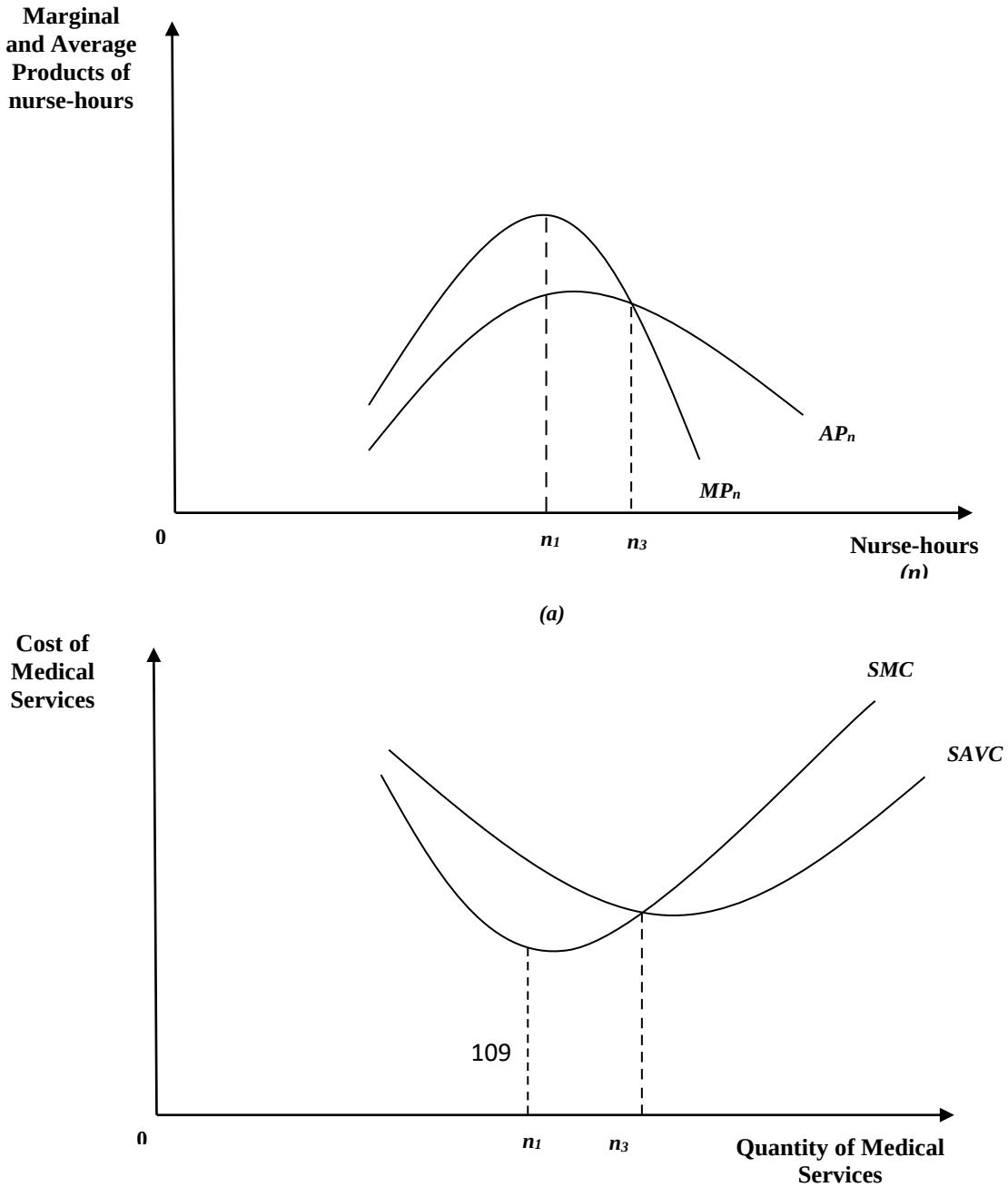
Notice on the right-hand side of equation (2.13) that short-run marginal costs equal the wage rate divided by the marginal product of nurse-hours. The short-run average variable costs,

SAVC, of production equal the short-run total variable costs, STVC, divided by the quantity of medical output. Because STVC is the total bill (that is, w*n):

$$SAVC = STVC/q = (w*n)/q = w*[1/AP_n] = w/AP_n \text{ -------- } (2.14)$$

Such that SAVC equals the wage rate divided by the average product of a nurse-hour. The short-run marginal and average variable costs are inversely related to the marginal and average products of labor, respectively. Thus, MC and AVC increase as the MP and AP fall, and vice, versa. Figure (2.13a) shows the graphical relation between the per-unit and cost curves

Fig. 2.13: Relationship between the Per-Unit Product and Cost Curves



The two graphs in figure (2.13) clearly point out the reciprocal relation between production and costs. For example, after point n_1, in figure 2.13a, diminishing productivity sets in and the marginal product begins to decline. As a result, the short-run marginal costs ($=w/MP_n$) increase beyond output level q_1 given a fixed wage. Similarly, the average product of a nurse-hour declines beyond n_3, so the AVC of production increase beyond q_3. Obviously, the shapes of MC and AVC curves reflect the law of diminishing marginal productivity. Because of this reciprocal relation, production and costs represent dual ways of observing various characteristics associated with the production process. It is clear from Equations (2.13) and (2.14) that the maximum points on the MP and AP curves correspond directly to the maximum points on the MC and AVC curve. Note in figure 2.13b that the short-run MC curve passes through the minimum point of the short-run AVC curve. In addition, the SMC curve lies below the SAVC curve when the latter is decreasing and above the SAVC curve when it is increasing.

In sample terms, the graph in figure 2.13b identifies how costs behave as the medical firm alters output in the short-run. Initially, as the medical firm expands output and employs more nurse-hours, both the MC and AVC of production decline. Eventually, diminishing productivity sets in due to the fixed input, and both MC and AVC increase. It follows that the MC and AVC of production depend in part on the amount of output a medical firm produces in the short-run. Besides the MC and AVC of production, decision makers are interested in the short-run ATC of operating the medical firm. From Equation 2.11, we can find the short-run ATC of production by summing the AVC and AFC. Short-run AFC (SAFC) are simply TFC (STFC) divided by the level of output, or

$$SAFC = STFC/q \quad \ldots\ldots\ldots\ldots \quad (2.15)$$

Because by definition the numerator in equation (2.15) is fixed in the short-run, the SAFC declines as the denominator, while medical services, increases in value. Consequently, the AFC of production decline with greater amounts of output because TFC (or overhead costs) are spread out over more and more units. Figure 2.14 shows the graphical relation among SMC, and SATC. Note that the MC curve cuts the ATC curve at its minimum point. (The minimum SATC lies to the right of the minimum SAVC). Also, the vertical distance between the ATC and VC

curve at each level of output represents the AFC of production, since TC include both VC and FC. The vertical distance between the two curves gets smaller as output increases because the SAVC approaches zero with increase in output. One implication of this is that ATC increase at some level of output because eventually the cost-enhancing impact of diminishing productivity outweighs the cost-reducing tendency of the AFC.

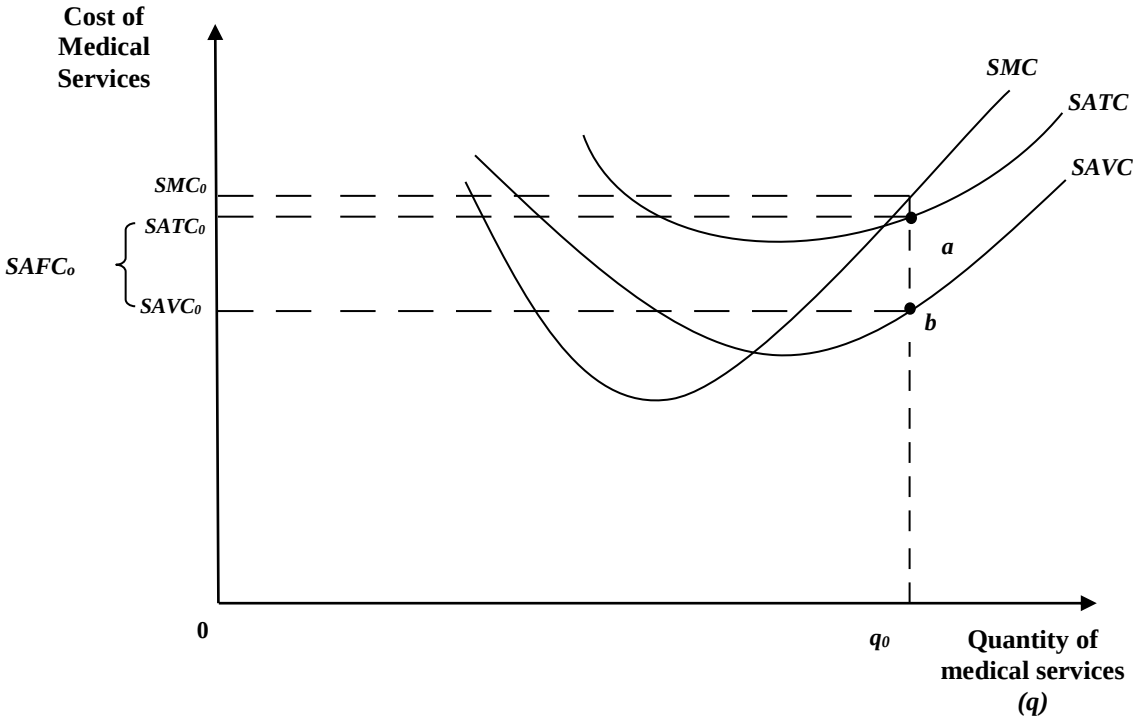


Fig. 2.14: Relation among Short-Run Marginal, Average Variable, and Average Total Costs

The unsuspecting reader may think that the medical firm should choose to produce at the minimum point on the SATC curve because average costs are minimized. As mentioned earlier, however, the level of output the medical firm chooses depends on the firm's objective (for example, to achieve maximum or break-even level of profits). Hence, a proper analysis requires some knowledge of the revenue structure in addition to the cost structure. However, assume that the firm has chosen to produce the level of medical output, q_0, in figure 2.14, let's identify the various costs associated with producing q_0, units of medical output. The identification of the per-

unit cost of producing a given level of output is a fairly easy matter; we can determine the per-unit cost by extending a vertical line from the appropriate level of output until it crosses the cost curve. For example, the average total cost of producing q_0 units of output is $SATC_0$, while the average variable cost is $SAVC_0$. The average fixed cost of producing q_0, units of output is represented by the vertical distance between $SATC_0$ and $SAVC_0$, or distance ab. In addition, SMC_0, identifies the marginal cost of producing one more unit assuming the medical firm is already producing q_0 units of medical services. Now suppose that instead of the per-unit costs, we want to identify the various total costs (i.e. STC, STVC and STFC) associated with producing q_0, units of output. We can do this by multiplying the level of output by the per-unit costs of production. For example, the rectangle $SAVC_0$-b-q0-0 in figure 2.14 measures the TVC of producing q_0 units of output, since it corresponds to the area found by multiplying the base of 0-q0 by the height of 0-SAVC0. Following similar logic, the TFC are represented by rectangle $SATC_0$,-a-$SAVC_0$, and total costs can be measured by area $SATC_0$-a-q_0-0.

3.3.3 Factors Affecting the Position of the Short-Run Cost Curves

A variety of short-run circumstances affect the position of the per-unit and TC curves. Among them are the prices of the variable inputs, the quality of care, the patient case-mix, and the amounts of the fixed inputs. Whenever any one of these variables changes, the position of the cost curves changes through either an upward or a downward shift depending on whether costs increase or decrease. For example, if input prices increase in the short-run, the cost curves shift upward to reflect, the higher costs of production (especially since $SAVC = w/AP_n$ and $SMC = w/MP_n$). If input prices fall in the short-run, the cost curves shift downward to indicate the lower production costs. Furthermore, if the medical firm increases the quantity of care or adopts a more severe patient case-mix, the cost curves respond by shifting upward. That is because a higher quality of care or a more severe patient case-mix means that a unit of labour is less able to produce as much output in a given amount of time. In terms of our formal analysis, a higher quality of care or a more severe patient case-mix reduces the average and marginal productivity of the labour input and thereby raises the costs of production. For example, a nurse can care for many more patients within an hour when these patients are less severely ill and quality of care is

of secondary importance. Conversely, a reduction in the quality of care or a less severe patient case-mix is associated with lower cost curves.

Finally, a change in the amount of the fixed inputs can change the costs of production. For instance, excessive amounts of the fixed inputs can lead to higher short-run costs. In sum, a specified short-run TVC function for medical services should include the following variables:

STVC = f(output level, input prices, quality of care, patient case-mix, quantity of the fixed inputs) (2.16)

These factors can explain cost differentials among medical firms in the same industry. Specifically, output influences short-run variable costs by determining where the medical firm operates along the cost curve, whereas the other factors affect the location of the curve. Most likely, high cost medical firms are associated with more output, higher wages, increased quality, more severe patient case-mixes, and / or an excessive quantity of fixed inputs.

3.3.4 The Cost-Minimizing Input Choice

A medical firm makes choices about which variable inputs to employ. Knowing that there is more than one way to produce a specific output, medical firms desire to produce with the least-cost or cost-minimizing input mix. For instance, suppose administrators desire to produce some given amount of medical services, q0, at minimum total cost, TC, using two variable inputs: registered nurses (RN), and licensed practical nurses (LPN), (we ignore the capital input in this example). These two inputs are paid hourly wages of W_R and W_L, respectively. The medical firm wants to minimize:

$$TC(q_0) = W_R*RN + W_L*LPN \quad \ldots\ldots\ldots\ldots \quad (2.17)$$

Subject to

$$q_0 = f(RN, LPN) \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.18)$$

by choosing the proper mix of registered nurses and licensed practical nurses. Taken equations (2.17) and (2.18) together, means that administrators want to minimize the total cost of producing q_0 units of medical services by choosing the “right” or efficient, mix of RNs and LPNs so that $TC(q_0)$ is as low as possible and sufficient amounts of the two inputs are available

to produce q_0. The efficient combination depends on the marginal products and relative prices of the two inputs. By using constrained optimization technique, the efficient mix of RNs and LPNs is chosen when the following condition holds:

$$MP_{RN}/W_R = MP_{LPN}/W_L \text{ ……. (2.19)}$$

Equation (2.19) means that the marginal product to price ratio is equal for both registered nurses and licensed practical nurses in equilibrium. The equality implies that the last dollar spent on registered nurses generates the same increment to output as the last naira spent on licensed practical nurses. As a result, a rearranging of expenditures on the two inputs cannot generate any increases in medical services, since both inputs generate the same output per naira at the margin. To fully appreciate this point, suppose this condition does not hold such that

$$MP_{RN}/W_R < MP_{LPN}/W_L \text{ ……. (2.20)}$$

In that case, the last naira spent on a licensed practical nurse generates more output than the last naira spent on a registered nurse. A licensed practical nurse is more profitable for the hospital at the margin, because the medical organization receives a "bigger bang for the buck." But as the organization hires more LPNs and fewer RNs, the marginal productivities adjust until the equilibrium condition in equation (2.19) holds. Specifically, the marginal productivity of the LPNs decreases, while the marginal productivity of the RNs increases due to diminishing marginal productivity. For example, suppose a newly hired RN can service six patients per hour and newly hired LPN can service only four patients per hour. At first blush, with no consideration of the price of each input, the RN might appear to be "better buy" because productivity is 50 percent higher. But suppose further that the market wage for an RN is ₦20 per hour, while an LPN requires only ₦10 per hour to work at the medical facility. Given relative input prices, the 50 percent higher productivity of the RN costs the medical facility 100 percent more. Obviously, the LPN is the better buy. That is, the last naira spent on an LPN results in the servicing of 0.4 additional patients per hour, while a naira spent on an RN allows the servicing of only 0.3 more patiently per hour.

SELF ASSESSMENT EXERCISE

Explain the factors that can cause cost differentials among medical firms in the same industry.

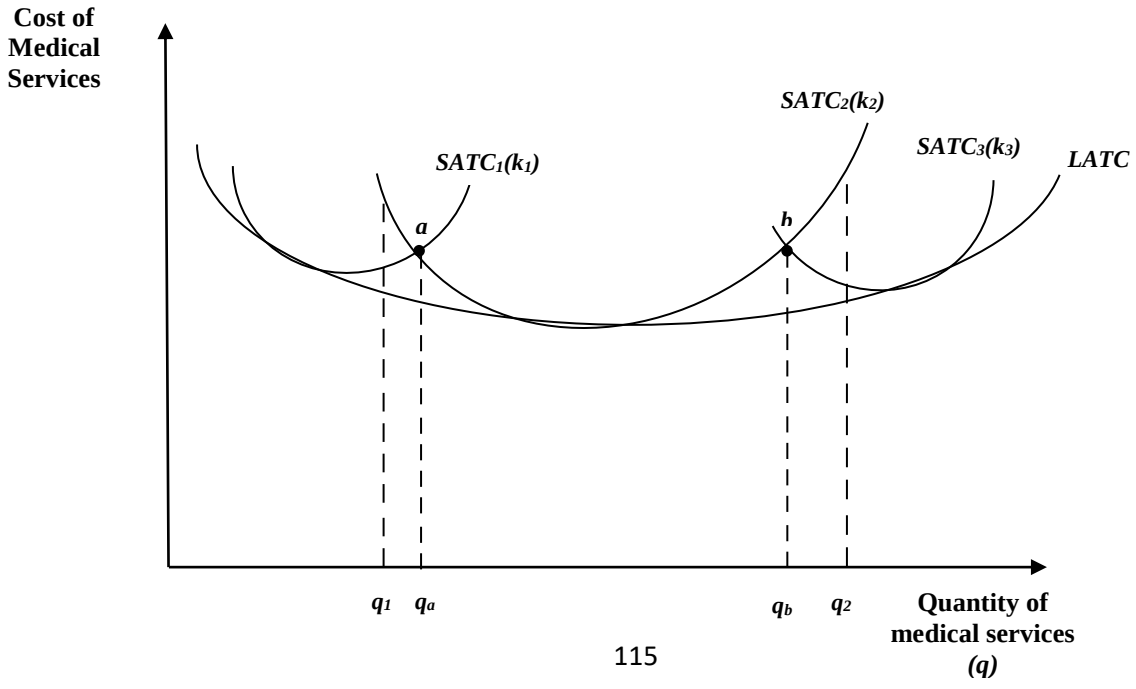
3.4 Long-Run Costs of Production

So far, we have focused on the short-run costs of operation and assumed that one input is fixed. The fixed input leads to diminishing returns in production and to U-shaped average variable and total costs curve. In the long run, however, when the medical firm is planning for future resource requirements, all inputs, including capital, can be changed. Therefore, it is also important to analyze the relation between output and costs when all inputs are changed in the long run.

3.4.1 Long Run Cost Curves

The long-run ATC curve can be derived from a series of short-run cost curves, as shown in figure (2.15). The three short-run ATC curves in the figure reflect different amounts of capital. For example, each curve might reflect the short-run ATC of producing units of medical services in physically larger facilities of sizes K_1, K_2, and K_3. If decision makers know the relation among different-size facilities and the short-run ATC, they can choose the SATC or size that minimizes the AC of producing each level of medical services in the long run. For example, over the range 0 to q_a, facility size K_1, results in lower costs of production than either size K_2 or K_3. Specifically, at output level q_1, $SATC_2$ exceeds $SATC_1$ by a significant amount.

Fig. 2.15: The Short-Run Average Cost Curves and the Long-Run planning Curve



Therefore, the administrators choose size K_1 if they desire to produce q1 of medical services at least cost in the long run. Similarly, from q_a to q_b, facility size associated with $SATC_2$, results in lower costs than either size K_1 or K_3. Beyond q_b units of medical services (say, q_2), a size of K_3 lower costs of production in the long run.

The three short-run cost curves in figure (2.15) paint a simplistic picture, since conceptually each unit of medical services can be linked to a uniquely sized cost-minimizing facility (assuming capital is divisible). If we assume a large number of possible sizes, we can draw a curve. Each point indicates the least costly way to produce the corresponding level of medical services in the long run when all inputs can be changed. Every short-run cost curve is tangent to the connecting or envelope curve, which is referred to as the long-run average total cost (LATC) curve. The curve drawn below the short-run average cost curves in figure (2.15) represents a long-run average total cost curve. Notice that the U-shaped long-run average cost curve initially declines, reaches a minimum, and eventually increases. Interestingly, both the short-run average cost curves have the same shape, but for different reasons. The shape of the short-run ATC curve is based on the law of diminishing productivity setting in at some point. In the long run, however, all inputs are variable, so by definition, a fixed input cannot account for the U-shaped long-run AC curve. Instead, the reason for the U-shaped LATC curve is based on the concepts of long-run economies and diseconomies of scale.

The **long-run economies of scale** refer to the notion that average costs fall as a medical firm gets physically larger due to specialization of labour and capital. Larger medical firms are able to utilize larger and more specialized equipment and to fully specialise the various labour tasks involved in the production process. For example, people generally get very proficient at a specific task when they perform it repeatedly. Therefore, specialization allows larger firms to produce increased amounts of output at lower per-unit costs. The downward-sloping portion of the LATC curve in figure (2.15) reflects economies of scale. Another way to conceptualize long-run economies of scale is through the direct relation between inputs and output, or returns to scale, rather than output and costs. Consistent with long run economies of scale is increasing

returns to scale. **Increasing returns to scale** result when an increase in all inputs results in a more than proportionate increase in output. For example, a doubling of all inputs that result in three times as much output is a sign of increasing returns to scale. Similarly, if a doubling of output can be achieved without a doubling of all inputs, the production process exhibits long-run increasing returns or economies of scale. Most economists believe that economies of scale are exhausted at some point and **diseconomies of scale** set in. **Diseconomies of scale** result when the medical firm becomes too large. Bureaucratic red tape becomes common, and top-to-bottom communication flows break down. The breakdown in communication flows means management at the top of the hierarchy has lost sight of what is taking place at the floor level. As a result, poor decisions are sometimes made when the firm is too large. Consequently, as the firm gets too large, long-run average costs increase. Diseconomies of scale are reflected in the upward-sloping segment of the LATC curve in figure (2.15)

Diseconomies of scale can also be interpreted as meaning that an increase in all inputs results in a less than proportionate increase in output, or **decreasing return to scale.** For example, if the number of patient-hour doubles at a dental office and the decision maker is forced to triple the size of each input, the production process at the dental office is characterized by decreasing returns, or diseconomies of scale. The production process may also exhibits **constant returns to scale**. **Constant returns to scale** occur when a doubling of inputs results in a doubling of output. In terms of long-run costs, constant returns imply a horizontal LATC curve, implying that long-run ATC is independent of output.

3.4.2 Shifts in the Long-Run Average Cost Curve

The position of the long-run average cost curve is determined by a set of long-run circumstances that includes the prices of all inputs (capital is a variable input in the long run), quality, and patient case-mix. When these circumstances change on a long-run basis, the long-run AC curve shifts up or down depending on whether the change involves higher or lower long-run costs of production. For example, an increase in the long-run price of medical inputs leads to an upward shift in the long-run AC curve. A cost-saving technology tends to shift the long-run average cost curve downward. Conversely, a cost-enhancing technology increases the AC of production in

the long-run and shifts the LATC curve upward. Higher quality of care and more severe patient case-mixes also shift the LATC curve upward.

3.4.3 Long-Run Cost Minimization and the Indivisibility of Fixed Inputs

Long-run cost minimization assumes that inputs can be costless adjusted. For an input such as an hourly labourer, employment adjustments are simple because number of workers can be changed easily. Capital inputs cannot always be easily changed, because they are less divisible. As a result, a medical firm facing a decline in demand may be unable to reduce the physical size of its facility. Medical firms may adjust slowly to external changes, not produce in LR equilibrium, and operate with excess capital relative to a LR equilibrium point. Figure 2.16 clarifies this point

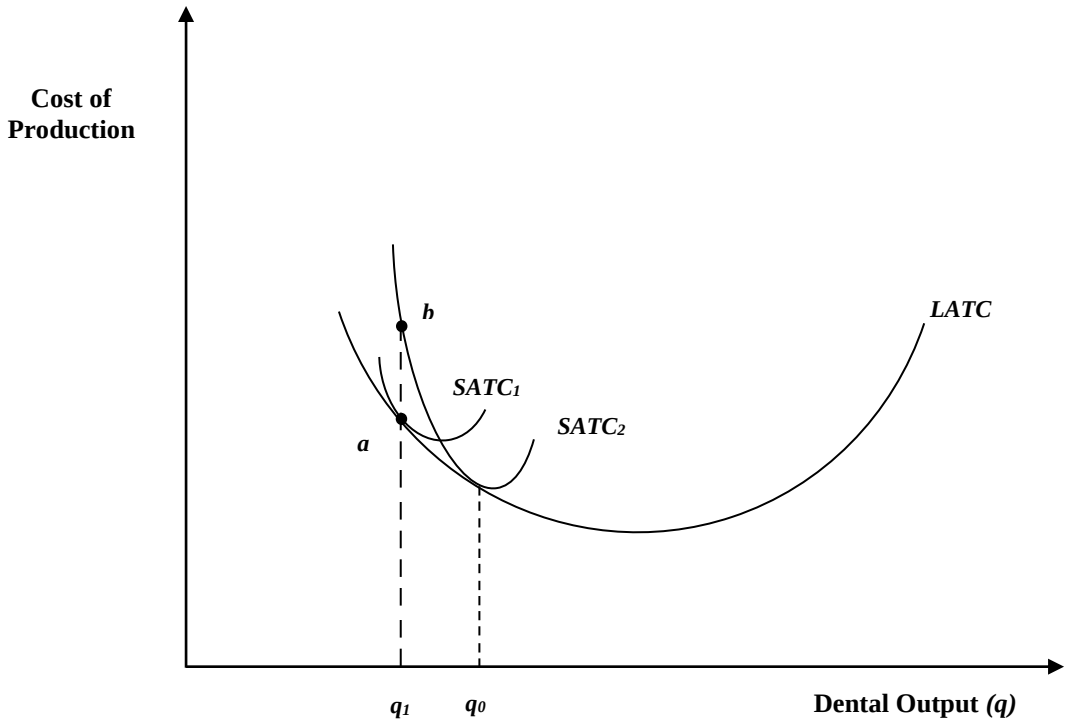


Fig. 2.16: Long-Run Disequilibrium of the Medical Firm

Suppose that initially a dental clinic produces q_0 amount of output (say, dental patient-hours) with a facility size of 1,200 square feet, as represented by the curve $SATC_2$. This represents a

long-run equilibrium point because the efficient plant size is chosen such that $SATC_2$ is tangent to the LATC curve at q_0; that is q_0 is produced at the lowest possible long-run cost and 1,200 square feet is the efficiently sized facility. Now suppose output sharply falls to q_1, due to a decline in demand. Long-run cost minimization suggests that the dental firm will reduce the size of its facility to that represented by $STAC_1$ and operate at point a on the LATC curve. It might do this by selling the old facility and moving into a smaller one. Because it may take time to adjust to the decline in demand, the dental clinic may not operate on the long-run curve at q_1, (point a) but instead continue to operate with the larger facility as represented by point b on $SATC_2$. The dental clinic incurs higher costs of production as indicated by the vertical distance between points b and a in figure (2.16).

Cowing and Holtmann (1983) derived a test to determine whether firms are operating LR equilibrium. Using a simplified version of equation (2.16), we can write a LRTC function as

$$LTC = STVC (q, w, k) + r*k \ldots\ldots\ldots\ldots\ldots\ldots (2.21)$$

Where all variables are as defined earlier. According to equation (2.21), long –run TC equal the sum of (minimum) short-run TVC and capital costs. The level of short-run TVC is a function of the quantity of output, the wage rate, and the quantity of capital (and other things excluded from the equation for simplification). A necessary condition for long-run cost minimization is that $\Delta STVC/\Delta k = -r$. The equality implies that the variable cost savings realized from substituting one more unit of capital must equal the rental price of capital in long-run equilibrium. That is, the marginal benefits and costs of capital substitution should be equal when the firm is minimizing the long-run costs of production. A nonnegative estimate implies that the cost of capital substitution outweighs its benefit in terms of short-run variable cost savings.

3.4.5 Neo-Classical Cost Theory and the Production of Medical Services

The cost theory described here is a neoclassical cost theory which under conditions of perfect certainty, assumes firms produce as efficiently as possible and possess perfect information regarding the demands for their services. Based on the underlying theory, the short-run or long-run costs of producing a given level of output can be determined by observing the relevant point on the appropriate cost curve. However, when applied to medical firms, this kind of cost

analysis may be misleading for two reasons. First, some medical firms, such as hospital or nursing homes, are not-for-profit entities or are reimbursed on a cost-plus basis or both. Therefore, they may not face the appropriate incentives to produce as cheaply as possible and, consequently, may operate above rather than on a given cost curve. Second, medical firms may face an uncertain demand for their services. Medical illnesses occur irregularly and unpredictably, and therefore medical firms such as hospitals may never truly know the demand for their services until the actual events take place. Accordingly, medical firms may produce with some amount of reserve capacity just in case an unexpected large increase in demand occurs. These two considerations are modifications that should be incorporated into the cost analysis when possible. A strong grounding in neoclassical cost analysis under conditions of perfect certainty is necessary before any sophisticated analyses or model extensions can be properly conducted and understood.

SELF ASSESSMENT EXERCISE

Discuss the long-run cost minimization for health care under the condition of perfect market.

4.0 CONCLUSION

Medical firms earn revenues from producing and selling medical output. The focus of this module is the various economics principles that guide the production process of a medical firm. The SR production function for medical services indicates that the level of medical services is a function of a variable input and fixed input. The production function identifies the different ways variable inputs and capital can be combined to produce various levels of medical services. The production function allows for the possibility that each levels of output may be produced by several different combinations of the variable and fixed capital inputs. Each combination is assumed to be technically efficient, since it results in the maximum amount of output that is feasible given the state of technology. Technical and economic considerations determine least-cost or economically efficient method of production. MP and AP curves can be employed to show the characteristics of the production process. The elasticity of substitution between any two inputs equals the percentage change in the input ratio divided by the percentage change in the ratio of the inputs' marginal productivities, holding constant the level of output.

5.0 SUMMARY

Cost theory is based on the production theory of the medical firm and relates the quantity of output to the cost of production. As such, it identified how (total and marginal) costs respond to changes in output. The reciprocal relation between production and costs focused on the short-run MC and AVC of production. The short-run marginal costs, SMC, of production are equal to the change in total costs associated with a one-unit change in output. A medical firm makes choices about which variable inputs to employ. Knowing that there is more than one way to produce a specific output, medical firms desire to produce with the least-cost or cost-minimizing input mix. The long-run average total cost curve can be derived from a series of short-run cost curves. This cost theory is a neoclassical cost theory under conditions of perfect certainty which assumes that firms produce as efficiently as possible and possess perfect information regarding the demands for their services. Based on this theory, the SR or LR costs of producing a given level of output can be determined by observing the relevant point on the appropriate cost curve. This kind of cost analysis may be misleading when applied to medical firms.

7.0 Tutored-Marked Assignment

University College hospital paid ₦250, 000 for a fiber-optic digital operating suite that it hopes will reduce the cost of performing minimally invasive surgeries. The new technology produces exceptionally clear digital images from a tiny camera placed inside the patient and close to the site of surgery, allowing surgeons to conduct the surgery while looking only at video monitors. The cameras transmit images to monitor that hang above the operating table, thereby reducing the amount of operating room clutter. After installing the new equipment, Sutter observed a 15-minute reduction in average operating room time, reduction in the average time under anesthesia, and quicker patient recovery times. Sutter has performed about 50 operations with the equipment thus far. Discuss how operating room total costs, average costs and marginal costs might change following the adoption of fiber-optic digital imaging equipment.

7.0 References/Further Readings

Folland S., A. Goodman & M. Stano (2010) The Economics of Health & Health Care, Sixth Edition, Prentice Hall, New Jersey.

Jacobs, P. (1991) The Economics of Health and Medical Care Maryland: Aspen Pub Inc. Jack,

Santerre E. & S.P. Neun (1996) Health Economics: Theories, Insights & Industry Studies, Irwin, Chicago.
Zweifel P., F. Breyer & M. Kifmann (2009) Health Economics, Second Edition, Springer Verlag Heidelberg.

Unit 4: Hospital Services and Efficiency

CONTENTS

1.0 INTRODUCTION

Hospital plays a key role in the economic problems of health care. This is due to the quantitative importance of the hospital industry. In many industrialized countries, hospital services account for the largest single block of health care expenditure. In 2005, most OECD countries spend more than a third of health care expenditure on hospital services. However, not the amount of health care expenditure is of primary interest to economists but the rules that determine the allocation of resources. One feature of the hospital sector deserves particular attention: to a large extent hospitals are non-profit institutions. Hospitals may not be interested in operating in a cost-efficient manner. From the point of view of a regulatory agency whose objective is to control hospital efficiency, this creates the need to gather and evaluate information on the production process of hospitals. Since the production function of a hospital is not known, the standard approach is to compare hospitals in a benchmarking study. Several methods have been proposed to conduct such studies. The estimation of a hospital cost function can be done with both parametric and a non-parametric estimation methods. To determine the efficiency of a hospital, the inputs and outputs need to be defined and measured. Inputs comprise the use of productive resources such labour, energy and raw materials. This unit deals with hospital services and efficiency and the willingness to pay as a measure of demand for medical care services.

2.0 OBJECTIVES

At the end of this unit, the students should;

(i) Understand the idea of hospital as a productive unit
(ii) Measure and compare hospital efficiency
(iii) Understand Willingness and Ability to Pay and Access to Health Care

3.0 MAIN CONTENT

3.1 The Hospital as a Productive Unit

3.1.1 Hospital Output: Health as a Latent Variable

In order to measure hospital output, it is not enough to describe the tasks that are carried out (surgery, radiotherapy, medication, wound dressing, and accommodation, etc.) or bundles of tasks such as medical, nursing or hotel services. All of these tasks are only a means to an end. One gets closer to true output by asking the question of what the patients (or referring physicians acting on their behalf) want, what they expect from hospitalization, and what taxpayers expect to obtain in return for their contribution to the financing of the hospital. In the majority of cases, expectations are in terms of a positive contribution to the patient's state of health, that is, the curing of disease (or control of its development) and the alleviation of pain. Even though there is little disagreement with regard to these objectives, the degree of their realization may hardly serve as a basis for the payment of hospital services. The difficulties lie with the measurement as well as the imputation of outputs to services performed.

In order to measure the extent of recovery of patients, their state of health would have to be evaluated not only at the beginning and at the end of hospital treatment but frequently years after the stay, using objective criteria. This is – except for obvious indicators such as survival and complication rates – a fairly hopeless undertaking because health is not only multi-dimensional but also contains a considerable subjective component. Even if this difficulty is surmounted, one should not simply tie payment for hospital services to the measured change in the state of health achieved during the period of hospitalization. For the relevant benchmark for assessing hospital performance is not the patient's actual state before admission, but the (hypothetical) state that would have been realized without hospital treatment at the end of the observation period. The importance of this distinction becomes very clear in cases where hospital treatment can only slow down the progressive course of an incurable patients hold expectations not only with regard to the final state of their health after hospital treatment but also with regard to their

physical and mental well-being during hospitalization (to the extent that the disease will permit), as life goes on in the hospital. This aspect takes on special significance if the disease itself can no longer be fought and only suffering can be alleviated, that is, when dealing with incurable and terminally ill patients. Objective and reliable measurement of subjective wellbeing during hospitalization, however, is just as difficult as measuring the influence of hospitalization on the patient's health status. Finally, the customers of a hospital not just cover people that will actually be treated as patients, but the whole population of its catchment area. The mere existence of a hospital provides people with the security that in case of accident or serious illness inpatient treatment will be available. This so called 'option demand' is satisfied by hospital beds held on reserve, along with the necessary staff and equipment.

3.1.2 The Multi-Stage Character of Production in the Hospital

As the final outcome of hospital activity (particularly the improvement of a patient's health status) can only be measured imperfectly, observable quantities suitable to serve as output indicators have to be identified for obtaining an operational definition of the term, 'efficient use of resources'. In the present context, it seems appropriate to list various indicators of hospital activity and to classify them according to the stage of production, using a scheme that may help to describe hospital activity from an economic perspective. The indicators commonly used are:

(i) Quantities of factors of production (hours worked by physicians, by the nursing staff and by other employees, drugs and dressings, electricity, fuel etc.);

(ii) Quantities of individual medical and nursing services performed (medications, injections, physical therapies, temperature measurements, meals etc.);

(iii) The number of patients or cases treated, possibly differentiated according to various types of diseases

(iv) The number of patient days, possibly differentiated according to intensity of care

3.1.3 The Heterogeneity of Hospital Output

Another problem with defining and measuring the output of a hospital stems from a considerable amount of heterogeneity, be it at the level of output indicators or of intermediate products.

Consider the number of cases treated in a hospital in the course of a year. Is it possible to describe adequately a hospital's outputs imply by counting the number of cases? Is it appropriate to say that a hospital that treats 1,000 cases achieves more than the one that treats 995? Obviously, the 1,000 cases of the first hospital might consist of 500 minor fractures and 500 uncomplicated tonsillectomies whereas the second hospital might be a heart unit specializing in transplantations. One must therefore take into consideration that the 'case treated' does not constitute a homogeneous quantity but a mental construct that needs to be specified through its characteristics. This means that a treatment case must be differentiated along various dimensions, for example:

(i) The type of illness that has called for hospital treatment (principal diagnosis);

(ii) The severity of the illness and complications arising during treatment;

(iii) The stage of the disease (e.g., in the case of cancer);

(iv) Concomitant diseases (secondary diagnosis);

(v) Patient characteristics reflecting her or his contribution to the 'production of recovery', such as age and possibly sex.

In view of these distinctions, to which still more could be added, a purist must come to the conclusion that heterogeneity of case mix can only adequately be taken into account by considering each patient as an output category of her or his own. Following this principle, however, one would forego the possibility of comparing the output vectors of two or more hospitals. This would put an end to the economic analysis of the hospital in the quest of, for example, measuring the degree of efficiency or determining a performance-based payment system. A reasonable compromise between the rigorous approach just outlined and the total abandonment of case differentiation may be achieved by dividing patients into a manageable number of groups, using the distinguishing characteristics mentioned above. This division is known as patient classification and purports to form a manageable number of patient groups that are as homogeneous as possible. Moreover, assignment to a group should be unique and reproducible using objective criteria. Obviously, there is a conflict between the criteria 'manageable number' and 'greatest possible homogeneity within each group' which can only be

resolved by weighing the relative disadvantages caused by their violation. The three most common patient classification systems are:

(i) The International Classification of Diseases (ICD) originally developed as a basis for mortality statistics and thus solely referring to (principal) diagnoses. In its three-digit version, the ICD consists of more than 900 groups, while aggregation to the level of 110 main groups is already very coarse. For instance, all benign neoplasms form one single group in this categorization.

(ii) Diagnosis Related Groups (DRGs) developed at Yale University in the 1970s with the explicit objective to create relatively cost-homogeneous groups. Besides the principal diagnosis, DRGs take into account the existence of concomitant disease and complications, the age of the patient, and the type of treatment (surgical or conservative), getting by with approximately 500 groups.

(iii) Patient Management Categories (PMCs), also developed in the United States (Pittsburgh) and consisting of a total of 840 groups compared to DRGs, PMCs put emphasis on concomitant diseases and the treatment strategies chosen by the hospital.

SELF ASSESSMENT EXERCISE

Discuss the indicators of hospital activity and classify them according to the stage of production.

3.2 Hospital Efficiency

3.2.1Regulation and Asymmetric Information

In a competitive market, there is no reason to measure the efficiency of firms. Only efficient firms make sufficient profits to cover their costs. The market for hospital services is different. Even though competition has been introduced in the hospital sector in a number of countries, hospitals continue to be subsidized by tax-payers undermining the incentives for efficiency. In addition, market entry is difficult because of public regulation (such as participation in the supply emergency services and capacity limitation schemes) and political opposition of public hospitals, who want to avoid competition. Conversely, market exit is frequently a highly unpopular political decision, in particular for public hospitals. For the majority of hospitals, their economic environment therefore cannot be expected to create incentives for efficiency.

An alternative approach to ensure the efficiency of hospitals is regulation. A regulatory agency that is fully informed about a hospital's technology and efforts to control costs could easily achieve this objective. But providers typically have an information advantage. Asymmetric information can exist both with respect to efforts to control cost ('hidden action') and with respect to the technology used ('hidden information'). The 'new economics of regulation' which applies the principal-agent methodology to the contractual relationship between regulators and regulated firms has shown how the presence of asymmetric information prevents regulators from implementing an efficient ('first-best') solution since efficient providers have both incentives and possibilities to imitate inefficient providers in order to get a favourable deal, allowing them to gain an information rent. Faced with this problem, regulators have two options:

(i) They can design contracts to obtain a second-best solution, striking a compromise between the reduction of the information rent and incentives for providers to control costs.

(ii) The regulatory agency can try to alleviate the information asymmetry which would limit the hospitals' possibilities to obtain information rents. In particular, information about the efficiency of a hospital can be gained by comparing it to other hospitals. Economists have developed two basic methods to conduct such a comparison. These are estimation of a hospital cost function and a nonparametric approach called 'Data Envelopment Analysis' (DEA).

3.2.2 Hospital Cost Functions

Considering the hospital as a productive unit suggests the application of elementary concepts of production theory to hospitals. One of the core concepts of production theory, which has great empirical relevance, is the cost function. It assigns minimal costs of production to each output bundle and contains the same information as the production function. The quantities on the right-hand side of the cost function are factor prices and output quantities. These variables – and in the case of a short-term cost function, the amount of fixed factors – may be considered exogenous, provided perfect competition prevails in the factor markets. This makes the cost function more easily amenable to econometric estimation than the production function which depends on input quantities. As these are chosen by the firm, they cannot be considered as

exogenous. The empirical estimation of cost functions is helpful for answering a number of important economic questions:

(i) From the shape of the cost function, one may recognize economies or diseconomies of scale, which are important for determining the optimal size of the unit. Optimal size is of relevance to policy because in some countries the hospital industry is subject to public regulation which prevents hospitals from freely choosing their number of beds.

(ii) Differentiation of the cost function with respect to the number of patients of a given category yields the marginal costs of treatment of this patient type. This information is useful for calculating prices in the context of performance-based payment.

(iii) Residuals of an estimated cost function are the difference between a hospital's actual and estimated cost. From their size, one may derive statements about their efficiency. This may facilitate the monitoring of hospital performance under cost-based payment.

Following Baumol et al. (1982), a cost function should satisfy the following requirements:

(i) As predicted by microeconomic theory, the cost function should be continuous, linear-homogeneous, non-decreasing and concave in prices.

(ii) In the multi-output case, the cost function should allow for zero values for some outputs lest all hospitals failing to produce all outputs must be dropped from the sample. This may lead to a sample selection problem. In addition, it would not be possible to derive results about economies of scope.

(iii) The cost function should be sufficiently flexible. It should allow for different degrees of economies of scale and scope. Such a flexible cost function is the translog cost function which is a second order Taylor-approximation of an arbitrary cost function.

Most importantly, the estimation of a microeconomic cost function proceeds on the assumption that the hospitals included in the sample all aim to minimize costs and that deviations from this target have the same property as an error term. This assumption is hardly sustainable for hospitals due to the predominance of government and community ownership, since the absence of the profit motive serves to diminish the pressure to minimize costs. To deal with the problem

of non-cost-minimizing firms, stochastic frontier estimation has been developed. This approach specifies the estimation equation for total hospital cost as

$$C_i = C(Y_i, W_i) + V_i + U_i \qquad (2.22)$$

Where: C_i denotes total cost of hospital i, Y_i a vector of output quantities, W_i a vector of input prices, V_i a normally distributed error term which measures statistical noise such as measurement error in the cost variable, and Ui a positive term which measures 'errors' in decision making which result in inefficiency and thus higher than minimum costs. To estimate equation (2.22), an assumption must be made concerning the distribution of the positive error term (such as half-normal, truncated normal or exponential) so that the equation can be estimated using maximum likelihood methods. In a second step, the inefficiency measures retrieved from this estimation can be regressed against various characteristics of the hospital and its market such as ownership type, teaching status, and the number of competing hospitals within its catchment area, to explore sources of hospital inefficiency. In estimating a cost function such as equation (2.22), a number of econometric issues have to be addressed:

(i) Choosing total cost as the dependent variable leads to heteroskedasticity, if the error term is positively correlated with the output variable. Furthermore, the regression can suffer from multicollinearity among the output variables if these vary directly with hospital size. To avoid these problems, both sides of the equation can be divided by an appropriate output measure, making average cost the dependent variable. This makes the output variable appear on both sides of the equation so that the estimation may be biased. When hospitals are considered as multi-product firms, it is unclear which output category should be used.

(ii) An appropriate functional form must be chosen for the function $C(\cdot)$. Candidates are the translog or the homothetic form, which are both flexible functional forms. A disadvantage of the translog function is that it does not allow for zero output and can only be used with broad output categories.

(iii) Hospitals are multi-product firms with several hundreds of different outputs even if patients are only classified by diagnosis-related groups. Especially with flexible functional forms the number of regressors (which is on the order of n^2 when there are n outputs) can

easily exceed the number of observations. Thus in past applications, the available output information was usually condensed into a small number of output categories such as the number of cases in the various hospital departments or just the number of inpatient and outpatient cases. In the latter case, often a scalar case-mix index (such as the average DRG weight) is added to control for output severity.

(iv) A further problem is the availability of information on factor prices which ideally should be used in the estimation of a cost function. This information is difficult to obtain in the health care sector.

SELF ASSESSMENT EXERCISE

Discuss econometric issues that need to be addressed in estimating hospital cost function.

3.3 Willingness and Ability to Pay and Access to Health Care

With respect to willingness and ability to pay, two postulates are commonly made when it comes to access to health care services. The first says that access should not depend upon ability to pay, i.e., income or wealth of a person. The second goes even further by requiring that not even willingness to pay should play a role, i.e., the amount of money that a person is willing to give up for the service. Instead medical criteria alone are to govern access to health care. To evaluate these hypotheses first note that willingness to pay simply reflects preferences for health as opposed to other goods if two persons with the same ability to pay are compared. To exclude willingness to pay as a criterion for access to health care services implies that preferences must not play a role. Given equal ability to pay, this is ethically questionable and not compatible with the idea of a liberal society. Thus, if one were to agree on excluding willingness to pay, one would want to limit application of this principle to persons who differ in their ability to pay. The mere fact that ability to pay differs between patients is not sufficient reason to ban it as a criterion governing access to health care services. The crucial point is whether differences in ability to pay are to be considered as unjust. To answer this question, one may consider the factors causing such differences, in particular (a) unequal personal effort, (b) unequal initial opportunities, (c) unequal ‘luck’ in life.

3.3.1 Health Goods, Market Failure and Justice

If only unequal personal effort were responsible for the differences in ability to pay, then there would be no reason to call the distribution of ability to pay unjust. A different verdict obtains if unequal initial opportunities and luck in life are predominant. In reality, it is likely that all three factors play a role, justifying a policy of redistribution in principle. But this does not imply that the access to health care services should be made completely independent of ability to pay. Rather, it seems more appropriate to treat ability to pay as such the primary target variable of social policy by paying transfers to the lowest income groups. To get political support from taxpayers, tying these transfers to the purchase of (social) health insurance may be advantageous, which guarantees access to health care services. On the other hand, if ability to pay is to be completely disregarded as a criterion for access, there are only two ways to accomplish this. First, one can try to completely equalize ability to pay. But this alternative seems hardly desirable because it involves high efficiency losses through taxation and transfers that cause important distortions in economic decisions, particularly with respect to labor supply and capital formation. Furthermore, it can be argued that enforcing an equal distribution of income and wealth gives rise to injustice to the extent that differences in income reflect differences in effort and services provided to other members of society, rather than the other two factors cited above. Secondly, one can attempt to suppress willingness to pay and thus ability to pay as a determinant of access to health care services. This idea of specific egalitarianism has been advocated by Williams (1962). A consequence of this view is that personal effort does not serve to obtain more health care services. A key argument in favour of specific egalitarianism is that in emergency situations involving life and death, ability and willingness to pay often coincide. In the face of death, however, citizens should be equal. Another situation calling for specific egalitarianism is an incident of a scale so large that available resources do not permit all victims to be treated. In such a situation, only an allocation of scarce resources according to medical criteria, in particular urgency and chance of survival, is deemed ethically acceptable. Failure to enforce specific egalitarianism would result in ability to pay only to decide who obtains treatment. An important question is how frequently such ‘life or death’ situations occur. More importantly, resources available for health care are not exogenously given, but determined

by demand and supply. For example, if the demand for physiotherapy by people with high ability to pay rises, the market mechanism will lead to increase in supply of physiotherapeutic services, through more services provided by existing physiotherapists or entry of new providers. The quantity of services and with it consumer surplus rises. These beneficial effects cannot be achieved if willingness to pay is excluded as a determinant of the allocation of health care.

3.3.2 Justice as an Argument in Favour of Government Intervention in Health Care

There are additional arguments against the concept of specific egalitarianism with respect to health care,

(i) Copayments create incentives for health-related behaviour. However, they are more easily borne by individuals who have a high willingness to pay. This means that copayments are not compatible with specific egalitarianism and must be ruled out if willingness to pay is to be excluded from the criteria determining access to health care services. As a consequence, people will ignore the financial consequences of their nutrition, exercise, smoking and drinking habits. To avoid an explosion of health care expenditure, the government would have to control health-related behaviour through compulsion, thus risking a conflict with basic values of a liberal society.

(ii) Medical services are not the only goods that have an impact on health. Other things such as adequate nutrition and housing play a comparable role. Thus, they should be allocated free of charge according to specific egalitarianism. Withdrawing such a wide spectrum of goods from the discipline of market mechanisms, however, would jeopardize the allocative efficiency of the economy as a whole.

(iii) The freedom of patients to decide on their own matters would be curtailed since decisions to treat would be based solely upon criteria emanating from a collective decision.

For these reasons, proposals to exclude willingness to pay as a criterion governing the access to health care services appear to be misguided. Only in emergencies where it is impossible to treat everybody in an adequate way does willingness to pay constitute an ethically questionable criterion for the allocation of health care services. In all other situations, the health care system is not a zero-sum game in which the well-to-do impose their demands at the expense of the

poorer members of society. Therefore, if ability to pay is inequitably distributed, it is a better strategy to influence it directly through taxes and transfers in order to guarantee an adequate provision of health care services for all citizens.

SELF ASSESSMENT EXERCISE

Write short not on willingness to pay as a criterion for access to health care.

4.0 Conclusion

Hospital output consists of improving or maintaining the patient's state of health on the one hand and the capacity to satisfy an option demand on the other hand. The former part of output is particularly difficult to operationalize and can only partially be attributed to the hospital. A hospital's 'output' can be described as the outcome of a multi-stage process, with each stage being assigned its specific concept. Patient classification systems try to do justice to the heterogeneity of the hospital output while making comparisons between hospitals possible. All systems seek to describe hospital output in some detail, if not with regard to treatment outcomes, that is, the improvement of health status, at least with regard to the difficulty of the task. The stochastic frontier approach to measuring hospital efficiency is adequate if the data is subject to measurement error and stochastic influences. As hospital output and quality are difficult to measure, it is problematic to simply equate the error term of this estimation with inefficiency.

A general exclusion of ability to pay or even willingness to pay from the criteria governing access to medical services is not desirable as it runs counter to the principles of a liberal society and would lead to an important loss of efficiency. Differences in ability to pay due to factors that are deemed unjustified can be addressed by taxes and income transfers. Only in emergency situations where a fixed amount of resources is available which is insufficient to treat everyone affected should willingness to pay be neglected in the allocation of health care services.

5.0 Summary

Hospital is a key element in the economic problems of health care due to the quantitative importance of the hospital industry. In many countries, hospital services account for the largest single chunk of health care expenditure. To determine the efficiency of a hospital, the inputs and

outputs need to be defined and measured. Inputs comprise the use of productive resources such as human labour, energy and raw materials. In order to measure hospital output, it is not enough to describe the tasks that are carried out (surgery, radiotherapy, medication, wound dressing, and accommodation, etc.) or bundles of tasks such as medical, nursing or hotel services. This unit takes a look at hospital as a productive unit, the heterogeneity of hospital output, hospital Efficiency, hospital cost functions, willingness and ability to pay and access to health care, health goods, market failure and justice.

6.0 Tutor Marked Assignment

1. Discuss the indicators of hospital activity and classify them according to the stage of production.
2. Discuss econometric issues that need to be addressed in estimating hospital cost function.
3. Write short not on willingness to pay as a criterion for access to health care.

7.0 References/Further Readings

Donaldson Cam and Karen Gerard (1993) Economics of Health Care Financing: The Visible Hand. Macmillan Press Ltd. London.

Folland S., A. Goodman & M. Stano (2010) The Economics of Health & Health Care, Sixth Edition, Prentice Hall, New Jersey.

Jacobs, P. (1991) The Economics of Health and Medical Care Maryland: Aspen Pub Inc. Jack,

Jones Andrew (2007) Applied Econometrics for Health Economists: A Practical Guide, 2nd Edition OHE

Santerre E. & S.P. Neun (1996) Health Economics: Theories, Insights & Industry Studies, Irwin, Chicago.

Zweifel P., F. Breyer & M. Kifmann (2009) Health Economics, Second Edition, Springer Verlag Heidelberg.

MODULE THREE: COST CONCEPTS, ECONOMIC EVALUATION AND HEALTH CARE FINANCING

Unit 1: Costs Concepts and Economic Evaluation
Unit 2: Health Care Financing
Unit 3: The Role of Government in Health Care

UNIT 1: Costs Concepts and Economic Evaluation

CONTENTS

1.0 INTRODUCTION

Economists define a cost as the value of resources used to produce a good or services. However, the way these resources are measured can differ. There are two main alternatives with respect to measurement of these resources: financial and economic costing. Financial cost represents actual expenditure on goods and services. Costs are described in terms of how much money has been paid for the resources used in the project or services. In order to ascertain the financial costs of a project, we need to know the price and quantity of all the resources used or the level of expenditure on these goods and services. Economists conceptualize costs in broader way. They define costs in terms of the alternative uses that have been forgone by using resources in a particular way. These economic or opportunity costs recognize the cost of using resources as these resources are unavailable for productive use elsewhere. The basic ideas are that things have a value that might not be fully captured in their prices. This unit discusses costs in health care services and health programmes and looks at its implication for economic evaluations.

2.0 OBJECTIVES

At the end of the unit, the student will be able to:

(i) understand the meaning and basis of cost as a concept.

(ii) be aware of the possibility of using cost concepts to undertake economic evaluation.

(iii) revisit the conceptual meaning of opportunity costs.

(iv) describe direct, indirect and intangible costs.

3.0 MAIN CONTENT

3.1 Costs Concepts

It is not difficult in many health programmes to identify resources inputs for which little or no money is paid: volunteers working without payment; health messages broadcasts without charge; vaccines or other suppliers donated or provided at large discount by organizations or individuals. Thus, the values of these resources to society, regardless of who pays for them, are measured by opportunity cost. Economic cost then includes the estimated value of goods or services for which there were no financial transaction or when the price of a specific good did not reflect the cost of using its productivity elsewhere. The main ways that financial and economic costs differ is in the way they treat:

(i) Donated goods and services

(ii) Others inputs whose prices are incorrect or distorted.

(iii) Valuation of capital items

The theory and the concept of cost arise from the fact that economic resources are scarce by nature. Had it not been for the scarcity of resources, the concept and theory of cost may not exist as such. Scarcity has two sides:

(i) The infinite nature of human wants

(ii) The finite or limited nature of resources available to produce goods and services.

3.1.1 Types of Costs

Costs can be defined in many ways (See figure 3.1 below), but generally can be considered as direct, indirect and intangible. Direct costs are those immediately associated with an intervention

such as staff time, consumables etc. Indirect costs might include a patient's work loss due to treatment. Intangible costs may be things like pain, anxiety, quality etc. Benefits, however, can be analyzed in three different ways reflecting the different types of economic analysis used in evaluation. First, benefits can be examined in terms of the immediate (direct) effects on health. These are usually clinically defined units appropriate to the area of study, such as 'lives saved', 'reduction in tumor size', 'change in blood pressure' etc. Second, benefits from an intervention can be considered in more generic terms such as the impact on general well-being/ happiness/ satisfaction, these are more generally labeled as 'utilities'. The utility of an intervention to an individual is its benefit. Measures such as the quality adjusted Life year (QALY) are used to quantify this third way. Benefits might be considered in the same terms as costs, which mean that benefits must be valued in monetary terms by some means.

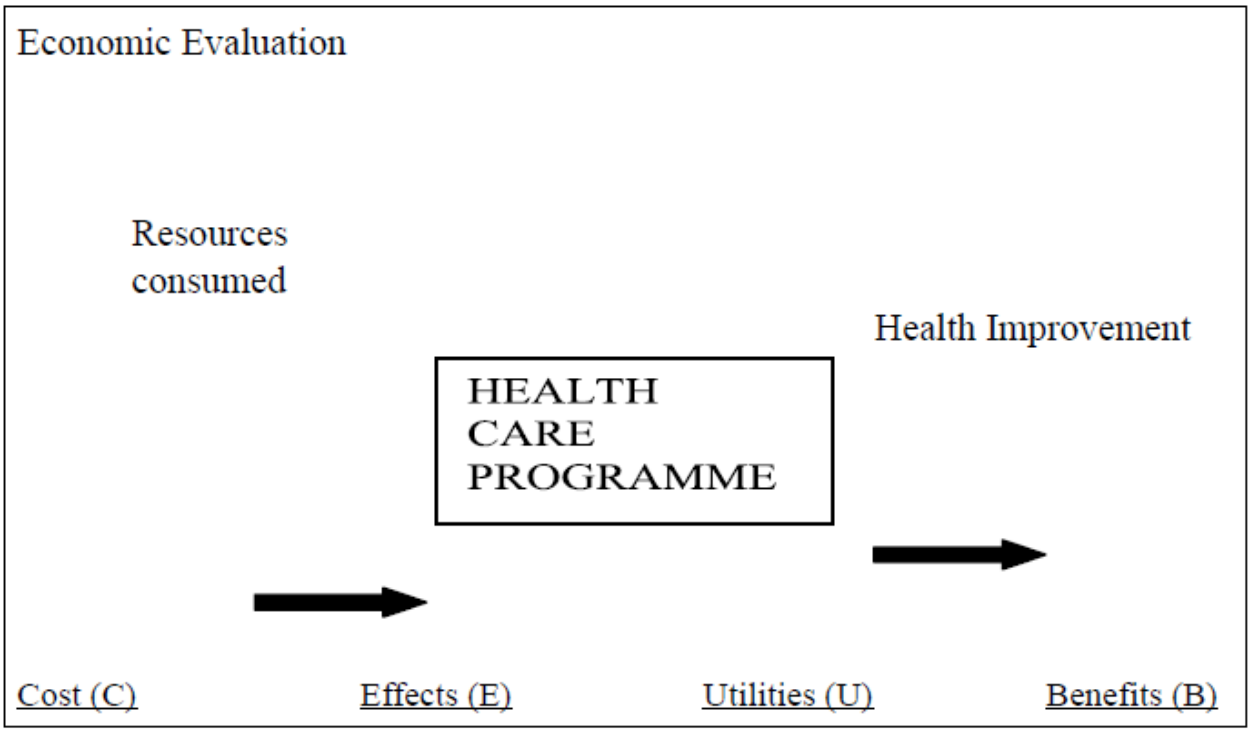


Figure 3.1 Evaluating Costs and Consequences
Sources: Drummond et al., 1997

Whatever kind of economic evaluation may be applied, the costs must be assessed. These are divided into costs borne

(i) by the ministry of health (like drug and equipment),

(ii) by patients and their relatives (like transport and food) and

(iii) by the rest of society (like health education).

The costs have to be valued in monetary terms:

(i) Direct costs, like wages, pose little problem.

(ii) But indirect costs (like time spent in hospital) have to have values imputed to them.

(iii) Costs must also be further subdivided in to average, marginal and joint costs, which help decisions on how much of a service, should be provided.

(iv) Capital costs (investment in plant, buildings, and machinery) are also important for due consideration, as discounting and inflation.

3.1.2. Inclusion of Costs

If the evaluation is being made from the widest perspective the viewpoint of society as a whole-then three main categories of costs must be considered:

(i) Health service costs: These will include staff time, medical supplies (including drugs), bed and food services in the case of inpatients, use of capital equipment, and overheads such as water, heating and lighting. These items may be divided into variable costs, which vary according to the level of activity (for example, staff time) and fixed costs, which are, incurred whatever the level of activity (for example, heating and lighting). In the long run, practically all costs become variable since those that are fixed in the short run may be varied-for example, by opening and closing wards, and by building new hospitals. In economic evaluation all such health service costs-both fixed and variable-are referred to as direct costs.

(ii) Costs Borne by Patients and their Families: These will include out of pocket expenses such as travel, and any cost resulting from caring activities undertaken by the family. These are both direct cost items. In addition, there may also be indirect costs (productivity costs) such as income lost because of absence from work (which is a production loss to society) and any psychological stress experienced by patients, or their families or both.

(iii) External costs: These occur when people not directly involved in a programme experience increased costs because of it. In most cases these effects are too small and diffuse to merit inclusion in the analysis, but there may be some occasions when they are large enough to require attention. For example, public health legislation enforcing anti-pollution standards or specifying water purification levels may lead to increase in manufacturing costs and consumer prices (as well as providing health benefits).

3.1.3 Valuation of Costs

How should costs be valued? Adequate valuation of costs must consider the following:

(i) The costs identified in physical units (such as hours of staff time, hours of operating theatre use, quantities of drugs and so on) must be valued in monetary terms.

(ii) For most direct cost items market prices will be available.

(iii) Nursing time can therefore be valued at the appropriate hourly rate;

(iv) Medical and surgical supplies can be valued at the prices charged by suppliers;

(v) Electricity and water can be valued at the appropriate tariffs; and so on.

Strictly speaking, economic evaluation should seek to value all inputs in terms of their opportunity costs-that is, their value in their next best use.

(i) These measures what is being given up to use resources in health care

(ii) Sometimes opportunity costs may diverge from market prices. Example, a nurse would otherwise be unemployed, and then his or her opportunity cost would be zero and not the hourly wage.

(iii) For most practical purposes, however, it is usual to use market prices unless there is strong evidence to suggest that they diverge appreciably from opportunity costs.

Indirect costs, for which there are no market prices, pose a more difficult problem of evaluation. Some method has to be used to impute values to them.

(i) This is known as “shadow pricing”, and time costs provide a good example.

(ii) When time is spent in hospital by a patient, or on caring by a relative, and this displaces work time, it is usual practice to use the relevant wage to value the lost time.

(iii) If it is not work time that is displaced, however, other measures must be used.

3.1.4 Average, Marginal, and Joint Costs

Most decisions in health care are not concerned with whether or not a service should be provided, or whether or not a particular procedure should be undertaken, but with how much of the service should be provided. That is, should existing levels of provision be expanded or contracted? For example, what family planning services should be made available? This decision requires that attention should be focused on marginal costs-that is, the change in total costs resulting from a marginal change in activity. In the short run, there is often an important difference between the marginal costs of an activity and its average cost, where the average cost is defined as the total cost divided by the total number of units of output. One context in which the distinction between average and marginal costs is important is in relation to duration of hospital stay of inpatients. Many new procedures have reduced the amount of time necessary for a patient to remain in hospital and thereby yield cost savings. When valuing these savings, however, it is important to bear in mind that using average costs /day will generally overstate the savings as the later days of a stay usually cost less than the earlier ones. It is the marginal costs/day that is the relevant measure. Yet another problem of cost measurement arises in connection with joint costs. Often a single production process can result in multiple outputs. For example, a single chemical analysis of a blood sample can diagnose the presence of many diseases. How should the cost be allocated to each diagnosis? Similarly, within a hospital setting, there are many common services (like medical records, radiology, operating theatres, laundry, catering, and cleaning) that contribute to a number of specialties. Economic evaluation requires some method for allocating the joint costs of these services to individual programmes or procedures. There are several methods, which may be used to do this. Most of them use some physical unit of utilisation, like the number of laboratory tests, hours of operating theatre use, or square meters of ward space, to allocate total laboratory and ward cleaning costs.

3.1.5 Capital Costs

Investments in buildings, plant, and equipment that yield a flow of services over a number of years give rise to capital costs. Generally, investment expenditure will be undertaken at the beginning of a project, but the use of items of capital equipment will generate annual capital costs over the lifetime of the asset. These costs have two elements vis: interest and depreciation.

- Interest costs should be included even if the asset was not acquired with borrowed money because tying up money in an item of capital equipment involves an opportunity cost-that is, interest foregone.
- Depreciation costs arise because of the wear and tear that an asset gets through use and the consequent reduction in the length of its life. (But land is a capital asset that is not assumed to incur depreciation costs).

Sometimes an item of capital expenditure is unique to a particular use and has little or no alternative use value (opportunity cost). In such cases, it is referred to as sunk cost. A hospital building or an item of medical equipment may, for example, have considerable value in its existing use, but little resale value. This can provide a powerful case for continuing to use existing assets instead of undertaking new investments because, in an economic evaluation, sunk costs should not be included among annual capital costs. In practice, this consideration is likely to be more important in the case of major capital developments than of individual procedures.

SELF ASSESSMENT EXERCISE

Discuss types of cost in economics and show their practical application to the health sector.

3.2 Issues in the Measurement of Costs

There are two practical problems in economic evaluation – measuring costs and measuring benefits. Measuring costs appears to be easier than measuring benefits. In almost all organisations there are some attempts to measure costs, if only for the purposes of financial control and accountability. But measuring costs accurately is often difficult, and there are important conceptual and practical problems to overcome. Cost in economics is the opportunity foregone, and in economic evaluation the aim is to measure opportunity cost. This implies that cost estimates produced for the purpose of economic evaluation will not be applicable to all

possible other purposes that cost estimates might serve. For example, if a health service manager is interested in the financial implications of the introduction of a new service; or even the maximum effect (in health or welfare terms) achievable for a given expenditure by the health service, a separate analysis of this would need to be commissioned. The justification for the choice of opportunity cost is that it takes into account the costs of all members of society, as they impinge on the social welfare function, and is consistent with the attempt to measure benefits in the same way. Alternative conceptions of cost, which are not consistent with benefit measurement, can lead to illogical conclusions. Nevertheless, if we believe that health services are 'under-funded' relative to other parts of the economy we might mean that money in health service hands has a higher value than elsewhere. This might provide a justification for analysing the costs to the health service only, of achieving health improvement. To measure opportunity cost, we need to know the context in which choices are made. Good costing exercises start from a clear understanding of how current or potential services operate, what resources are used for particular groups of patients, which are shared, and how the staff spend their time. For example, in one study that was assessing the cost of a long-stay mental health facility, it was found that the staff actually spent most of their time caring for a small group of patients with the most serious problems. This meant that very little direct support was provided for the majority of residents. The actual use of the staff resources, and therefore the way in which costs varied between different groups of residents, would not have been observed if costs had been calculated only from accounting data. It is likely that costs would have been assessed as being the same for all residents in the facility. A good understanding of current provision can also help to identify if there is any spare capacity, which would allow the service to be expanded at low cost. It can be important to assemble information about the choices of technology and organisational structure available at different scales of provision. It is unusual for it to be possible to assess the costs of new developments accurately without quite detailed knowledge of the technology, management and human skills needed.

Some costs are fixed; some can be changed but only slowly. Some elements of cost can be easily observed and are obviously related to a particular activity, but others, especially buildings and

land, senior staff, equipment and administration, may not vary directly with the level of activity, and it may be difficult to apportion these costs. The simplest approach to calculating costs looks in detail at all the inputs into a service, multiplies by the unit cost of each, and thereby calculates total cost. This can present an accurate account of the direct costs of a particular service, although overhead costs may be hard to allocate. A drawback of this approach is that it does not demonstrate clearly how costs are likely to behave in the event of changes in scale, case mix or technology. There is therefore an argument for trying to estimate cost functions from information on costs and outputs in a larger number of service providers. These data are analysed using statistical methods to identify how costs vary with the level and mix of output, and to identify the factors that affect costs. It is quite common for costing exercises to use a mixture of approaches, since there are usually constraints on access to appropriate data. In some studies a cost-function approach is used to calculate the unit costs to be applied to activity data.

Costing is not a simple technical exercise – it is too important to leave to accountants alone. Understanding the services provided as well as the financial data and analysis are important. Some examples may help. Many health interventions exhibit economies of scale, so that increasing the output may allow a lower cost service to be developed. In these circumstances the average cost based on current services will overestimate the costs of an expansion. Most emergency services exhibit economies of scale due to the possibility of using the capacity more intensively. An example is neonatal care. Since the need for such services is inherently unpredictable, most centres aim to keep at least one bed free at any time. The proportion of empty beds is therefore lower when there are fewer, larger centres. In cases like these it would be very misleading to cost services at the average if changes in the scale of provision were being contemplated. Other services are less likely to show significant scale economies, such as palliative care for people dying of cancer, and many parts of primary and secondary care. In such examples, the advantages of centralisation are often balanced by the disadvantages to patients of greater geographical distance between home and facility. It may take time to adapt to higher levels of output, so scale economies may not be realised promptly. The technology used may be lumpy (e.g. bits of equipment come only in certain sizes, so that expansion beyond a

certain threshold requires a large additional investment). New approaches to provision of certain services may involve a large change in the scale of provision.

Tradition plays a large part in how services are organised. Many patterns of care owe more to historical accident than careful and rational planning. This means that it is important to understand what inputs are really necessary. For example, immunisation schedules may be rationalised, grouping a number of vaccines within a single administered dose. Among the reasons for doing this might be a more efficient use of staff time, but it is unlikely that staffing levels of an immunisation programme will be immediately adjusted. After some time, it might be apparent that there is a little slack in the immunisation programme than in another programme, and staff might be reallocated. There are several reasons why costs of care for different patients may vary.

(i) First, there may be characteristics of the patient that lead to longer hospital stays or more interventions, and therefore higher costs.

(ii) Second, habits and traditions in different hospitals can differ, and this can lead to variation in cost. For example, the policy may be to keep all emergency admissions in hospital for at least one day, or alternatively it could be to review all patients admitted and discharge those not receiving active investigation or treatment. Given the somewhat limited evidence about the outcomes related to different patterns of care, it can be difficult to impose lower cost (but possibly lower quality) approaches to care.

(iii) Third, the providers may have different levels of technical efficiency, and so any given service will have a different unit cost. If we are seeking the opportunity cost, in principle we should be interested in identifying the lowest feasible cost of providing a given service. Differences that are explained by patient characteristics must be taken into account. It is less obvious how we should treat different clinical policies – normally they vary most where the evidence is weakest, and we often do not know if lower-cost practices reflect greater efficiency or lower quality. In principle the opportunity costs should not allow any X-inefficiency (technical inefficiency), so we should try to identify the cost in an efficient care provider.

Technically, cost that results from inefficiency is not a part of opportunity cost. Simply by using the resources efficiently, it is possible to increase welfare. However, if we are convinced that it is impossible to eliminate X-inefficiency within a particular time scale, then it may be appropriate to include some element of inefficiency in the estimates of cost. In this case, in practice these are the minimum costs of providing the programme, in the short term at least.

Identifying the appropriate concept and measure of cost can be particularly difficult when economic evaluation is carried out as part of clinical trials and studies. Patients recruited into a study are normally heterogeneous, and some variation in costs is likely. Normally they are not completely typical of patients who are likely to receive the treatment. Large trials normally recruit patients from many centres, and clinical policies and efficiency will also be important. To assess the cost-effectiveness of a new intervention we need to calculate the costs for those patients likely to be provided with the service, in the ways and places they are likely to receive the service. It is important therefore to know how costs vary with such factors as age, sex, disease severity, co-morbidities, case mix and scale of provision. To do this properly we need large enough samples of patients for variations to be understood, and for it to be possible to calculate confidence intervals for the estimates of cost. Since little is currently understood about the patterns of costs and these factors, it is not yet easy to estimate ex ante the sample sizes needed for costing studies, but it is clear that in some cases the variations are large. There has been widespread criticism of the lack of data on confidence intervals in economic evaluation studies, and a range of methods is often applied to estimate these.

Ideally costing studies calculate unit costs of services from a range of settings, but this is not always feasible. Where a range of different providers show very different unit costs, and these are not explained by characteristics of patients or effectiveness of treatment, there is an interesting issue of which estimate should be used. Since opportunity cost is the objective, there is a case for choosing the lowest observed estimate of unit costs, as discussed above. However, differences in unit costs by institution may not only reflect differences in technical efficiency as this perspective suggests. Since interventions are administered through a given infrastructure,

there is a need to match the ideal infrastructure for this particular intervention and the ideal infrastructure for the health system as a whole. This intervention may be efficiently delivered in medium-sized health centres, whereas others are most efficiently delivered in small or large ones. Additionally, health infrastructure as a whole has to balance technical efficiency questions from a health service perspective with patient access costs. Patterns of human settlement do not present standard problems capable of producing a single 'best' solution to health unit size.

On the assumption that the most important determinant of cost variation between units is technical efficiency, some costing studies use data envelopment analysis (DEA) or stochastic frontier techniques, which aim to show costs of the most efficient care providers. Both these approaches aim to estimate cost functions in terms of the lowest observed costs rather than as the average of those observed. DEA is a non-parametric technique, and simply joins up the lowest cost observations to describe the function. Since there is likely to be measurement error there are advantages in using a method that takes this into account. Stochastic frontier analysis aims to do this. Using these techniques the relative efficiency of different hospitals can be estimated by comparing observed cost with the lowest observed cost for a comparable provider. A typical measure of relative efficiency is the ratio of the cost of a service and the cost of the lowest cost observed service.

As with all statistical methods of estimating costs, the concern must be to ensure that differences in case mix are properly controlled for, so that the lowest observed cost is genuinely an example of greater efficiency and not simply the result of easier cases. With that proviso, there are many advantages in frontier methods to estimate cost. The estimate of cost is the lowest for a comparable provider and should therefore contain less X-inefficiency than the average provider. Thus, the frontier estimate can be viewed as being closer to opportunity cost than the average unit cost for all providers. Of course, such techniques can only identify relative efficiency, since the comparison is with the most efficient observed provider, and not with one that is necessarily efficient in absolute terms.

The costs we are aiming to estimate will usually be associated with adding a new service, or expanding an existing service. Where we are expanding an existing service, whether increasing the level of activity within a unit, or expanding the service from one set of units to others, information derived from cost functions can be very helpful. When a cost function is estimated it can be used to identify costs at higher or lower levels of output, and with different mixes of cases. By comparing the costs at the present level of activity with the costs at the level after implementation of an expanded service, we can obtain estimates of additional (or incremental) cost. The incremental cost is a similar concept to marginal cost, but in this case the change in service volume may not be small.

Where we are adding a new service – which is not yet provided anywhere in the health system – existing cost data are probably not very useful, and we are likely to be evaluating experimental provision (as where the economic evaluation is attached to a clinical trial – see below), or building up a hypothetical picture of costs. Nevertheless, our interest is still in incremental cost. Whereas, when we are expanding an existing programme, economies of scale cause divergence between average and incremental cost, when we are introducing a new programme, economies of scope cause this divergence. In principle a focus on incremental cost is useful since in assessing options we really want to compare differences in costs and benefits between options. When costs are estimated using measures of changes in activity and a vector of (average) unit costs, the estimated costs or savings are likely to be over or under estimates of incremental cost. It is, of course, possible to make adjustments to the unit cost vector to reflect any economies of scale or scope, and therefore to derive estimates that are closer to incremental cost. If we know the change in output associated with a development, the incremental cost (calculated from a cost function or from a hypothetical model of a new activity) can be used in the cost vector in place of average cost. There is continuing controversy about the best estimates of incremental costs. In the short run capital costs are not relevant to measuring incremental costs, since there will be no change in capital (and other fixed) costs. If the current service has excess capacity, then there may be little or no need to invest in new facilities and equipment, and there may be no need for additional staff. Under these circumstances the incremental cost will only include consumables.

However, in the long run all efficient services will adapt capacity to that which is most efficient. Changing the volume of a service will therefore mean that the fixed costs will change in the long run. For this reason, many economists argue that the correct basis for calculating incremental costs includes any changes in capital and other fixed costs. In many cases this means that the short-run AC is a better proxy for long-run incremental cost than the short-run incremental cost.

In circumstances of economies of scale and scope, divergence between average and marginal or incremental cost applies to the long term. For example, where the underlying cause is 'lumpy' investment requirements, or 'indivisibilities' (units of investment are large), there will be no long-term reconciliation between the two measures of cost. In these circumstances, it is clearer that adjustments for incremental cost have to be made, although there is debate over the extent to which long-run economies of scale and scope exist.

When average cost is being used as an estimate of long-run incremental cost, it is important to check that the change in activity is unlikely to lead to a major change in the most efficient technology of provision. For example, if a new universal vaccination programme replaces a smaller selective one, the whole organisation of the service, and probably the equipment and staff in use, will change. Average costs are unlikely in these circumstances to be a useful basis for estimating the incremental costs of the additional services. In general, for small changes in the volume of a service it is safe to use short-run average cost as a proxy for long-run incremental cost unless the technology is such that there is spare capacity in the current provision, and the service can be efficiently expanded without additional investment.

3.2.1 Sources of Variation in Cost Measures, Confidence Intervals and Assessing Sample Sizes for Costing

There are many reasons why costs vary for the same service in different locations. There is a useful resemblance here with the measurement of the effectiveness of different treatments in clinical trials and studies. The statistical principles for judging the comparative effectiveness of different interventions are widely accepted. Before the start of a clinical trial there is a calculation of the sample that will be needed to give a particular probability of demonstrating a

given difference of effect with a given level of statistical significance. Clearly this calculation is dependent on assumptions about the likely distribution of effects, and this assumed variability in effect for any given treatment is one factor in determining the sample size needed.

In the case of clinical trials it is normal for the basic unit to be the patient. In most studies patients are allocated to different treatments (using random allocation if feasible), and variations coming from different facilities or staff skills matter little since in each site the patients are allocated at random. A problem arises in cases where randomisation has to be by hospital or district rather than by patient, since local facilities or skills may play important roles. There can be similar problems in assessing costs. Since costs for a particular patient depend on disease severity, co-morbidity, hospital size, location and efficiency, it is not clear whether we need a large sample of patients or hospitals to assess the range within which costs are likely to lie.

There is a growing understanding of the ways costs vary as a result of differences in patient characteristics. Many costing studies, particularly in clinical trials, calculate costs for each patient, and this gives data on the degree of variability. Such evidence can give a basis for the calculation of sample size that will allow costs for each category of patients to be assessed with reasonable reliability. When economic evaluation is being carried out alongside clinical trials or studies, this should be done. Some studies have shown that the distribution of service use is highly skewed in certain patient groups, especially in mental health and in cases where some patients receive treatment involving high-technology equipment.

It is usually desirable, but not always feasible, to assess unit costs of services from many different hospitals. In terms of interpreting the results of economic evaluation there are two reasons to be interested in understanding variation in cost between facilities. First, it may be that a particular service or intervention is cost-effective only if provided in a low-cost facility. Knowledge of the structure of costs can allow judgements to be made about where such developments should be located. A good example could be haemoglobinopathy screening. Given the large economies of scale in testing, the service is only likely to be cost-effective if testing can be centralised. Second, unless we know the variation in unit costs in different

facilities, there is a risk that the assessment of cost-effectiveness reflects the chance that the evaluation was done at a low or at a high-cost location. This is somewhat analogous to drawing conclusions about the efficacy of a new treatment from case reports or small studies.

Since it is often not possible to calculate unit costs for services in more than a few centres, it can be impossible to explore the range of likely costs using conventional statistical methods. It is still useful to present evidence of variation in unit costs, but confidence intervals for cost variation are only possible if it is possible to include in the study data from a large enough range of providers to allow the distribution to be analysed. But we should remain interested in the consequences of any errors in estimates of cost, and we should try to ensure that strong recommendations reflect our level of confidence in the estimates.

3.2.2 Using Sensitivity Analysis on Costs

While it is desirable where possible to calculate mean costs and confidence intervals around the mean, since variability may be related to location of services, it is often impossible to do this. In these circumstances it is still desirable to explore the consequences of variation in costs. This is best done by sensitivity analysis. There are two ways in which sensitivity analysis can be used.

(i) First, a range of plausible assumptions can be tested out (such as plus or minus 15 per cent), to see if this is likely to affect the conclusions of the analysis. If there is some basis for judging plausible levels of variation this is appropriate.

(ii) An alternative is to start from the other end, and ask the question 'What size of variation in cost would be needed to change the conclusions?' If the conclusion remains the same with even large variations in cost, this may be grounds for accepting the results as robust.

3.2.3 Costing in Economic Evaluation

The normal approach to calculating costs in economic evaluation is to estimate the number of cost-generating events for each patient, and to multiply this matrix of different events for different patients by a vector of unit costs. As suggested above, this unit cost vector may be calculated using a range of methods, from accounting or budget data or estimates of cost

functions. In many cases simple approaches have been considered to be adequate, and most studies do not take into account changes in costs with time or technical progress. It was argued at the start of this chapter that costing requires understanding of circumstance as well as technique. It may be quite acceptable to assume that costs for a particular service will remain stable over time. Equally, there are some instances in which such an assumption leads to serious errors. For example, in most surgeries in industrialised countries, the length of hospital stays has fallen consistently, and this trend seems likely to continue and to be capable of exploitation in other countries. Failing to take this into account may lead to overestimates of the costs of surgical options in the future. New technologies may reduce in price over time, and may be the subject of learning, suggesting that health workers' skills may develop in such a way that they use the technology more efficiently. Patients and potential patients learn more about the service and how to use it, contributing to reduced costs. For example, a new technology such as the treatment of bed nets with insecticide to combat malaria may require aggressive marketing at first, but rely on word of mouth later once its uptake has reached high levels. Drugs become much cheaper when patents expire. Costing studies should take all these factors into account.

Costing cannot be an exact science, but costs estimated using sensible approaches by people who are well informed about context are more likely to reflect the true foregone opportunities.

SELF ASSESSMENT EXERCISE

Discuss issues involve in the measurement of costs in a health care intervention programme.

3.3 Economic Evaluation

Economic evaluation may be based on the viewpoint of an individual patient, the hospital, the government, or the society at large. Hence, it is important to determine at the beginning from whose viewpoint an economic evaluation is to be carried out. The broadest viewpoint is that of society, as this will include all the costs and benefits. Adopting this approach has two main implications that distinguish it from approaches with more limited perspectives.

(i) Firstly, it usually involves measuring and valuing items that do not have market prices attached to them, such as the time costs that patients incur when undergoing treatment and recuperating.

(ii) Secondly, it means that certain costs, or cost savings, or both, should not be included in the evaluation because they are transfers from one sector to another rather than a net cost to society e.g. free health care.

3.3.1 Characteristics of Economic Evaluation/ Analysis

First, it deals with both the inputs and outputs, sometimes called costs and consequences, of activities. It is the linkage of costs and consequences, which allows us to reach our decision. Second, Economic analysis concerns itself with choices. Resource scarcity, and our inability to produce all desired outputs, necessitates that choices must be made in all areas of human activity. These choices are made on the basis of many criteria, sometimes explicit, but often implicit. Economic analysis seeks to identify and to make explicit set of criteria, which may be useful in deciding among different uses of scarce resources. Economics evaluations:

(i) Always compare any health care programme with an alternative, for example, no treatment or routine care.

(ii) Always measure the benefits produced by all alternatives compared.

(iii) Always measure the cost of any programme. The above characteristics of economic evaluation/analysis lead us to define economic evaluation as the comparative analysis of alternative courses of action in terms of both their costs and consequences.

Therefore, the basic tasks of any economic evaluation are;

(i) to identify,

(ii) measure,

(iii) value,

(iv) compare the costs and consequences of the alternatives being considered.

Economic evaluation of health care programmes aims to aid decision-making with their difficult choices in allocating health care resources, setting priorities and moulding health policy. But it

might be argued that this is only an intermediate objective. The real purpose of doing economic evaluation is to improve efficiency: the way inputs can be converted into outputs (saving life, health gain, improving quality of life, etc.).

The choice of what health care to provide is about what economists call allocative efficiency. This means that we strive for the maximization of benefits (however we decide to measure this) subject to given available resources. So, from a fixed resource we aim to get as much out of a range of health care programmes as possible. This will mean that we will need to compare very different interventions, say health promotion advice to quit smoking versus prescribing Relenza versus a procedure on an ingrown toenail. Thus, allocative efficiency is about finding the optimal mix of services that deliver the maximum possible benefit in total. Resources will be directed to interventions that are relatively efficient at converting inputs into health benefits and away from those that require larger input for relatively low health gain. This approach may be constrained by certain equity considerations, to ensure that certain groups do receive health care.

The choice of how to provide health care is about what economists call technical efficiency. This means that we might strive for minimum input for a given output. For example, if we have decided that performing tonsillectomies on children is worthwhile, part of an allocative efficiency, then we may need to examine the efficiency of how we do this. So, if the output we wish to achieve is to remove a child's tonsils then we might choose between, say, a day case procedure or an inpatient stay. This is an issue of technical efficiency since the output or 'outcome' is fixed, but the inputs will differ depending on which policy we adopt. The day case approach may perhaps require more intensive staff input and more follow-up outpatient visits. If this was the case, then inpatient tonsillectomy may be the more technically efficient strategy.

Thus, with any given health care programme an economic evaluation is aiming to make explicit the total resources consumed specifically by a programme and the total benefit generated by that programme. Drummond et al (1997) defines economic evaluation as "the comparative analysis of alternative courses of action in terms of both their costs and consequences." It differs from other forms of analysis since it considers both costs and consequences and is comparative.

Evaluation needs to be comparative as an intervention can only be labeled as good or bad relative to some benchmark or alternative even if this alternative is a 'do nothing' strategy. If an evaluation is not comparative and does not consider both costs and consequences, then it is only a partial evaluation. It is a description of either the costs or the benefits of one intervention in isolation. This is most uninformative since it is one-dimensional and without a context by which to judge relative performance. If both costs and consequences are considered, but no comparator is provided, then the study is again only a partial evaluation, described as a cost-outcome study. It lacks context and is of limited use. If alternatives are compared, but only in terms of costs or benefits and not both then again the study only provides a partial evaluation and can be labeled an effectiveness study or a cost analysis. It would be comparative, but only across one dimension. Hence, an economic approach can be considered a full evaluation technique.

Whatever the approach, the same three-stage process for the assessment of all costs and benefits can be applied. All relevant cost and benefit variables must be

i) identified,
ii) quantified and
iii) valued.

At the start of an evaluation, it must be determined which costs and benefits are sufficiently important to merit inclusion in the study. This should be separate from the measurement stage so as to avoid the study being entirely data driven (i.e. the more intangible consequences of an intervention might be considered equally important). The identification of relevant benefits and costs will define the variables in the study. These can be broadly classified into changes in resource use, changes in productive output and changes in health state.

The next stage is to measure changes in these variables brought about by the intervention in question. Often it is important that this is done before valuation, as it is necessary to know the magnitude of gains or losses before values can be attached. Presenting variables in terms of 'natural' quantities or frequencies (i.e. hour's worked or clinical units) can also be very useful in

terms of generalisability. Others can use these data and apply values relevant to their own setting (i.e. different cost structures or health values).

The differential timing of costs and benefits must also be considered in an evaluation. The effects of health treatments do not always occur at the same point in time. Costs may be incurred today, but the benefit may not arrive until next year (i.e. preventive treatments, health promotion), part of this future benefit might be that future costs will be avoided. ₦100 spent today may not have the same value as ₦100 spent next year because of inflation; interest on savings and a positive rate of time preference. People may just prefer to have ₦100 in their pocket today rather than ₦100 in a week or a month or a year, because it offers them more choices. This can be incorporated into economic evaluation by the notion of discounting future costs and benefits to their present day value.

SELF ASSESSMENT EXERCISE
Discuss the basic characteristics of economic evaluation/ analysis.

3.4. Types of Economic Evaluation

The different ways of looking at benefits combined with cost analysis represent the different techniques of economic evaluation. The basic types of economic evaluation are cost effectiveness analysis (CEA), cost minimisation analysis (CMA), cost utility analysis (CUA) and cost benefit analysis (CBA). When to see each of the above techniques will depend on the nature of the question to be addressed, which may be a choice between alternative clinical strategies for a condition: timing of an intervention; settings for care; types and skill-mix of personnel proving care; programmes for different conditions; or other ways to improve health.

3.4.1. Cost-Effectiveness Analysis

When different health care interventions are not expected to produce the same outcomes both the costs and consequences of the options need to be assessed. This can be done by cost-effectiveness analysis, whereby the costs are compared with outcomes measured in natural units-for example, per life saved, per life year gained, and pain or symptom free day. Many cost-effective analyses rely on existing published studies for effectiveness data, as it is often too

costly or time consuming to collect data on costs and effectiveness during a clinical trial. Where there is uncertainty about the costs and effectiveness of procedures sensitivity analysis can be used, which examines the sensitivity of the results to alternative assumptions about key variables. CEA is concerned with technical efficiency issues, such as: what is the best way of achieving a given goal or what is the best way of spending a given budget. Comparisons can be made between different health programmes in terms of their cost effectiveness ratios: cost per unit of effect. Under CEA effects are measured in terms of the most appropriate uni-dimensional natural unit. So, if the question to be addressed was: what is the best way of treating renal failure? Then the most appropriate ratio with which to compare programmes might be 'cost per life saved'. Also, if we wanted to compare the cost-effectiveness of programmes of screening for Down's syndrome the most appropriate ratio might be 'cost per Down's syndrome fetus detected'. The advantages of the CEA approach are

(i) it is relatively straightforward to carry out

(ii) It is often sufficient for addressing many questions in health care. However, it is not comprehensive. The outcome is uni-dimensional under this analysis, but often health programmes generate multiple outcomes.

(iii) For example, in Downs' syndrome screening, foetus detected is one outcome, but miscarriages avoided might be another very relevant outcome measure, especially if, say, blood testing is being compared to amniocentesis. But this cannot be incorporated into this form of analysis. So, CEA not only assumes that the outcome of the health programme is worthwhile per se, but also that it is the most appropriate measure. A further problem with CEA is comparability between very different health programmes. Cost per foetus detected may be a useful way to compare the efficiency of blood testing versus amniocentesis, but how would these be compared to, say, drugs aimed at reducing cholesterol. Health programmes with different aims cannot be compared with one another using CEA: cost per unit reduction in cholesterol cannot meaningfully be compared with foetus detected. Hence, CEA is useful when comparing programmes within like areas, where common 'currencies' can be used.

If the outcomes of alternative procedures or programmes under review are the same, then attention can focus upon the costs in order to identify the least cost option. Then, the method of evaluation will be cost-minimisation analysis. If, however, the outcomes are not expected to be the same, then both the costs and consequences of alternative options need to be considered. Cost-effectiveness analysis is one method of economic evaluation that allows this to be done.

3.4.1.1 Measures of Effectiveness

In order to carry out a cost effectiveness analysis it is necessary to have suitable measures of effectiveness. These will depend on the objectives of the particular interventions under review. In all cost effectiveness analysis, however, measures of effectiveness should be defined in appropriate natural units and, ideally, expressed in a single dimension.

Common measures used in several studies have been “lives saved” and “life years gained”. Thus, Boyle and colleagues, in their study of neonatal intensive care of very low birth weight babies, measured effectiveness in terms of mortality rates at the time of discharge of newborn infants from hospital. Their study compared two periods-one before the introduction of neonatal intensive care, and another after its introduction-and measured cost effectiveness in terms of additional costs per life saved. Several other measures of effectiveness have been used by different researchers. These have included the number of pain or symptom free days resulting from alternative drug regimens in the treatment of duodenal ulcers; and the number of episodes of fever cured and deaths prevented in the treatment of chloroquine resistant malaria in African children. Most of the above mentioned studies express effectiveness in terms of a single dimension and thereby permit direct comparison between alternative procedures in terms of their marginal cost per unit of outcome. Sometimes, however, the alternatives under examination have multiple outcomes. Nonetheless, many of these choices can be dealt with in the cost-effectiveness analysis framework. Thus, if one procedure emerges as less costly and of equal or greater effectiveness than all the other options on each dimension of effectiveness, it is clearly the most cost effective option. For example, the comparison of day surgery with overnight inpatient care for cataract surgery, measured outcomes in terms of the number of both operative and postoperative complications, and in terms of visual acuity of patients three to six days and

10 weeks to six months after surgery. Patient satisfaction was also elicited through a questionnaire. As day surgery emerged as the more effective option on practically all of these effectiveness measures, and was subsequently less costly, the evidence suggests that it is the preferred option. One argument for carrying out analysis in this way-that is, not always seeking to combine outcome measures into a single unit, is that the variations across a number of dimensions are made clear to decision makers rather than being concealed within an aggregate measure. This can sometimes permit more informed decision-making.

Table 3.1: Measures of Effectiveness

• Cases treated appropriately
• Lives saved
• Life years gained
• Pain or symptom free days
• Cases successfully diagnosed
• Complications avoided

3.4.2 Costs-Minimisation Analysis

Cost-minimisation analysis (CMA) is an appropriate evaluation method to use when the case for an intervention has been established and the programmes and procedures under consideration are expected to have the same or similar outcomes. In these circumstances, attention may focus on the cost side of the equation to identify the least costly option. Cost –Minimisation

(i) is concerned only with technical efficiency;

(ii) can be regarded as a narrow form of cost effectiveness analysis;

(iii) evidence is given on the equivalence of the outcomes of different interventions; and

(iv) as outcomes are considered to be equivalent no different decisions can be made on the basis of costs.

The advantages of cost minimisation analysis include:

(i) Simple to carry out, requires costs to be measured, but only that outcomes can be shown to be equivalent.

(ii) Avoids needlessly quantifying data.

The disadvantages are:

(i) Can only be used in narrow range of situations.

(ii) Requires that outcomes be equivalent.

An example of CMA can be as in comparing two programmes involving minor surgery for adults. Both accomplish the outcome of interest, and from an examination of effectiveness data differ in no other significant respects except that one requires hospital admission for at least one night, while the other (a day surgery programme) does not. If we identified the common outcome of interest – operations successfully completed – we would find that it could be achieved to the same degree (i.e. identical number of surgeries) in either programme, though at different costs. The economic evaluation is then essentially a search for the least cost alternative. Analysis such as this is often called cost-minimization analysis. We might also be interested in the distribution of costs (e.g. in this case to what extent does the day-surgery programme shift costs to the patient), but our principal efficiency comparison will be made on the basis of cost per surgical procedure.

3.4.2.1. Discounting Benefits (in Cost-Effectiveness Analysis)

Costs incurred at different points in time need to be "weighted" or discounted to reflect the fact that those that occur in the immediate future are of more importance than those that accrue in the distant future. This raises the question: should the benefits or effects of alternative procedures also be discounted? In answering this issue there is a difference among economists. If a zero discounting (no discounting applied) were adopted, the main consequence would be to change the relative cost effectiveness of different procedures. Using a positive discount rate means that projects with long lasting effects receive lower priority. If a positive rate is replaced by a zero rate, procedures such as neonatal care-which lead to benefits over the recipient's entire future lifetime-will, become relatively more cost effective. In practical terms, it is probably true to say that while the case for using a zero discount rate for benefits has powerful intellectual and may gain empirical support in the future, it will be too hasty to recommend that positive rates be discarded in economic evaluations. In general

(i) cost-effectiveness analysis is a form of economic evaluation in which the costs of alternative procedures or programmes are compared with outcomes measured in natural units-for example, cost per life year saved, cost per symptom free day.

(ii) Effectiveness data are collected from economic evaluations built in alongside clinical trials. In the absence of dedicated trials researchers need to draw on the existing published work.

(iii) Sensitivity analysis should be applied when there is uncertainty about the costs and effectiveness of different procedures. This investigates the extent to which results are sensitive to alternative assumptions about key variables.

(iv) There is debate among economists about whether benefit measures should be "time discounted" in the same way as costs. If they are not, projects with long lasting effects will become relatively more cost effective-for example, maternity services and health promotion. But it will be probably wrong to recommend this as a standard practice.

3.4.3. Cost-Utility Analysis (CUA)

CUA is concerned with technical efficiency and allocative efficiency (within the health care sector). It can be thought of as a sophisticated form of CEA, since it also makes comparisons between health programmes in terms of cost-effect ratios. However, CUA differs in the way it considers effects. CUA tends to be used when quality of life is an important factor involved in the health programmes being evaluated. This is because CUA combines life years (quantity of life) gained as a result of a health programme with some judgment on the quality of those life years. It is this judgment element that is labeled utility. Utility is simply a measure of preference, where values can be assigned to different states of health that represent individual preferences. This is done by assigning values between 1.0 and 0.0, where 1.0 is the best imaginable state of health (completely healthy) and 0.0 is the worst imaginable (perhaps death). States of health may be described using many different instruments which provide a profile of scores in different health domains. EuroQol EQ-5D for example, simplifies health into just five domains (such as mobility, self-care, usual activities, pain/discomfort and anxiety/depression).

- Each domain is given a score from 1 to 3,
- So the health profile would read 11111 for the best scores in all domains
- 33333 for the worst.

This approach of using utility is not restricted to similar clinical areas, but can be used to compare very different health programmes in the same terms. As a result, 'cost per QALY gained' league tables are often produced to compare the relative efficiency with which different interventions can turn resources invested into QALYs gained. It is possible to compare surgical, medical and health promotion interventions with each other. Comparability then is the key advantage of this type of economic evaluation. For a decision-maker faced with allocating scare resources between competing claims, CUA can be very informative. The key problem with CUA is the difficulty of deriving health benefits.

Table 3.2: Advantages and Disadvantages of Cost per QALY gained 'League Tables'

Pros
• reveals opportunity cost
• common currency
• comparison across diseases
• considers length and quality of life
• investment type problem- "best returns"
• underlying principle – buy "cheap" QALYs not "expensive" QALYs
Cons
•What of equity?
• What of equality of access?
• only health service costs
• What of other health benefits?
• patient information/ reassurance
• Comparability of C-U-A studies
• Lack of them
• Apply locally?

3.4.3.1 When should CUA be used?

The following are a number of situations CUA may be used:

(i) When health-related quality of life is the important outcome. For example, in comparing alternative programmes for the treatment of arthritis, no programme is expected to have any impact on mortality, and the interest is focused on how well the different programmes will be improving the patient's physical function, social function, and psychological wellbeing;

(ii) When the programme affects both morbidity and mortality and we wish to have a common unit of outcome that combines both effects. For example, treatments for many cancers improve longevity and improve long-term quality of life, but decrease quality of life during the treatment process.

(iii) When the programmes that are being compared have a range of different kinds of outcomes and we wish to have a common unit of output for comparison. For example, if a health planner who must compare several disparate programmes applying for funding, such as expansion of neonatal intensive care, a programme to locate and treat hypertension, and a programme to expand the rehabilitative services provided to post-myocardial infarction patients;

(iv) When we wish to compare a programme to others that have already been evaluated using cost-utility analysis.

3.4.3.2 When should CUA not be used?

The following are a number of situations where the CUA may not be used:

(i) When only intermediate outcome data can be obtained. For example, in a study to screen employees for hypertension and treat them for one year, intermediate outcomes of this type cannot be readily converted into QALYs for use in CUA.

(ii) When the effectiveness data show that the alternatives are equally effective in all respects of importance to consumers (e.g. including side-effects). In this case, cost-minimization analysis is sufficient; CUA is not needed;

(iii) When the effectiveness data show that the new programme is dominant; that is, the new programme is both more effective and less costly (win-win). In this case, no further analysis is needed;

(iv) When the extra cost of obtaining and using utility values is judged to be in itself not cost effective. This is the case above in points (ii) and (iii). It would also be the case even when the new programme is costlier than the old, if effectiveness data show such an enormous superiority for the new programme the incorporation of utility values could almost certainly not change the result. It might even be the case with a programme that is

costlier and only somewhat more effective, if it can be argued that the incorporation of any utility values will show the programme to be overwhelmingly cost-effective.

3.4.3.3 Measuring Quality

Measuring a person's quality of life is difficult. Nonetheless, it is important to have some means to have for doing so since many health care programmes are concerned primarily with improving the quality of a patient's life rather than extending its length. For this reason, various quality of life scales has been developed in recent years. The Nottingham health profile is one quality of life scale that has been used quite widely in Britain. This comprises two parts:

(i) The first measures health status by asking for yes or no responses from patients to a set of 36 statements related to six dimensions of social functioning:

- ❖ Energy,
- ❖ Pain,
- ❖ Emotional reactions,
- ❖ Sleep,
- ❖ Social isolation,
- ❖ Physical mobility. These responses are then "weighted" and a score of between 0 and 100 is assigned to each dimension.

(ii) The second part asks about seven areas of performance that can be expected to be affected by health:

- ❖ Employment,
- ❖ Looking after the home,
- ❖ Social life, Home life,
- ❖ Sex life,
- ❖ Hobbies,
- ❖ Holidays. The Nottingham health profile has been applied, for example, in studies of heart transplantation, rheumatoid arthritis and migraine, and renal lithotripsy.

Other widely used measures include the sickness impact profile and the quality of wellbeing scale. Recently, a new outcome measure, the Sf-36 health survey questionnaire, has been

gaining popularity. After testing it on 1980 patients in two general practices it was considered to be a promising measure which is "easy to use, acceptable to patients, and fulfils stringent criteria of reliability and validity". Although all of these scales embody some form of scoring scheme, they do not usually generate a single quality of life score. This means that, although they are of considerable value in assessing the outcomes of interventions in the case of particular diseases or disabilities, they cannot be used to compare outcomes between different programmes. To do this, generalisable measure of quality is necessary. One of the earliest measures to be developed-and one which has subsequently been used widely to calculate QALYs –is the Rosser index.

3.4.3.3.1 Rosser Index

Rosser Index described health status in terms of two dimensions: disability and distress. The states of illness are classified into eight categories of disability and four categories of distress. By combining these categories of disability and distress 32 (8 times 4), different states of health were obtained. Rosser then interviewed 70 respondents (a mixture of doctors, nurses, patients and healthy volunteers) and, by using psychometric techniques sought to establish their views about the severity of each state relative to other states. The final results of this exercise were expressed in terms of a numeric scale extending from 0 = dead to 1 = perfect health. With this classification system it becomes possible to assign a quality of life score to any state of health as long as it is placed in an appropriate disability or distress category.

3.4.3.3.2 Quality – Adjusted Life – Years (QALY)

One of the features of conventional CUA is its use of the QALY concept. Results are reported in terms of cost per QALY gained. QALYs:

(i) combine life years gained with a measure of the quality of those years.

(ii) Quality is measured on a scale of 0 to 1, with 0 equated to being dead and 1 equated to the best imaginable state of health.

(iii) Combine all dimensions of health & survival into a single index.

$$\text{CU ratio} = \frac{\text{Cost A} - \text{Cost B}}{\text{QALY A} - \text{QALY B}}$$

3.4.3.3.3 QALY Concept

The advantage of the QALY as a measure of health outcome is that it can simultaneously capture gains from reduced morbidity (quality gains) and reduced mortality (quantity gains), and combine these into a single measure. Moreover, the combination is based on the relative desirability of the different outcomes.

The QALY approach, which forms a key part of most cost-utility analyses, has been the subject of some criticism. It has been accused of discriminating against elderly people, making illegitimate interpersonal comparisons, disregarding equity considerations, and introducing bias into quality of life scores. Rival measures that are claimed to be sound theoretically, such as "healthy years equivalents" (HYEs), have also been put forward. It has been claimed that under most assumptions QALYs and HYEs will lead to identical project rankings. Amid all this debate it is as well to bear in mind that decisions have to be made about the allocation of resources and cost-utility analysis is probably the most sophisticated form of economic evaluation available at present. However, the technique and interpretation of research findings should recognise that cost utility-analysis is still at a fairly early development stage and treat it accordingly.

3.4.3.3.4 DALY Concept

The Disability-Adjusted Life Year is a measure akin to the QALY in aggregating survival and quality of life effects, but normally advanced as a method of estimating the burden of illness associated with a disease, rather than the cost-effectiveness of health care interventions.

3.4.4. Cost-Benefit Analysis

Cost benefit analysis is the most comprehensive and theoretically sound form of economic evaluation and it has been used as an aid to decision making in many different areas of economic and social policy in the public sector for more than fifty years. Cost-Benefit analysis (CBA) estimates and totals up the equivalent money value of the benefits and costs to the community of projects to establish whether they are worthwhile. These projects may be dams and highways or can be training programmes and health care systems. The main difference between cost-benefit analysis and other methods of economic evaluation that were discussed earlier in this series is that it seeks to place monetary values on both the inputs (costs) and

outcomes (benefits) of health care. Among other things, this enables the monetary returns on investments in health to be compared with the returns obtainable from investments in other areas of the economy. Within the health sector itself; the attachment of monetary values to outcomes makes it possible to say whether a particular procedure or program offers an overall net gain to society in the sense that its total benefits exceed its total costs. Cost-effectiveness and cost-utility analysis do not do this because they measure costs and benefits in different units. CBA requires programme consequences to be valued in monetary units, thus, enabling the analyst to make a direct comparison of the programmes incremental cost with its incremental consequences in commensurate units of measurement, be they Birr, dollars, or pounds. CBA compares the discounted future streams of incremental programme benefits with incremental programmes costs; the difference between these two streams being the net social benefit of the programme. In simple terms, the goal of analysis is to identify whether a programme's benefits exceed its costs, a positive net social benefit indicating that a programme is worthwhile. CBA is a full economic evaluation because programme outputs must be measured and valued. In many respects CBA is broader in scope than CEA/CUA. Because CBA converts all costs and benefits to money, it is not restricted to comparing programmes within health care, but can be used to inform resource allocation decisions in the sectors of the economy. CBA is broader in scope and able to inform questions of allocative efficiency, because it assigns relative values to health and non-health related goals to determine which goals are worth achieving, given the alternative uses of resources, and thereby determining which programmes are worthwhile.

(i) Both costs and benefits are assigned a monetary value. The benefits of any intervention can then be compared directly with any costs incurred. If the value of benefits exceeds the costs, then it is potentially worthwhile to carry that intervention out.

(ii) It is concerned with allocative efficiency.

(iii) It is concerned with the question, is a particular goal worthwhile? It can answer questions such as should extra money be used for heart transplants or improving housing?

(iv) Method requires that all resources and benefit generated by an intervention need to be assigned a monetary value. Therefore, it needs to cost things which have no market value, i.e., changes in health, quality of life, length of life, pain, etc.

(v) Methods of valuing

- ❖ willingness to pay (WTP)
- ❖ human capital approach (HCA)

The net welfare gain or net value of a project X (NVX) is equal to

$$NVX = WTPX – WTY$$

Where: y refers to the next best alternative project, if the latter cannot be defined.

$$NVX = WTPX – WTPXi$$

Where: WTPXi refers to society while WTP is for the inputs used alternatively in the economy. If NVX is positive then, project X may be undertaken. When several projects compete with each other, the one with the highest NV needs to be selected in order to maximize welfare. This shows the CBA for projects that have benefits or costs in the current period. It is evident that projects may also entail future benefits and future costs. Some modifications in the calculation of net value will be required in this case. Note that individuals prefer a net value of ₦1 received now to ₦1 in the future. It follows that one cannot simply add up benefits or costs that are related to different points in time. A social discount rate, denoted as r, will enable us to add up a stream of net benefits, namely, ₦1in year one will be worth ₦1tr in year two, ₦$1tr^2$ in year three etc.; conversely, ₦1 in year two is worth ₦(1/1tr) in year one, ₦1 in year three is worth ₦$(1/1tr)^2$ in year one etc. The value in year one of a naira received or paid in the future is called the present value of that naira. Making use of the social discount rate r, we can calculate the net present value (NPV) of a project.

$$NPV = [(Bt – Ct) / (1+r) – 1]$$

Where: ‘B’ and ‘C’ refer to benefits and costs and‘t’ is the time index. ‘Bt’ is equal to the WTP for the nth project at time ‘t’, while ‘Ct’ has to be understood as the benefits forgone in period ‘t’. Note that if NPV > 0 society’s welfare will increase; hence the project can be adopted. If several projects are competing with each other the one with the highest NPV should be chosen.

SELF ASSESSMENT EXERCISE

Discuss the various methods of economic evaluations

4.0 CONCLUSION

Costs can be defined in many ways but generally can be considered as direct, indirect and intangible. Costs in economics are broader and economics considers both the present values and future value of costs. However, it is important to determine at the outset an economic evaluation to be carried out. It may be based on the viewpoint of an individual patient, the hospital, the government, or society at large. The broadest viewpoint is that of society in general, as this will include all the costs and benefits. For this reason, it is the preferred approach. Economic evaluation considered many possibilities and various alternative resources can be used.

5.0 SUMMARY

This unit discussed extensively costs concept and economic evaluation. It considered present value and future value of alternatives course of action. It also considered measurement of values and various methods of economic evaluation.

6.0 TUTOR-MARKED ASSIGNMENT

1. Costs are incurred in all economic activities – why?
2. Explain the cost implications of ill health.
3. Define cost benefit analysis and explain its difference from cost effectiveness analysis.
4. Define the terms net present value and discount rate.

7.0 REFERENCES/FURTHER READINGS

Donaldson Cam and Karen Gerard (1993) Economics of Health Care Financing: The Visible Hand. Macmillan Press Ltd. London.

Folland S., A. Goodman & M. Stano (2010) The Economics of Health & Health Care, Sixth Edition, Prentice Hall, New Jersey.

Jacobs, P. (1991) The Economics of Health and Medical Care Maryland: Aspen Pub Inc. Jack, Williams (1964) Principles of Health Economics for Developing Countries. WBI Development Studies. The World Bank, Washington D. C.

Santerre E. & S.P. Neun (1996) Health Economics: Theories, Insights & Industry Studies, Irwin, Chicago.

Zweifel P., F. Breyer & M. Kifmann (2009) Health Economics, Second Edition, Springer Verlag Heidelberg.

Unit 2: Health Care Financing

CONTENTS

1.0 INTRODUCTION

The precise definition of what services and activities comprise of the health sector is necessary to guide data collection and to make comparisons of health systems across countries or at different times. The following pairs of items show the difficulty of drawing a line between aspects of the health sector/non-health sector. Which should be included within the definition of the health sector? heath services, environmental services (e.g. water, sanitation, environmental pollution control, occupation safety etc.), hospitals, social welfare institutions, education and training, pure medical research, medical social work, social work, formally trained medical practitioners, traditional medical practitioners. In practice, the boundaries of the health sector vary considerably between countries and different definitions have been developed for different purposes. In developing countries, the definition tends to be broader than in developed countries due to greater deficiencies in certain areas (e.g. environmental health) and extensive use of the traditional health sector. A useful rule of thumb is to include all finance/ expenditure whose primary intention (regardless of effect) is to improve health. Financing refers to raising revenue to pay for a good or service. It is the function of a health system concerned with the mobilization, accumulation and allocation of money to cover the health needs of the people, individually and collectively, in the health systems. The whole processes of health care finance

involve where the money came from, how it was collected and used to pay the providers for their services. This unit discusses health care financing.

2.0 OBJECTIVES

By the end of this unit, students should be able to:

(i) identify the factors that influence the choice of a financing system;

(ii) explore the different sources of financing in the health service sector; and

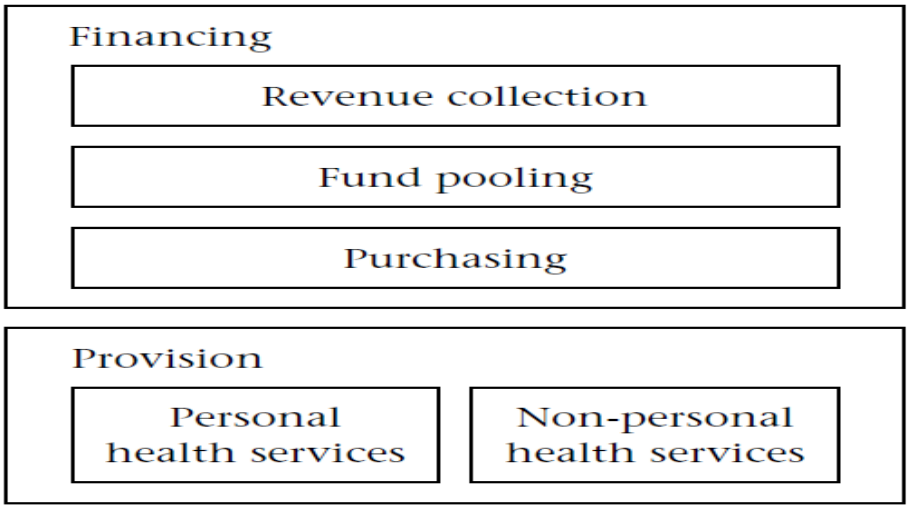
(iii) understand the strong and weak points of different financing mechanisms.

3.0 MAIN CONTENT

3.1 Forms of Health Care Financing

Financing refers to the ways in which money is raised to fund health activities as well as how it is raised to achieve a nation's health objectives. Health financing is a collection of funds from various sources (e.g., government, households, businesses and donors), pooling them to share financial risks across large population groups and using them to pay for services from public and private health care providers. Five methods of financing health activities are general and earmarked taxes, social and private insurance, community financing and out–of-pocket payments. But, a financing strategy which determines how these different methods are combined is based on the amount of funds available for health care, who controls the resources, and who bears the financial burden. The strategy chosen has implications for the health status and financial risk protection of various income and age groups. Sustainable health care systems are built on reliable access to human, capital and consumable resources. Securing these inputs requires financial resources to pay for investment in buildings and equipment, to compensate health service staff for their time and to pay for drugs and other consumables. How these financial resources are generated and managed – the process of collecting revenue and pooling funds – raises important issues for policy-makers and planners faced with the challenge of designing systems of funding that meet specific objectives related to social policy, politics and economics. Most countries feel constant pressure because expenditure is increasing and resources are scarce.

According to WHO (2002) there are several factors nations have to consider in their selection of health care financing methods. These include their fiscal capacity, equity, efficiency in raising funds, and the economic effects of raising the fund. Capacity depends on context – the fiscal capacity of any method will depend on the economic structure of the society (the proportion of workers in the formal sector, and on the government's administrative capacity to collect taxes or social insurance contributions). Therefore, when nations search for financing strategies to improve the performance of their health systems, they need to know the relative strengths and weaknesses of the five financing methods. The health care system can be broken down into functional components, as shown in Figure 3.2: revenue collection, fund pooling and the purchasing and provision of health care. Functions can be integrated and separated in various combinations. In some cases, the functions are integrated within a single organizational entity; in others, one entity may collect and pool the funds while other bodies purchase and provide the services. Resources are then allocated between these different entities.



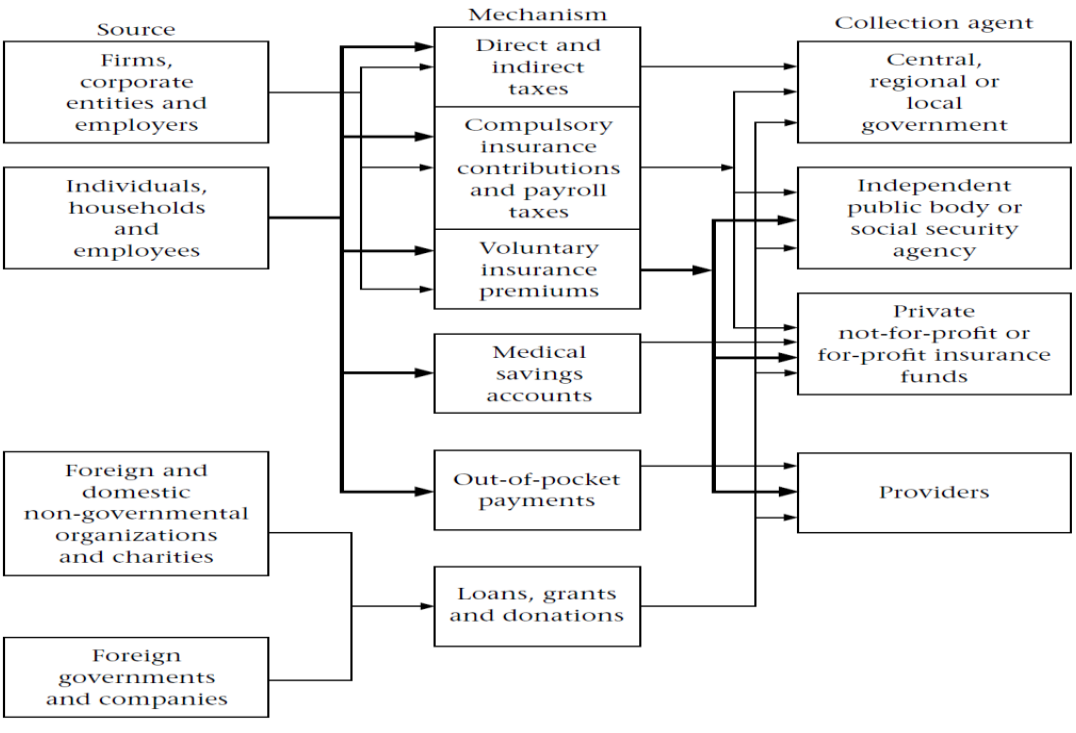
Source: adapted from Murray and Frenk (2000)

Figure 3.2: Functions of health care systems

3.1.1 Revenue collection

The process of revenue collection is specifically concerned with who pays, the type of payment made and who collects it. Figure 3.3 illustrates the diversity of sources of funding, contribution mechanisms and collection agents and how these interrelate. Funds derive primarily from the population (individuals and corporate entities). The funding mechanisms include taxation, social

insurance contributions, private insurance premiums, individual savings, out-of-pocket payments and loans, grants and donations. Collection agents can be private for-profit, private not-for-profit or public. Taxes can be levied on individuals, households and firms (direct taxes) or on transactions and commodities (indirect taxes). Direct and indirect taxes can be levied at the national, regional or local levels. Indirect taxes can be general, such as a value-added tax, or applied to specific goods, such as an excise tax. Some social or compulsory insurance contributions are, in fact, a payroll tax collected by government. Taxes can be general or hypothecated – that is, earmarked for a specific area of expenditure. Social health insurance contributions are usually related to income and shared between the employees and employers. Contributions may also be collected from self-employed people, for whom contributions are calculated based on declarations of income or profit. Contributions on behalf of elderly, unemployed or disabled people may be collected from designated pension, unemployment or sickness funds, respectively, or paid for from taxes.



Source: adapted from Kutzin (2001)

Figure 3.3: Examples of Funding Sources, Contribution Mechanisms and Collection Agents

Private health insurance premiums are paid by an individual, shared between the employees and the employer or paid wholly by the employer. Premiums can be: individually risk rated, based on an assessment of the probability of an individual requiring health care; community rated, based on an estimate of the risks across a geographically defined population; or group rated, based on an estimate of the risks across all employees in a single firm. Government may subsidize the cost of private health insurance using tax credits or tax relief.

Medical savings accounts are individual savings accounts into which people are either required to, or given incentives to deposit money. The money must be spent on personal medical expenses. Medical savings accounts are usually combined with high-deductible catastrophic health insurance. Patients may be required to pay part or all of the costs of some types of care in the form of user charges. These charges may be levied as a co-payment (a flat-rate payment for each service), co-insurance (a percentage of the total cost of the service) or a deductible (a ceiling up to which the patient is liable after which the insurer covers the residual cost). The collection agent is the provider, such as a physician, hospital or pharmacist.

3.1.2 Fund pooling

Revenue collection must be distinguished from fund pooling, as some forms of revenue collection do not enable financial risks to be shared between contributors, such as medical savings accounts and out-of-pocket payments. Fund pooling is the ‘accumulation of prepaid health care revenues on behalf of a population’. The importance of fund pooling is that it facilitates the pooling of financial risk across the population or a defined subgroup. Examples of this include social health insurance contributions collected by funds and retained by them and national, regional or local taxes that are collected and retained. If different agents carry out these functions, a mechanism is required to distribute resources from the collection agent to the pool. If there are multiple pools, allocation is increasingly being adjusted according to the risk profile of the population covered by each pool. This process is referred to as ‘risk adjustment’. Risk adjustment in competitive social health insurance systems has developed mainly from a concern to prevent cream-skimming. Within tax-financed systems, risk-adjusted capitation methods developed from a concern to ensure equity of access by ensuring a fair allocation of resources to

territorial health authorities based on the needs of the population. Irrespective of the source of funds, the underlying rationale for allocating based on risk-adjusted capitation is the same – to ensure that each pool has the 'correct' relative level of resources for the population for which it is responsible. Under private health insurance, funds are pooled between subscribers of the same insurance provider. The extent of risk pooling is limited with actuarial premiums related to an individual's risk. If premiums are community rated, pooling is between high-risk and low-risk members in the same geographic area.

Medical savings accounts prevent pooling by keeping funds in individual accounts. Medical savings accounts are supplemented with catastrophic insurance for expensive treatments. User charges are paid at the point of service and are not a form of pooling. The revenue generated by user charges is handled differently depending on how the system is designed. For example, the individual health care provider may retain the money as income. It may be retained at the level of a clinic or hospital and, together with other revenue, contribute to the cost of maintaining local service provision. If the user charges are given to, or levied by, the insurer or government, they may be used to meet any gap between premium or tax revenue and expenditure.

3.1.3 Purchasing

Purchasing means 'the transfer of pooled resource to service providers on behalf of the population for which the funds were pooled'. In some systems, separate agents purchase services (e.g. Primary Care Trusts in England); in this case, the resources have to be allocated to the purchasers. Pursuing widely held objectives of equity and efficiency requires allocating resources according to health care need. However, many health care systems continue to allocate resources based on political negotiation, historical precedent or the lowest bids.

SELF ASSESSMENT EXERCISE

Discuss different forms of health financing.

3.2 Sources of National Health Care Financing Systems

Health care financing is a broad term used to define alternative arrangements for paying, allocating, organizing and managing health resources. It includes:

(i) defining a level/ quality of care preferably a minimum basic health services packages to be provided, in an accessible and equitable manner;

(ii) identifying different modalities of financing to establish a financially sustainable system; and

(iii) institute different mechanisms for mobilizing funds and rationalizing the use of available resources including cost and risk –sharing mechanisms/ insurances plans.

3.2.1 Financing Strategies

The financing mechanisms are grouped into broad and complementary strategies. It includes improving government health sectors efficiency, generating additional and new sources of revenue, encouraging private and non-governmental organizations participations, development of social and private health insurance, promotion of community participation, encouraging bilateral and multilateral agencies participation, alternate financing options for the urban areas and organizational mechanisms for implementation of the health care and financing strategies. National health care financing systems has pluralistic nature in funding. Therefore it has different sources of health care funding which are

(i) Public sources

- ❖ Direct government budgeting
- ❖ National health services and public services health systems
- ❖ Social health insurances sponsored or mandated by the government
- ❖ Community financing

(ii) Private sources

- ❖ Direct payment by households
- ❖ Private voluntary health insurance
- ❖ Employers based health insurances
- ❖ Payments by community and other local organizations

(iii) External financing

- ❖ Foreign aid or development loans

3.2.1.1 Government Financing

3.2.1.1.1Public and QUASI-public sources of Finance

a) **General tax revenues:** General tax revenue is used in almost every country of the world to finance certain components of health care; and in developing countries; it is often the most important source of financing. However, low tax ratios (the proportion of national income collected as tax) in these countries mean that it is often insufficient by itself to support health care. Although tax ratios tend to increase in line with development, this depends in larger part on a country's political will to increase the tax burden. In developing countries general tax revenue is composed largely of duties on imports and exports and sales taxes. Taxes on business transactions, profits and incomes are all of lesser importance. General tax revenue is currently not the most reliable source of finance for the health sector in developing countries. This results from factors such as the low political priority frequently given to the health sector in national budget decisions; the instability of government finance in countries heavily dependent upon taxes on imports and exports; the frequent use of public expenditure as a tool of macro-economic policy; and frequent disparities between budgeted funds and their actual availability or disbursement. The net yield is usually high unless bureaucratic overheads are high. The equity impact of tax systems is dependent on both the proportional burden of taxation and on the use which is made of the revenue raised. Tax systems can be progressive, falling more heavily on the rich than the poor and, therefore, equitable; but they may also be regressive falling more heavily on the poor than the rich, and inequitable. Developing countries are assumed to have regressive financing systems because they tend to rely on indirect taxation. But in practice their tax systems may be progressive because the poorest sections of society fall outside the formal economy and indirect taxes may be levied primarily on luxury items consumed predominantly by the wealthier population groups. Available evidence on the burden of taxation is inadequate to permit often used inequitably in health systems. Health systems are comminuted by high-technology urban-based care and so the rural populations (and the urban poor) have inadequate access to any form of care. There is a limit to what can be collected in tax

revenue and how much can be allocated to the health sector without conflict with wider primary health care objectives. Taxes that make the poor poorer could seriously damage their health status and undermine their productivity.

b) **Deficit Financing:** General tax revenue may be supplemented by deficit financing that is the decision to borrow and spend funds in the present and repay them over some period of time. Deficit finance may be raised nationally or internationally, through mechanisms such as the issuing of certificates or long-term low-interest loans. The cost enjoying the use of those funds in the present rather than the future is the interest that needs to be paid on the loan. In developing countries high inflation rates (affecting the real of interest on loans) and lack of confidence in the government's abilities to honor eventual redemption of the bonds may make it difficult to use deficit financing as a source of support for health systems. When it is used, deficit financing is typically for specific construction projects (e.g. hospitals water and sewage systems). Unless such projects deliver well, their services or contribute directly to increased output that can be taxed to service the debt, the deficit must be repaid from general tax revenue. Thus, the agency doing the deficit financing must be endowed with the authority to impose additional taxes or fees, or be given a claim on general tax revenue in order to service the debt. Deficit finance may also be raised from abroad in the form of bilateral or multilateral AID loans, typically given for a project life of between three and five years, and thereby constituting a short-term source of support. Although useful for many developing countries in helping to develop and expand health care infrastructure, foreign aid is often limited to support import components.

c) Earmarked Taxes Most tax revenues are paid into a national pool and then shared out between different areas of government expenditure. Some governments, however, may "earmark" a particular tax for a particular purpose. For example, taxes on the sale of particular products may be earmarked for health services at either national or local level. The problem with such taxes is that they are often difficult to administer, may be politically unpopular and are also often unpopular with tax administrators because they limit their freedom of action. They can be regressive if, as often the cases, taxes are levied on items such as beer, cigarettes, recreational events, or foodstuffs; but they can be progressive if

they are imposed on luxury, products purchased primarily by the more influential sections of society. A clear advantage of this source of finance is that a tax is visibly assigned to priority funding of certain activities or programs. Although not a major source of health sector finance, they may constitute an important source of finance for specific projects.

d) Social Insurance: Social insurance can finance health care, as well as other needs such as invalidity and old age support. It is conventionally financed by imposing mandatory insurance payments on employed workers as a percentage of their wages, and by imposing a similar higher payroll tax on their employers. In order to include those workers outside the modern employment sector insurance payments may also be calculated on measured income or wealth other than wages, such as the value of crops produced. Allowance will then have to be made for the fact that cash income is only available seasonally, when crops are sold. In their capacity as employers, governments may either run their own social insurance scheme or contract such schemes from private insurance companies. The total financial contribution to social insurance schemes is (in theory) determined actuarially on the basis of the incidence of illness, the conditions of eligibility for benefit, and the value of those benefits. Individual contributions are not determined, however, on the basis of expected risks or claims, but in some proportion to income. As risks are pooled, there is an unequal benefit distribution in favor of high-risk (high-need) workers. The main problems of social insurance are related to issues of equity and efficiency. It is easiest to cover those in regular employment, who may be as little as 5 to 15% of the population in developing countries; and there are often marked inequalities in the quantity and quality of services available to those covered by insurance relative to those who are not overall, it is argued that social insurance reinforces the mal-distribution of resources between rural and urban areas in developing countries. It provides extra funds for largely urban, employed workers and leaves the large rural population and the informally employed urban population even further handicapped than before its introduction. Critics of social insurance also argue that it undermines both public and private health care by competing with these sectors for limited supplies of real medical resources (e.g. personnel). Finally, it tends to promote or reinforce high-cost, hospital-based, doctor-centered, curative care.

e) Lotteries and Betting: These may be used as sources of earmarked income for health and social services in developing countries. Often administrated by quasi-public bodies under national or local government regulation, these typically non-profit schemes rarely constitute an important component of overall health sector finance. Largely supported by the incomes of the poor and thereby constituting a form of regressive taxation, they typically have low net yields because of the payment of prizes and high administrative costs. The typical net yield from lotteries is between 10-30% of gross receipts.

3.2.1.1.2 Private Financing

Private financing for health care can be direct or indirect

a) Direct payment: This is personal payments made directly to a wide range of providers, including private practitioners, traditional healers and private pharmacists. User fees, whether for government-provided or for privately provided health services, are an out-of-pocket payment and are therefore considered here as health finance from a private source. Similarly, charges to contributions or prepayments by members of community financing schemes are also considered as coming from private (non-government) sources.
b) Indirect payment: This is payments for health care services by employers (e.g. payment by large and privately owned industrial complexes in developing countries or sharing of health care costs by employers in industrialized countries) and health financing by other non-government bodies such as local charity fund-raising for health causes.

3.2.1.2 Health Insurance

(i) Private Health Insurance: Private health insurance differs from social insurance in two main ways. First, private health insurance typically does not include pensions for invalidity or old age. Second, the price (or premium) charged for private health insurance is not based on the pooled risks of a large population, but on personal risk characteristics and the likelihood of illness in the individual or group covered. As a result, premiums are likely to vary for different individuals or groups. Schemes may be profit or non-profit making and may be organized for individuals or groups, the letter often benefiting from lower premiums (resulting from lower per capita administration costs as well as a degree of risk-sharing). In many countries the larger

employers act as an organizing body for health insurance, and may pay part of the premium as a fringe benefit. However, in order to control the level of utilization of services, individuals are often required to pay for part of the cost of medical care on a direct fee-for-service basis. In countries where demand is sufficiently high, commercial insurance companies may be active.

Private insurance is not subject to the political allocation process and may channel extra funds into the health sector. However, it suffers from problems of two coverage because of its cost and the exclusion of bad risks, or enhancing inequity and promoting the growth of high-technology health care, inappropriate to developing countries.

(ii) **Employer-Financed schemes:** In some instances employers may directly finance health care for their employees. They may, for instance, pay for private sector health services, employ medical personnel directly, or provide necessary facilities and equipment. Oil companies, mining and mineral industries, and large-scale export-centered agricultural enterprises usually provide for the health needs of their workforce. Benefits are seldom extended to families as employers are primarily concerned with maintaining the productivity of the work force. In developed countries the primary focus is on accident prevention and occupational health, and in developing countries also, employers may have a legal obligation to provide first aid or occupational health services (e.g. sugar and coffee plantations in Latin America, tea and rubber estates in Asia and Cocoa farms and mines in Africa). Problems with employer-financed schemes relate to the quality of care provided, the possible fragmentation of services, difficulties enforcing employer liabilities, and the fact that viability depends upon the performance of the employing agency. Nowhere is employer finance a predominant source of support for health, although employer schemes are often a precursor to national social insurance schemes.

(iii) Charity and voluntary contributions: It can take the form of financial support or in-kind donations (e.g. personal services, physical facilities, equipment and supplies), and may originate from business enterprises, wealthy families, religious organizations or private individuals. Often these resources are channeled through foundations or religious bodies. The problem with this source of finance are often indirect for example, donors may have different priorities from the recipient nation and may not recognize their most urgent health needs. They prefer to finance visible evidence of their support such as physical facilities and equipment and thereby

committing the recipient, country or contributions may also take the place of, or reduce other sources of finance. For example, contributions may be eligible for tax relief; reducing general tax revenues for use elsewhere (the effects may be minor). Charitable contributions have played an important role in health services provision in the past, and in some African countries and are still major sources of health care finance, channeled through religious agencies. The general trend, however, is for governments to support or take over mission health services. Thus, the role of charitable and voluntary contributions is decreasing, although it may still be important in times of emergency and can be a useful supplement to other forms of health finance.

(iv) Community Financing and Self-Help: Primary health care initiatives in developing countries stress the importance of national self-reliance and community participation in health care delivery. By mobilizing underutilized national and local resources (e.g. organizational skills, manpower and cash) and by developing affordable and culturally appropriate delivery systems, it is hoped that basic health care will become universally accessible. Consequently, some governments and many non-governmental agencies are turning to communities for organization, participation and financial support and communal self-help is increasingly thought of as an important source of financial support for health services in developing counties. The challenge is to develop new types of local institutions that can coordinate; and for health services in developing countries to systematically utilize the community resources. Self-help can take many forms, such as labour, local insurance support for volunteer health workers, and drug cooperatives.

(v) Direct household expenditure: Household income is ultimately the source of most health care finance, but direct expenditure constitutes a specific category of financing that should be considered separately. Included in this category are any payments a consumer may make directly to health care providers such as fees for services, or prices paid for goods and supplies.

Direct household expenditure is not independent of other sources of finance. Government services may charge user fees (often nominal) for certain services. Even with insurance coverage, there is often a requirement for some degree of copayment, which tends to increase the amount that would otherwise have been spent on health. Health insurance benefits, moreover, may have an upper ceiling, with household requirements in excess of this level. The

extent to which these payments represent a real ability and willingness to pay for health care is, however, unclear. Willingness to pay does not necessarily reflect ability to pay. Current levels of household expenditure partly result from the existing pattern of government health care provision, and the limited access to free/cheap government health care (particularly in rural areas). People may use and buy non- government (e.g. mission, private, traditional) health care partly because they have no cheap or good quality government alternative. Low-income groups tend to delay use of health services until illness is severe, presumably in part to avoid payment, but such delay generally only increases the necessary expenditure. High health care bills may sufficiently undermine their economic position to push them further into poverty. Health care payments also sometimes displace expenditure for other basic necessities of life (e.g. food), because there is only limited ability to pay for the range of household needs. Utilization of, and payment for health services is, moreover, likely to depend heavily on the perception of their relevance to a specific health need and the extent to which they provide a service that people value. Use of traditional healers for example, may reflect a belief in the relevance of their treatments for certain diseases rather than a general willingness to pay for any type of health care. Perceptions of poor quality in government services certainly undermine their use and therefore, willingness to pay for them. Private services may be more oriented to the preferences and circumstances of households, for instance, providing for pay. Raising the level of direct household expenditure for health care, for example, by user fees, will clearly have a negative impact on equity (by influencing both the distribution of the payment burden and the benefited failed). It may be mitigated by the introduction of an exemption mechanism for the poor, although such a mechanism may itself reduce the demand for health care made by low-income groups because they may not wish to be identified as “poor”. Moreover, such willingness to pay as exists is attached primarily to curative services, and so can only extend the provision of preventive care if it is possible to re-allocate resources within the health sector. Finally, the potential yield from user fees is unclear. It is dependent on the level and type of fees, the bureaucratic structure required to implement them, the existence of exemption mechanisms, and the impact of fee systems on the demand for care and the rates of collection. The administrative

difficulties of implementing a fee system (e.g. how is ability to pay assessed? Who assesses it? Who collects the fees? How is abuse of the system restricted?).

(vi) Health Insurance: Provides the means by which risks or uncertain events are shared between many people. Premiums are paid to an insurance institution which compensates any insured victim of the event for any financial loss resulting from the event. Insurance therefore, helps to lessen and spread risks, and it relies on the fact that what is unpredictable for an individual is highly predictable for a large number of individuals. It follows that for insurance to be feasible, there must be enough individuals insured to spread the risks widely, and the uncertain events must be relatively independent of each other. That is, the principle is one of insurance based on probabilities, not one of prepayment for known future events; though in practice, a prepayment element for health care exists since certain types of utilization are highly predictable. For a health insurance scheme to be cost covering, the level of its premiums needs to be related to the statistical frequency with which the population covered requires care, and to the average cost of claims, plus an allowance for administrative costs and a profit margin (for commercial institutions). Insurance has redistributive consequences, their nature and magnitude depending on the financing of the schemes and the way in which premiums are assessed. Because the occurrence of the event being insured against is uncertain, some participants will draw out more than they pay in, thus resulting in redistribution from the healthy to the sick. Other distributive effects will depend on whether the insurance is organized privately or through collective mechanisms, and on the method of distributing the costs over the population.

Health insurance can be financed and organized in a variety of different ways. It can be purchased by an individual or group through the private market, from either profit or non-profit firms, and under these circumstances is conventionally termed private or voluntary health insurance. Health care itself would usually be delivered by independent providers, but sometimes by facilities owned by the insurer. In the case of private or voluntary health insurance, the level of an individual's premium would be based on the actuarially determined likelihood of illness of that individual. In contrast, group insurance is often based on a firm or co-operative and the premiums related to the risk of the group of employees in it, not of individuals. All subscribers will pay similar premiums and such insurance may well be made

compulsory by the firm to prevent low risk or high income employees opting out. In some countries (for instance the united states and Australia) there are examples of the imposition of community rating on private insurers; that is, within a given geographical area, premiums are not permitted to vary according to health risk or occupation. Premiums are often paid at least in part by employers, health insurance being considered a fringe benefit, though labour legislation making it compulsory for employers to provide their workers with some form of medical care is increasingly being introduced in developing countries.

SELF ASSESSMENT EXERCISE

Discuss various sources of health care financing.

3.3 Factors Influencing Health Care Financing

The form and level of health care financing are now major policy issues for most developing countries and it is essential that decision makers have a clear understanding of the implication of alternative approaches to financing health care. There is an increasing interest in how health services are funded in all countries. The following factors, among others, influence the health services sector and should be given due attention in health care financing.

(i) Demographic Changes: These have major effects on health care provision; firstly, demographic change may lead to variations in the health coverage of the population. Rapid population growth rates can cause tremendous strains on the provision of social services including health care. Secondly, the age structure of the population has an important significance to the provision of health care. There are higher health service unit costs associated with the young and the old. The antenatal, obstetric and under five age groups are all heavy users of health care, as are the elderly with their higher incidence rate of chronic illness. Third, demographic factor relates to the relationship between economic producers and dependents of a country. High dependent ratio means an increased burden on the productive population for providing health care.

(ii) Economic Recession: This can be expressed by low or even negative growth rates, increasing debt burdens and high inflation rates. This has severe implications for the ability of governments to maintain, let alone expand, expenditure on health care. Such effects on

the supply of health care are worsened by the increased need for health care brought about by the recession itself through the links between poverty and ill health.

(iii) Rising Expectation: The rising expectation of health care consumers especially the middle classes, to receive high-technology medical care similar to that available in the industrialized world is high.

(iv) Concerns about equity: Governments committed to the principles of primary health care have a major responsibility to improve levels and depths of coverage. The concerns for equity may influence the choice and system of health care financing. To extend basic health care at a time when there is such strong middle class pressure may only be available by providing substantial additional resources to the health sector.

(v) Disease-pattern changes: Disease-pattern change may result due to changes in average income levels or due to changes in social development. Thus, as standards of living rise and morbidity patterns change, these changes are likely to have an effect on health care financing. In addition to shifts in disease patterns, the advances of medical technology have led to the possibility of treatment for health problems previously accepted as untreatable. This again places further pressures on health-care providers.

(vi) Efficiency: Given the limited resources available for health in developing countries, it is essential to taste and use resources as efficiently as possible.

(vii) Displacement effects: Rather than generating additional resources for the health sector, new or expanded financing mechanisms may merely displace funding from other sources. Displacement is not necessarily an undesirable consequence if the new or expanded source of finance is more efficient or more equitable than the one it partially displaces. Examples of displacement effects include foreign assistance which may displace government support for health care; counter-funding often a precondition for foreign assistance, which may divert funds away from existing priority projects; health insurance schemes, which may in some instances displace earth than additional to the total of resources being allocated to health care (e.g. displacing direct payments); charitable contributions which may be withdrawn when other sources are developed; and government allocations which may be reduced when other sources of finance (such as user fees) are developed.

(viii) Wider effects of the health sector: Health sectors may account for a sizeable share of national resources and are often major employers. Consequently, the activities of the health sector may have spill-over effects on the economy as a whole. These include external effects on costs (e.g. inflation through the repercussions of high increases in stag pay); foreign exchange problems through heavy foreign borrowing for development projects or for payments for imports such as pharmaceutical or equipment opportunity costs such as the attraction of scarce manpower into the health sector at the expense of other professions, and disincentives to investment and employment (e.g. as a result of financing health services through high taxes on certain economic activities, enterprises or sectors). These external effects may also be positive as in the case or improved productivity resulting from reduced death and disability in the work force.

In selecting a system of financing health care some criteria should be used. The first three criteria outlined below are general, while the last two have particular importance within the context of primary health care:

a) Viability and ease of using the system: This implies bureaucracy and cost simplicity, social acceptability ad technical feasibility
b) Revenue generating ability: Net revenue minus earning ability = Revenue minus operating costs. The administration of user-changes for example, may include the costs of billing, accounting and the safe storage and collection of funds. Even where additional staff is not employed and existing staff are used, it implies an opportunity cost to the health service in terms of alternative activities which the staff could have been engaged in had they not been involved in the revenue generating scheme.
c) Effects on service provision: Systems of financing, for example which involve three parties – the patient, the provider and an insurance company – may lead to over-provision of certain services.
d) Effects on equity: That is equal access to care for those in equal need.

e) Participation in decision-making: This is a concept that stresses community participation which creates an opportunity for a direct relationship between the consumer and the provider; an example of a financing system suitable of such participation is user charges.

SELF ASSESSMENT EXERCISE

1. What are the factors that influence the choice of a financing system?
2. Show how demographic conditions can affect the choice of health care financing systems.
3. Explain the difference between public and private goods.

3.4 Health Financing in Nigeria

Health expenditure financing in Nigeria is made up of public and private sources. The public source of financing health expenditure is either part of the general public tax revenues or is collected from the health insurance in the form of contributions to the social health insurance. The private source of health expenditure funding can be divided into three parts: private health insurance, out-of-pocket payments and all other private means.

Out-of- pocket payments are usually mentioned under the heading of private sources but it may also supplement private as well as public health insurance services. Government, private sector (including households and donors) and private sources of which household contributes the most are the three principal main sources of health care financing in Nigeria. Table 3.3 shows that private health expenditure as a proportion of gross domestic product (GDP) is between 3.4% and 4.2% while public health expenditure as a proportion of GDP is between 1.3% and 1.9% from 1995 to 2012 respectively.

Table 3.3: Health Expenditure in Nigeria: 1995-2012

Years	**1995**	**2000**	**2005**	**2010**	**2012**
Private Health Expenditure (%of GDP)	3.4	3.0	4.7	4.1	4.2
Public Health Expenditure (%of GDP)	1.3	1.5	1.9	1.5	1.9
Public Health Expenditure (%of GOVT EXP)	8.4	4.2	8.3	5.5	6.7
Public Health Expenditure (% of TOTAL HEALTH EXP)	28.2	33.5	29.2	26.3	31.2
Total Health Expenditure (%of GDP)	4.7	4.6	6.6	5.6	6.1

Source: WHO World Health Statistics, Various Years.

It further shows that public health expenditure as a percentage of total government expenditure is between 4.2% and 8.4% while public health expenditure as percentage of total health is between 26.3% and 33.5% from 1995 to 2012. Total health expenditure as a percentage of GDP is about 4.6% and 6.6% between 1995 and 2012. This shows that public health expenditure was still below private health expenditure and public health expenditure as a proportion of total health expenditure was still around 30% in the periods considered. Table 3.4 further shows the structure of health care financing in Nigeria being made up of public-private mix. The public source of financing health care in Nigeria comprises public expenditures by federal government (FG), state governments (SG) and local governments (LG) on health care while private financing comprises households and firms' out-of-pocket expenditures on health, private health insurance, donor agencies or development partners' expenditures on health and health expenditures by departments of private firms. From the table, households are the major source of health care financing in Nigeria. Households' health expenditures range between 60% and 74% from 1998 to 2005 while public health expenditures are between 15% and 27%. Donor agencies and development partners financing are between 4% and 16% while that of other departments of private firms stay around 1%. Health insurance is mainly from private health insurance and constitutes about 2.4% of total health care financing. The increasing enrolment in social health insurance and emergence of micro-health insurance scheme indicates a possible improvement in the contribution of health insurance to health care financing in Nigeria.

Table 3.4: Structure of Health Care Financing in Nigeria: 1998-2005 (N'mn)

Years	**Public Health Expenditure**		**Out-of-Pocket Payments**				**Health Insurance**				**All other Private Means**				**Total**
			Households		Firms		Social		Private		Donor Agencies/Developments Partners		Firms		
		%		**%**		**%**		**%**		**%**		**%**		**%**	
1998	23,502.1	14.9	108720.0	69.3	2808.9	1.8	-	-	-	-	20,551.0	13.1	1,499.1	0.9	157081.1
1999	29,882.9	16.6	118,782.4	66.0	4,283.8	2.4	-		-	-	24,911.9	13.8	2,030.2	1.1	179,891.2
2000	40,391.3	18.8	129,872.1	60.4	7,238.1	3.4	-	-	-	-	34,899.0	16.2	2,808.7	1.3	215,209.1
2001	69,765.9	27.2	157,601.7	61.5	11,456.7	4.5	-	-	-	-	14,269.1	5.6	3,190.1	1.3	256,283.4
2002	60,211.9	21.6	183,598.4	65.9	13,836.4	4.9	-	-	-	-	17,104.0	6.1	3,981.5	1.4	278,732.2
2003	123,681.8	18.7	489,464.6	74.0	1,504.1	0.2	-	-	15,655.5	2.4	27,872.2	4.2	3,484.0	0.5	661,662.2
2004	208,207.9	26.4	518,070.3	65.7	1,591.9	0.2	-	-	18,788.9	2.4	36,037.9	4.6	6,026.8	0.8	788,723.9

2005	254,174.4	26.0	656,115.8	67.2	2,016.2	0.2	-	-	21,335.4	2.2	36,296.7	3.7	6,749.1	0.7	976,687.6

Source: Soyibo, Olaniyan and Lawanson, 2009.

SELF ASSESSMENT EXERCISE

Discuss the various sources of health financing in Nigeria.

4.0 CONCLUSION

Every country wants a health care system that offers good health outcomes, affordable services, satisfied consumers and providers, and medical and financial equity. These objectives are hard to attain, when budget constraints are binding at low levels of overall expenditure, in particular in the public sector. Because health expenditures are largely out of pocket in low-income countries and there is limited capacity to increase domestic public expenditures, donors are expected to finance most of the scale-up. But even if donors make long term commitments, health expenditures will eventually have to be absorbed within each country's domestic resource.

5.0 SUMMARY

This chapter dealt with health care finance and various sources of health care financing. It examined mechanisms for increasing resources for health and the major restrictions on each method in countries. Public and private financing arrangements for pooling health care revenues are also discussed. Insurance markets suffer from market failures, particularly those associated with imperfect information. These problems are often described as those of moral hazard and adverse selection (those who anticipate needing health care will choose to buy insurance than others, which leads to higher costs, lower profits, higher premiums and fewer users).

6.0 Tutor-Marked Assignment

1. Describe the different systems of financing the health service sector.
2. Explain the difference between public and private goods.
3. Outline the problems of health insurance as a system of health care financing.
4. What are the drawbacks of private financing?
5. Discuss the weaknesses of government financing.
6. Are communities financing schemes applicable to Nigeria?

7.0 References

Donaldson Cam and Karen Gerard (1993) Economics of Health Care Financing: The Visible Hand. Macmillan Press Ltd. London.

Folland S., A. Goodman & M. Stano (2010) The Economics of Health & Health Care, Sixth

Edition, Prentice Hall, New Jersey.
Jacobs, P. (1991) The Economics of Health and Medical Care Maryland: Aspen Pub Inc. Jack,
Phelps Charles E. (1992) Health Economics, New York: Harper Collins Pub Inc.
Zweifel P., F. Breyer & M. Kifmann (2009) Health Economics, Second Edition, Springer Verlag Heidelberg.
Soyibo, A., Olaniyan, O.A. and Lawanson A.O. (2009) "National Health Accounts of Nigeria, 2003 – 2005 Incorporating Sub-National Health Accounts of States. VOl.1. Federal Ministry of Health, Abuja.
World Health Organization World Health Statistics, Various Years.

Unit 3: The Role of Government in Health Care

CONTENTS

1.0 Introduction

In recent years' health reform has shot up to the top of political agenda throughout the world. For developed industrial countries and many middle-income developing countries, reasons include rapidly rising costs, the large number of people still not covered by health insurance and the fear of AIDS. For developing countries, the main reason is a better understanding of the importance of health for improving the productivity of workers and of the potential for enormous gains in health at very low cost. Governments all over the world have played a vital role in bringing about the great advances in health over the past many years. Public health measures are responsible for eradicating smallpox and have been central to the reduction in deaths caused by other vaccine-preventable childhood diseases. Expanded and improved clinical care by government doctors and nurses has saved millions of lives from infectious diseases and injuries. Better prenatal and delivery services organized by governments have lowered the rate

of serious complications of pregnancy and childbirth for millions of mothers. Despite these remarkable improvements, however, enormous health problems remain. Absolute levels of mortality in developing countries are still unacceptably high; child mortality rates are about ten times higher than those in the established market economies. According to the World Bank Development Report in 1993, if death rates among children in poor countries were reduced to those prevailing in the rich countries, 11 million fewer children would die each year. Almost half of those preventable deaths are as a result of diarrheal and respiratory illness exacerbated by malnutrition. In addition, every year seven million adults die of conditions that could be inexpensively prevented or cured; tuberculosis alone causes two million of these deaths. Over 400,000 women die from the direct complications of pregnancy and childbirth. Maternal mortality ratios are on average 30 times as high in developing countries as in high income countries. There are several major problems with the way health systems are now run and financed and if solutions are not found, the pace of progress in reducing the burden of premature mortality and disability will be slowed. The appropriate nature and extent of government involvement will vary from country to country, in part depending on income levels.

2.0 Objectives

At the end of this unit, the students will be able to:

(i) Understand the role of government as affecting the resource allocation pattern in health and the extent to which it can influence the overall performance of the sector.

(ii) Analyze the possible measures that can be taken to alleviate the health problems of developing countries.

(iii) Appreciate the problems of health policy in developing countries.

3.0 Main Content

1.1 Problems of Health Policy

Some of the common problems of most countries in their policy are misallocation, inefficiency and cost allocation.

a) Misallocation: One of the most important aspects of economics in making health policy is the appropriate allocation of material, financial and human resources. This implies optimal

distribution of economic resources among competing needs. This calls for the proper identification of the need. Sometimes public money is spent on health interventions with low cost effectiveness such as cancers and critical and highly cost effective interventions such as treatment of tuberculosis and sexually transmitted diseases remain under funded.

b) Inequity: The poor lack access to basic health service and receive low quality care. Government spending for health goes disproportionately to the affluent in the form of subsidies to sophisticated public tertiary care hospitals and to private hospitals.
c) Inefficiency: Much of the money spent on health is wasted because brand name pharmaceuticals are purchased instead of generic drugs, health workers are badly deployed and supervised and hospital beds are under-utilised
d) Cost explosion: In some middle income developing countries health care expenditures are growing much faster than income as increasing number of specialists, the availability of new medical technologies and expanding health insurance linked with fee-for-service payments together generate a rapidly growing demand for costly tests, procedures and treatments. As countries alike rethink the best way to provide health care in the century ahead some argue that governments should step up their financing while allowing more participation by non-government organizations and the private sector in supplying services.

SELF ASSESSMENT EXERCISE

1. Describe the major problems of health systems in developing countries.

3.2 Solutions to the Problem of the Health Sector

Governments need to be involved in finding solutions to the problem of the health sector based on the problems mentioned. The poor cannot always afford the health care that would improve their productivity and well-being. Some actions to promote health are pure public goods or care large positive spillover effects. Market failures in health insurance also mean government intervention can raise welfare by improving the way those markets function. Clearly, governments have a responsibility to spend wisely and to evaluate carefully exactly what form their involvement should take. The World Bank recommends four main policies to overcome the existing weakness of health systems in developing countries.

(i) Governments should finance a nationally defined package of essential public health and clinical care, especially for the poor, and should ensure the widespread and efficient delivery of such a package.

(ii) The public sector should devote far fewer resources or none at all, to financing health services outside of the essential package which are of lower cost effectiveness.

(iii) Governments should promote such types of health insurance that not only achieve broad coverage of the population, but also build in payment mechanisms that control the cost of health services.

(iv) Governments should encourage diversity and competition in the supply of health inputs, particularly drugs, supplies and equipment, as a means of improving quality and driving down costs. They should also foster a competitive private sector to provide the full range of health services including financial publicity.

3.2.1 A Basic Health Package

Government action in many areas of public health has already had an important payoff. The challenge now is to expand coverage of interventions with high cost-effectiveness. School based health services information on family planning and nutrition programs to reduce tobacco and alcohol consumption regulation, information, public investments to improve the household environment and AIDS prevention could be explored. At the same time, governments should also put together a package of essential clinical services, depending on local needs and the level of income. The World Bank Development Report (1993) has come out with a suggested minimum package of health services which is affordable by developing countries at their levels of health spending and would reduce the burden of disease by just over 30 percent in low income countries. Eleven clusters of interventions or individual interventions are included in the package, apart from being cost-effective these services address diseases responsible for a large share of the disease burden in developing countries.

However, the exact content of each country's essential package will be largely determined by the epidemiology profile of the country (the distribution of disease burden across diseases) and the cost effectiveness of the corresponding interventions. The size of the package (number of

intervention cluster) will depend on the financial resources available for health care. Clustering interventions improve cost-effectiveness through at least three mechanisms:

(i) Synergism between treatments or prevention activities is common, particularly in pediatric care.

(ii) Joint production costs can substantially reduce the amount of resources needed were interventions to be provided separately.

(iii) The optimal use of specialized resources, such as hospital beds, requires a screening process to refer the most severe cases from the first level of care to other facilities.

An efficient health cluster should include interventions that can be given to the same individual, at the same time, and through the same mode of delivery (outreach community health worker, health center or hospital).The expanded program on immunizations, for example, is a very efficient one because it includes six vaccines provided through the same delivery system to the same individuals ,often at the same time hence, an essential health package approach is an important measure which governments can be encouraged to do.

3.2.2 Value for Money

When it comes to health policy, one of the most difficult decisions for developing nations to make is how best to put together a mix of health services that will be financed by public spending. Ideally, they would like to offer as much health care as possible to as many people as possible. In practice, however, they end up concentrating resources in urban hospitals which provide a wide range of services for a few, leaving other population groups, particularly in rural areas, with a relatively little access. This allocation of public spending is inequitable and inefficient. Costly treatments are prescribed that prolong life only slightly, while large populations are denied inexpensive services that extend life greatly, such as immunization. As a way-out of resolving this problem, the World Bank Development Report recommends that governments design and finance national health package embracing essential public health and clinical services that will substantially reduce the burden of disease (the present value of future streams of disability-free life lost as a result of death, disease or injury) at affordable costs. This

means that government will need to review the value of interventions they offer so that they can reallocate resources in the most cost effective manner.

No matter how health services are organized and paid for, what they actually provide are health interventions. Debates about whether health services should concentrate on “vulnerable groups” such as children, pregnant women and the elderly, or about the relative role of hospitals versus health centers, or about preventive versus curative activities are at bottom debates concerning the proper mixture of interventions. In health, as in every other sector, customers want value for the money spent. That is why the first step in designing a country’s essential health package is to determine the cost effectiveness of a health intervention- the net gain in health compared with doing nothing divided by the cost. Indeed, the developing countries that have been the most successful in improving health for a given level of spending have concentrated their public monies on highly cost effective interventions.

3.2.3. Redirecting Public Spending

The World Bank Development Report pointed out the need for widespread and fundamental reform of health policies and health systems. It called for changes in the level and composition of government spending for health in public and private institutions responsible for delivering health services and in insurance, cost recovery and mechanisms for financing health care. Public financing of an essential clinical package can be justified because the package creates positive spillover effects and reduces poverty. However, the case for government financing of discretionary clinical health services outside of the essential national package is far less compelling. In fact, if governments reduced or eliminated public funding of these services they would actually increase in both efficiency and equity. One important way to direct government spending away from discretionary care is to recover costs in government hospitals especially from the wealthy and the insured. Even in low- income countries such as Nigeria, Kenya, Pakistan and the Philippines where insurance may account for less than 5 percent of total health spending, a combination of limited private insurance and the ability of upper income groups to pay makes feasible for governments to charge for discretionary care delivered in public

hospitals. In middle – income countries, where insurance becomes more important, there is ever-greater potential for cost recovery.

Governments should also phase out public subsidies to insurance which generally benefit the better-off. There are strong efficiency arguments for directing government funding to public health interventions because of the public good nature of these services and a number of the essential clinical services, including treatment to tuberculosis effects. In addition, there are equity grounds for financing the basic health package. The poor are disproportionately affected by the disease burden the package addresses. This means that making public financing of this package with universal government finance leads to public subsidies to the wealthy, who can afford to pay for their own services, with the result that fewer government resources go to serve the poor. One way to solve this problem is by targeting public spending to the poor. In low-income countries, where current public spending for health is less than the cost of the minimum package, some targeting is almost inevitable. In countries where the wealthy do not use government financed services because of the greater quality and convenience of privately financed services, targeting may be fairly easy. The most sophisticated facility required to deliver the minimum elements of the package is a "district" hospital which serve as the first level of referral from health centers. These hospitals offer basic surgery, emergency services and some outpatient care. Generally, they can have 100-400 beds and serve 50,000-200,000 inhabitants; the minimum package requires access to health centers and district hospitals throughout the country. On the average it requires about 1 district hospital bed, 0.1 to 0.2 physicians per 1,000 population and 2 to 4 nurses per physician. Governments can direct public spending to support the nationally defined essential package in several ways:

a) Where services are publicly financed and provided, government can reallocate public spending towards inputs-drugs, supplies and equipment, staff and facilities that support the package. In many countries extending lower-level facilities are necessary steps to delivering the package. At the same time, governments can eliminate or greatly reduce financing of inputs for less cost-effective services. This might include losing wards or converting specialized hospital physicians.

At the same time, provider's treatment decisions would not be micromanaged; they would be influenced by the nature of input availability. The specialized staff and equipment for example, would be available for treating malaria in young children. Budgetary and salary incentives could also be used to reward individual providers, facilities that achieved good coverage of the population with the services in the package. Where services are publicly financed, but privately provided, governments should reimburse only for those services in the essential package.

b) Where services are publicly financed, but privately provided, governments should reimburse only for those services in the essential package. This model of health care delivery is growing. It is still uncommon, however, in developing countries. At present the regulatory capacity to oversee such arrangements is poorly developed

3.2.4 Controlling Costs

Where subsidies in discretionary clinical services for the better-off are cut or public insurance is universal and pays for a more comprehensive set of services in the national package, governments must cope with the problem of escalating health care costs. These costs can crowd-out spending on other sectors of the economy or raise the price of labour threatening a country's international competitiveness. The sources of excess health costs are complex and much debated. Health services are labour intensive, and their productivity grows slowly compared to the other areas of the economy. In the United States, higher levels of underlying morbidity and greater hospital amenities relative to other industrial countries are part of the answer. But two types of inefficiencies are also important, high administrative costs and unnecessary use of an ever-expanding array of costly technologies of diagnostic tests and surgical procedures. These inefficiencies appear to be linked to two basic features of the US health system. Open-ended free-for all service compensation for health providers encourages the development of new equipment, drugs and procedures since neither providers nor patients have strong incentives to hold down utilization or spending. A complex system of multiple insurance institutions and other payers, each with its own procedures, raises administrative costs substantially.

These findings concerning health costs escalation in industrialized countries are especially relevant for middle-income developing countries which are under pressure from medical professionals, manufactures and consumers to use new medical techniques. They face difficult policy choices related to provider compensation. One approach to controlling health costs is to pay a fixed amount for each person (capitation) as is done by health maintenance organizations. Another approach used in several industrial countries is to provide each hospital or network of physicians with a fixed total budget. In countries where there is expanded insurance system, insurers can jointly negotiate uniform fees for physicians or they can set fixed payments for specified medical procedures.

3.2.5 Promoting Competition

Although governments have a fundamental responsibility for financing basic health services, they need not be responsible for delivering those services. Experience suggests that diversity and competition lead to better results. In a competitive system people seeking health services can choose from a variety of providers-public, private non-profit and private for-profit. As developing countries move towards such a system, they face a wide range of policy options as regards the impact of different providers (Public and private) in terms of allocative efficiency, technical efficiency and the potential to reach the poor. Non-governmental organizations (NGOs) provide a major share of health services in developing countries especially for low-income households in the poorest countries. Data from Africa suggested that the NGOs are often more efficient. Governments that have excluded NGOs or heavily restricted their operations have seen essential services deteriorating. Where such bans or barriers to NGO activity exist, they should be removed. Beyond this, there are opportunities for governments to form constructive partnerships with NGOs to deliver essential clinical services. Some governments in Africa, such as Tanzania and Lesotho, allow appropriately located religious mission hospitals to serve as district hospitals. They then make them responsible for a full range of public health and clinical services and for performing district wide functions such as the health planning, supervision of lower-level clinics and community activities as well as the maintenance of emergency transport. In return, the government pays some of the NGO costs.

In Africa and Asia where traditional medicine remains an important part of the health care system, governments could make greater use of traditional practitioners. Successful examples include the use of healers in Thailand to screen for malaria and distribute anti-malarial drugs and the promotion of modern contraceptives in Kenya. Traditional birth attendants have also been enlisted to improve pregnancy outcomes in Bangladesh. At the same time governments can improve the equity and efficiency of their own health programs and facilities thereby increasing their responsiveness to local needs through decentralization and the use of managerial incentives. Governments can also reap efficiency gains by converting public hospitals into semi-autonomous foundations or public enterprises. These foundations are less restricted by public sector procedures in managing their costs and collecting charitable donations for investments and operational costs. Government finance of public health and of a nationally defined package of essential clinical services would leave the remaining clinical services to be financed privately or by social insurance within the context of a policy framework established by the government. Governments can promote diversity and competition in the provision of health and insurance by adopting policies that:

(i) Encourage social or private insurance (with regulatory incentives for equitable access and cost containment) for clinical services outside the essential package.

(ii) Encourage suppliers (both public and private) to compete to deliver clinical services and provide input, such as drugs, to public and privately financed health services. Domestic suppliers should not be protected from international competition.

(iii) Generate and disseminate information on provider performance, on essential equipment and drugs, on the costs and effectiveness of interventions and on the accreditation and status of institutions and providers.

Increased scientific knowledge has accounted for much of the dramatic improvement in health that has occurred during the 20th century by providing information that forms the basis of household and government action and by under printing the development of preventive, curative

and diagnostic technologies. Investment in continued scientific advances will amplify the effectiveness of each element of the suggested three-pronged approach.

3.2.6. Strengthening Household Capacity

Within the household, health improves as people escape poverty and get better education. Beyond the household, every society's health services are affected by its national income and its ability to acquire and apply new scientific knowledge which depends on the level of schooling.

3.2.6.1 The role of income

Life expectancy is believed to be strongly associated with income per capital. The higher a country's income per-capita, the more likely its people are to live long, healthy lives. Income growth has more impact in poor populations because additional resources buy basic necessities, particular food and shelter that yield especially large health benefits. The relationship between income and life expectancy has improved over the course of the century as advances in science and medicine have made it increasingly possible to realize greater health for a given income.

Because poverty has a powerful influence on health, it is not just income per capital that is relevant. The distribution of income and the number of people in poverty matter as well. In industrial countries life expectancy depends much more on income distribution than on income per capital and it has been rising faster in countries with improving income distribution. In developing counties, the variation in the prevalence of poverty and per capital public spending on health goes a long way toward explaining differences in life expectancy. Moreover, the adverse effect of poverty on health can be seen in health differences across rich and poor neighborhoods and families, even within the same city. The strong link between income level and health highlights the costs to health of slow economic growth.

3.2.6.2 The Role of Education

Households with more education enjoy better health for both adults and children. A mother's schooling is a powerful determinant of child health. The advantages that a mother's schooling confers on her children's health are felt even before birth and they continue to operate throughout the childhood years. Better-educated mothers marry and start their families later diminishing the health risks of early childbearing. They also tend to practice better domestic

hygiene and make more effective use of health services. In general, they are better at getting information on health and acting on it. Among adults, health depends strongly on personal habits and lifestyles. Since educated people tend to make choices that are better for their health, there is a strong relation between schooling and health. In Brazil, adults with primary schooling or less are about five times as prone to high blood pressure as those with post-secondary schooling. Educated people are quick to modify their behavior as new health threats arise (such as AIDS) or in response to new information about health. In the United Kingdom, for example, the proportion of smokers among adults declined by 50 percent between 1958 and 1975 among the most educated, but hardly changed among the least educated.

Given these strong links between better health and income and education, the policy implications are clear; governments should work to boost economic growth, reduce poverty and expand schooling (especially for girls – one of the most effective ways of strengthening women's ability to care for their families). It is difficult to reduce poverty and thereby improve health status without economic growth, so establishing sound economic policies is one of the most valuable things a government can do.

SELF ASSESSMENT EXERCISE

What are the possible solutions by governments to the problems of the health sector?

4.0 Conclusion

Governments have played a vital role in bringing about the great advances in health over the years. Despite these efforts, enormous health problems remain. There are several major problems with the way health systems are now run and financed and if solutions are not found, the pace of progress in reducing the burden of premature mortality and disability will be slowed. The appropriate nature and extent of government involvement varies from country to country, in part depending on income levels. Some of the common problems of most countries in their policy are misallocation, inefficiency and cost allocation. Governments have to engage in finding solutions to the problem of the health sector based on the problems mentioned. The poor cannot always afford the health care that would improve their productivity and well-being. Some actions to promote health are pure public goods or care large positive spillover effects. Market

failures in health insurance also mean government intervention can raise welfare by improving the way those markets function. An efficient health cluster should include interventions that can be given to the same individual, at the same time, and through the same mode of delivery.

5.0 Summary

This unit discussed different problems of health policy and various solutions to the problem of the health sector. It also considered how best to put a mix of health services together when making health policy. It stressed the need to efficiently utilizes available fund and ensure equity in financing disbursement. This is because inefficiency may make reforms efforts costly.

6.0 Tutor-Marked Assignment

1. Provide an overview of Nigerian Health Policy from 1960 – date.
2. What is the relevance of encouraging diversity and competition in improving the health system?
3. Describe the implication of "Value for money" with respect to the purchasing process of government organizations.
4. Discuss the general strategies in the health sector development programme of Nigeria.

7.0 References/Further Readings

Donaldson Cam and Karen Gerard (1993) Economics of Health Care Financing: The Visible Hand. Macmillan Press Ltd. London.

Folland S., A. Goodman & M. Stano (2010) The Economics of Health & Health Care, Sixth Edition, Prentice Hall, New Jersey.

Jacobs, P. (1991) The Economics of Health and Medical Care Maryland: Aspen Pub Inc.

Jack, Williams (1964) Principles of Health Economics for Developing Countries. WBI Development Studies. The World Bank, Washington D. C.

Jones Andrew (2007) Applied Econometrics for Health Economists: A Practical Guide, 2nd Edition OHE

Phelps Charles E. (1992) Health Economics, New York: Harper Collins Pub Inc.

Santerre E. & S.P. Neun (1996) Health Economics: Theories, Insights & Industry Studies, Irwin, Chicago.

Zweifel P., F. Breyer & M. Kifmann (2009) Health Economics, Second Edition, Springer Verlag Heidelberg.