DEVELOPMENT OF AN INSTRUMENT FOR THE EARLY DETECTION OF THYROID ORBITOPATHY

BY

MARK LINDER, B.Sc.H.

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE
IN
THE FACULTY OF GRADUATE STUDIES

(Department of Health Care and Epidemiology)

We accept this thesis as conforming

to the required standard

The University of British Columbia
July 2000
© Mark Linder, 2000
In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Health Care and Epidemiology

The University of British Columbia
Vancouver, Canada

Date July 31, 2000
Abstract

Graves' Hyperthyroidism (GH) is an autoimmune disease that affects 1 in every 100 Canadians according to the Thyroid Foundation of Canada. A significant number of these patients go on to develop another closely related autoimmune condition called Thyroid Orbitopathy (TO). In this disease, the extra-ocular muscles become inflamed and enlarged beyond their normal size, leading to a variety of clinical changes that may include (but are not limited to) swelling, pain, exophthalmos, strabismus and double vision. Patients suffering from moderate or serious cases of TO are most often referred to an ophthalmologist at a relatively advanced stage of disease. At that point, major interventions such as surgery or orbital radiotherapy are usually required. Although these treatments are usually quite effective, the patient is rarely restored to the quality of life enjoyed before getting sick. However, if patients are identified early in the disease process, potential therapeutic interventions exist which may lessen or prevent the development of serious disease.

The objective of this study was to design and develop a symptom-based self-administered clinical screening rule for the early diagnosis of thyroid orbitopathy. The study took place between September of 1998 and July of 2000 and was broken down into two distinct phases. In the first phase, a symptom-based questionnaire was designed, developed and refined through literature search, expert consultation and focus groups. In the second phase, an inception cohort of patients with Graves' Hyperthyroidism was recruited and given a questionnaire as well as a standard TO clinical exam by participating ophthalmologists. The resulting data was then analyzed in order to identify symptoms related to presence of disease. Data collection began in January of 1999, and continued until April of 2000. A total of 50 patients with GH, referred from a single endocrinologist practice, participated in the study.

The final rule, dubbed the Vancouver Orbitopathy Rule, was designed with a view to maximizing sensitivity. As a result, the final diagnostic analysis yielded a sensitivity of 1.0, a specificity of 0.78 and a positive predictive value of 0.60. However, these parameters were measured based on the same data that was used to design the rule. In order to calculate the true measure of the test's clinical utility and diagnostic accuracy, a new validation study must be carried out.
Table of Contents

ABSTRACT .............................................................................................................................. ii
TABLE OF CONTENTS ........................................................................................................... iii
LIST OF TABLES .................................................................................................................... v
LIST OF FIGURES .................................................................................................................. vi
ACKNOWLEDGEMENTS ......................................................................................................... vii

1.0: INTRODUCTION ........................................................................................................... 1

1.1 GRAVES’ HYPERTHYROIDISM ....................................................................................... 1
  1.1.1 Pathogenesis and Clinical Presentation ................................................................. 1
  1.1.2 Other Causes of Hyperthyroidism ........................................................................ 2
  1.1.3 Epidemiology of Graves’ Hyperthyroidism .......................................................... 3
  1.1.4 Clinical Management of Graves’ Hyperthyroidism ................................................. 4

1.2 THYROID ORBITOPATHY ............................................................................................. 6
  1.2.1 Thyroid Orbitopathy and Graves’ Hyperthyroidism .............................................. 6
  1.2.2 Classification of Thyroid Orbitopathy ................................................................. 6
  1.2.3 Diagnosis of Thyroid Orbitopathy ...................................................................... 10
  1.2.4 Treatment of Thyroid Orbitopathy ..................................................................... 11
  1.2.5 Epidemiology of Thyroid Orbitopathy ............................................................... 13

1.3 DIAGNOSTIC TESTING ............................................................................................... 16
  1.3.1 Introduction ......................................................................................................... 16
  1.3.2 Properties of diagnostic tests ............................................................................. 17
  1.3.3 Diagnostic Decision Analysis ............................................................................. 22
  1.3.4 Evaluation of diagnostic tests ............................................................................. 25
  1.3.5 Subjective Diagnostic Tests ................................................................................. 28
  1.3.6 Screening Programs ............................................................................................. 30
  1.3.7 Thyroid Orbitopathy and Screening .................................................................. 35

1.4 INSTRUMENT DEVELOPMENT AND DESIGN ......................................................... 36
  1.4.1 Introduction ......................................................................................................... 36
  1.4.2 Desirable properties of an instrument ................................................................. 37
  1.4.3 Steps in instrument development ....................................................................... 42

1.5 STUDY RATIONALE .................................................................................................... 49

1.6 HYPOTHESIS ............................................................................................................... 49

2.0: STUDY DESIGN .......................................................................................................... 50
2.1 INTRODUCTION ................................................................. 50
2.2 QUESTIONNAIRE DESIGN AND DEVELOPMENT ......................... 51
  2.2.1 Devising the items .................................................. 51
  2.2.2 Scaling Responses and Selecting the items ....................... 52
  2.2.3 Final Draft ....................................................... 54
2.3 INCEPTION COHORT RECRUITMENT DESIGN ARCHITECTURE .......... 55
  2.3.1 Patient Selection and Referral .................................. 55
  2.3.2 Questionnaire Administration .................................... 55
  2.3.3 Clinical Evaluation ............................................... 56
  2.3.4 Database Organization ............................................ 58
2.4 STATISTICAL ANALYSIS .................................................. 59
  2.4.1 Sample Size Considerations ..................................... 59
  2.4.2 Questions used .................................................. 59
  2.4.3 Chi-squared tests ................................................ 60
  2.4.4 Development of the Decision Rule ............................... 62
3.0: RESULTS ........................................................................... 63
  3.1 PARTICIPATION RATE ................................................... 63
  3.2 DESCRIPTIVES OF STUDY POPULATION ............................... 64
  3.3 SIGNIFICANT SYMPTOMS .............................................. 65
  3.4 FINAL RULE CONSTRUCTION .......................................... 66
    3.4.1 Limitations of the data ......................................... 68
4.0: CONCLUDING REMARKS .................................................. 69
  4.1 CONCLUSIONS .......................................................... 69
  4.2 SOURCES OF ERROR .................................................. 71
  4.3 LIMITATIONS OF THE STUDY ......................................... 72
  4.4 CURRENT AND FUTURE STUDIES ..................................... 73
5.0: REFERENCES .................................................................... 75
APPENDIX 1 ........................................................................ 81
APPENDIX 2 ........................................................................ 83
APPENDIX 3 ........................................................................ 88
APPENDIX 4 ........................................................................ 90
APPENDIX 5 ........................................................................ 93
APPENDIX 6 ........................................................................ 96
List of Tables

1.1 Common Clinical Manifestations of Thyrotoxicosis .................................................. 2
1.2 Detailed Classification of Eye Changes of Graves' Disease (NOSPECS) ..................... 8
1.3 The 10 items of the clinical activity score (CAS) ...................................................... 10
1.4 Complications of corticosteroid therapy ................................................................. 13
1.5 Diagnostic Test 2 X 2 table ...................................................................................... 18
1.6 Terminology ............................................................................................................. 18
1.7 Principles of diagnostic decision analysis .................................................................. 22
1.8 Eight guides for deciding the clinical usefulness of a diagnostic test ....................... 25
1.9 Seven guides in the evaluation of a screening program ............................................. 32
1.10 Characteristics of a Good Screening Test .................................................................. 33
2.1 Chi-Square Test ....................................................................................................... 60
2.2 Expected Values in the Chi-Square Test .................................................................. 61
3.1 Ethnic Background ................................................................................................. 64
3.2 Gender Distribution ............................................................................................... 64
3.3 Age Distribution ..................................................................................................... 64
3.4 Smoking Behaviour ............................................................................................... 64
3.5 Distribution of disease ............................................................................................ 64
3.6 Disease Status ........................................................................................................ 64
3.7 Significant Symptoms ............................................................................................. 65
3.8 Final Rule ................................................................................................................ 66
3.9 Measures of Diagnostic Accuracy ............................................................................ 67
List of Figures

1.1 ROC Curve ................................................................. 21
1.2 Diagnostic Trees ......................................................... 23
1.3 Diagnostic Trees for radionuclide angiography .................. 24
1.4 Natural History of Disease ............................................. 30
1.5 Validity and Reliability .................................................. 39
Acknowledgements

First and foremost, I would like to offer a warm thanks to Dr. Jack Rootman who not only conceived of, initiated and funded this project, but provided a steady source of support, dialogue and open constructive criticism throughout the course of the study.

Secondly, thanks to Dr. Martin Schechter who always managed to squash in time to address my questions or concerns promptly despite often being several time zones away.

Thirdly, to the remaining members of my committee, Dr. Peter Dolman and Dr. Joel Singer, and to Dr. Ted Wilkins and Dr. Peerooz Sayeed. Thank you for your advice and support throughout this project.

Fourthly, great thanks are due to the wonderful administrative staff in both the departments of Health Care & Epidemiology and in the department of Ophthalmology. In particular, thanks to Daniella Ciucci for providing assistance and humour whenever it was most needed.

Finally to my parents Barbara and Jeffery and to my brother Glen, thank you for always being there for me. I couldn’t have done this without you
1.0: Introduction

1.1 Graves' Hyperthyroidism

1.1.1 Pathogenesis and Clinical Presentation

Graves' Hyperthyroidism (GH) accounts for 70-85% of all cases of hyperthyroidism. In this condition, the immune system targets the Thyroid Stimulating Hormone receptor (TSH-R) of the thyroid gland. Antibodies are formed which constitutively bind to the receptor, and activate it. The result of this interaction is an unregulated release of thyroid hormone (thyroxine or T4) as well as growth of the thyroid gland which can drastically alter the metabolism of the patient. (Table 1).

High circulating levels of thyroid hormones can affect almost all organ system. Typically, hyperthyroid patients are gaunt and restless, talk rapidly, and display emotional lability. Other classic signs and symptoms include sweating, heat intolerance, palpitations, insomnia, and warm, fine skin. Prominent eyes or a stare may be produced by increased thyroid hormone levels, but infiltrative eye signs signal the presence of Graves' disease. It is important to note that thyrotoxicosis may also present in an atypical fashion. Older patients in particular tend to show fewer findings and may present with only weight loss, cardiac symptoms, or a change in mental function. Some patients have "apathetic" hyperthyroidism and lack almost all of the usual clinical manifestations of thyrotoxicosis. Their behaviour may lead one to consider, erroneously, a psychiatric diagnosis.

Differential diagnosis of GH from other forms of hyperthyroidism (see section 1.2) is made possible through laboratory tests and thyroid imaging. In particular, laboratory tests such as Radioactive Iodine (RAI) uptake, Serum Total Thyroxine (T4) and Serum TSH may be used to determine presence or absence of disease. Presence of GH is likely when RAI-uptake and serum T4 are elevated and TSH level is low. Diagnostic imaging through the use of radionuclides, ultrasound or magnetic resonance imaging enables clinicians to determine degree of enlargement and presence of nodules or tumours in the gland.
Graves' Hyperthyroidism was named for Robert J. Graves, the 19th century Irish physician who first described patients with the particular symptom set that characterized the disease. Several years later, a German clinician by the name of Karl Von Basedow went further in describing the "Merseburg triad" of goiter, exophthalmos and palpitations. In recognition of these observations, the disease has been variously labeled Graves' Disease or Graves-Von-Basedow Disease. The term Graves' Disease, distinct from GH, may refer to both ophthalmic and non-ophthalmic forms of the disease. (The ophthalmic form of the disease-Thyroid Orbitopathy-is introduced in section 1.5).

Table 1.1 Common Clinical Manifestations of Thyrotoxicosis

<table>
<thead>
<tr>
<th>SYMPTOMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nervousness</td>
</tr>
<tr>
<td>Fatigue</td>
</tr>
<tr>
<td>Increased perspiration</td>
</tr>
<tr>
<td>Heat intolerance</td>
</tr>
<tr>
<td>Tremor</td>
</tr>
<tr>
<td>Hyperactivity</td>
</tr>
<tr>
<td>Palpitation</td>
</tr>
<tr>
<td>Appetite change (usually increase)</td>
</tr>
<tr>
<td>Weight change (usually loss)</td>
</tr>
<tr>
<td>Menstrual disturbances</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GENERAL SIGNS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hyperactivity</td>
</tr>
<tr>
<td>Tachycardia or atrial arrhythmia</td>
</tr>
<tr>
<td>Systolic hypertension</td>
</tr>
<tr>
<td>Warm, moist, smooth skin</td>
</tr>
<tr>
<td>Stare and eyelid retraction</td>
</tr>
<tr>
<td>Tremor</td>
</tr>
<tr>
<td>Hyperreflexia</td>
</tr>
<tr>
<td>Muscle weakness</td>
</tr>
</tbody>
</table>

*adapted from Werner and Ingbar's The Thyroid, pg 524

1.1.2 Other Causes of Hyperthyroidism

Other relatively common causes of hyperthyroidism include thyroiditis, toxic multinodular goiter and toxic adenoma. These account for the majority of the non-Graves cases, and are described very briefly here. For a fuller description see Falk1.
Thyroiditis patients usually experience transient hyperthyroidism as a result of the release of thyroid hormones from the thyroglobulin stores in the gland, due to inflammatory processes. This kind of disease is distinguished from GH by a low RAI-uptake, and accounts for between 5% and 25% of cases of hyperthyroidism.

Toxic multinodular goiter usually occurs in older patients, and can often be detected through palpation of the neck. A thyroid scan generally provides clear evidence of its presence. Metabolism is altered through autonomous release of thyroxine from the nodules.

Toxic Adenoma also affects older patients and may be detected by thyroid scan. It is distinct from a multinodular goiter in that there is one hyperactive nodule, usually greater than 3cm in diameter, that is active enough to suppress the function of the rest of the gland.¹

1.1.3 Epidemiology of Graves' Hyperthyroidism

A recent large scale epidemiological report out of Johns Hopkins University analyzing the prevalence of 24 autoimmune diseases in the US, placed hyperthyroidism on top of the list with an estimated prevalence of 1151.5 per 100,000⁴. The Canadian experience appears to be similar. A 1993 report from the Thyroid Foundation of Canada estimated that 1 in 100 Canadians suffer from hyperthyroidism. The American study further estimated that 87.9% of those affected were female. This is corroborated by the Whickham Study, a large British analysis that estimated a male to female ratio of 0.10.⁵

Estimates of the age specific incidence of the disease vary considerably.² Study estimates of peak incidence range from 20 to 57 years of age depending on the year and country in which the study takes place.²,⁵
1.1.4 Clinical Management of Graves' Hyperthyroidism

Graves' Hyperthyroidism, like most autoimmune diseases, has a general tendency towards spontaneous remission over time. Nevertheless, a number of therapies are offered early on because of the deleterious effects of the condition on multiple organ systems and on the quality of life of the patient. The most common courses of treatment include anti-thyroid drugs, radioactive iodine therapy and surgery.

Anti-thyroid drugs inhibit the synthesis of thyroid hormone, leading to a gradual reduction in serum thyroid hormone concentrations. The medication is prescribed for a period of weeks or months with the expectation that the condition will eventually improve by itself, following which the patient may be slowly weaned off the drug. The most commonly prescribed drugs today are the orally-administered methimazole and propylthiouracil, both of which are extremely effective in controlling the disease.

Radioactive Iodine therapy has become the most widely used treatment for adults with GH in the United States. It most commonly involves the administration to the patient of a drink or a capsule containing radioactive Iodine (\(^{131}\text{I}\)) to the patient. The iodine is quickly metabolized and taken up by thyroid follicular cells which are irradiated and killed by short range (1-2mm) beta particles. As would be expected, a common result of this therapy is hypothyroidism. (This is considered by many to be the desirable clinical endpoint of this therapy). Thus, treated patients are often placed on a lifelong course of thyroid hormone replacement medication.

Subtotal thyroidectomy is the oldest form of therapy for thyrotoxicosis. It is usually defined as a surgical procedure in which the bulk of the thyroid gland is removed, leaving a rim of a few grams of each lobe posteriorly. Hypothyroidism occurs in most cases following surgery, however in at least 5% of patients, recurrent thyrotoxicosis develops. In these cases alternative therapies must be considered.
Opinions on the effectiveness and benefits of surgery vary a great deal. Werner and Ingbar’s The Thyroid, the pre-eminent text of thyroid disease, suggests very limited use of surgery in favour of other therapies:

The procedure is not without risk of complication and is as a result only performed in special circumstances: children, adolescents and pregnant women who are allergic to or noncompliant with antithyroid drugs, patients with large goiters, and patients who prefer ablative therapy but are apprehensive about radioiodine therapy.\(^2\)

On the other hand, Falk’s text on thyroid disease provides a chapter extolling the virtues of this procedure, citing extremely high efficacy and low complication rates.\(^1\)

Ultimately, the best procedure for a patient must be arrived at through informed shared decision making between doctor and patient.
1.2 Thyroid Orbitopathy

1.2.1 Thyroid Orbitopathy and Graves' Hyperthyroidism

The majority (~90%) of patients suffering from GH display sub-clinical evidence of extra-ocular muscle involvement. This is supported by computed tomography (CT) scans of the orbits of these patients, in which the extra-ocular muscles are inflamed and enlarged beyond their normal size. For a significant number of these patients (about 50% by one estimate) the eye muscle involvement is sufficient to cause clinical changes that may include but are not limited to swelling of the lids and conjunctiva, pain, itching, ophthalmoplegia, exophthalmos, strabismus and double vision. The disease is called Thyroid Orbitopathy (TO), and is almost always found associated with GH. (Various other names for this condition include Thyroid Ophthalmopathy, Endocrine Ophthalmopathy and Graves’ Orbitopathy).

Although the precise mechanism of pathogenesis remains unknown, TO is presumed to be an autoimmune condition. Current hypotheses point to cross-reactivity of TSH receptor antibodies with a similar antigen on the extra-ocular muscles.

As would be expected, there is a close temporal relationship between GH and TO. One report estimates that in about 80% of patients, the diseases occur within 2 years of each other. Another measured 66% of patients presenting with TO within 6 months of GH diagnosis. While it is true that the majority (~90%) of TO cases are preceded by GH, TO may present with no associated thyroid disorder. This kind of presentation is termed euthyroid Grave’s disease, and often results in GH developing months or years later. In cases where no GH develops, it is believed that subclinical autoimmune thyroid disease is still taking place.

1.2.2 Classification of Thyroid Orbitopathy

Clinical presentation of TO may be divided into two categories: infiltrative or non-
infiltrative disease. The majority of patients experience non-infiltrative TO, presenting with lid lag, stare and/or eyelid retraction. These symptoms may also simply be manifestations of GH. Generally in these individuals, orbital involvement is not very serious, and can be controlled through topical interventions such as eyedrops or sunglasses.

Patients with infiltrative disease manifest symptoms related to soft tissue features, myopathy, severe proptosis with corneal exposure or the crowded orbital apex syndrome. It is these manifestations that can variously lead to a range of serious symptoms including strabismus, swelling of the lids, conjunctival inflammation, exophthalmos (protrusion of the eyes) and in some cases, loss of vision. As would be expected, infiltrative disease can become extremely debilitating, resulting in impairment to activities of daily living and profound impact on quality of life.

Werner's NOSPECS model introduced in 1969 by the American Thyroid Association represents the oldest general classification system for TO. (Table 1.2) This system divides TO into a series of classes: classes 0 and 1 (non-infiltrative disease) and classes 2 - 6 (infiltrative disease). Furthermore, extent of involvement in each class is divided into absent, minimal, moderate or marked. (0, a, b and c respectively). A number of criticisms of the NOSPECS system have led to numerous modifications in an effort to improve the system. The most common criticism highlights the pointed lack of correlation of NOSPECS grade to severity of disease. Others call into question the summarizing of TO into an index, (resulting in two very different patients possibly being assigned the same classification), and the inability of NOSPECS to indicate disease activity or remission. In 1992, an international convention of thyroid associations agreed that the NOSPECS system could continue to be used for clinical examination, but that more objective measures should be used in clinical trials and in reporting the clinical features of the disease such as status of eyelids, extra-ocular muscles,
Table 1.2 Detailed Classification of Eye Changes of Graves' Disease (NOSPECS)

<table>
<thead>
<tr>
<th>Classes (0-6)</th>
<th>Grades (o,a,b,c)</th>
<th>Ocular signs and symptoms</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>No signs or symptoms</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Only signs, no symptoms (signs limited to upper lid retraction and stare, with or without lid lag and proptosis)</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>Proptosis associated with class 1 only (specify if difference of 3.0 mm or more between eyes; or progression under observation of 3.0 mm or more, grade o included)</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>Absent (20.0 mm or less is normal)</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>Minimal (21.0 – 23.0 mm)</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>Moderate (24.0 – 27.0 mm)</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>Marked (28.0 mm or more)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Soft-tissue involvement (symptoms of excessive lacrimation, sandy sensation, retrobulbar discomfort, and photophobia, but not diplopia; objective signs as follows)</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>Absent</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>Minimal (edema of conjunctivae and lids, conjunctival injection, and fullness of lids, often with orbital fat extrusion palpable laterally beneath lower lids)</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>Moderate (above plus chemosis, lagophthalmos, lid fullness)</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>Marked</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Proptosis associated with class 2 through class 6 only (specify if inequality of 3.0 mm or more between eyes, or if progression of 3.0 mm or more under observation)</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>Absent (20.0 mm or less)</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>Minimal (21.0 – 23.0 mm)</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>Moderate (24.0 – 27.0 mm)</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>Marked (28.0 mm or more)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Extraocular muscle involvement (usually with diplopia)</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>Absent</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>Minimal (limitation of motion evident at extremes of gaze in one or more directions)</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>Moderate (evident restriction of motion without fixation of position)</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>Marked (fixation of position of a globe or globes)</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Corneal involvement (primarily due to lagophthalmos)</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>ABSENT</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>Minimal (stippling of cornea)</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>Moderate (ulceration)</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>Marked (clouding, necrosis, perforation)</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Sight loss (due to optic nerve involvement)</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>Absent</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>Minimal (disc pallor or choking, or visual field defect; vision 20/20 – 20/60)</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>Moderate (disc pallor or choking, or visual field defect; vision 20/70 – 20/200)</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>Marked (vision less than 20/200)</td>
<td></td>
</tr>
</tbody>
</table>
proptosis and optic nerve function. In this report, an objective assessment of disease activity was suggested:

Disease activity at one time may be assessed by assigning 1 point for the presence of each of the following signs and symptoms: spontaneous retrobulbar pain, pain on eye movement, eyelid erythema, conjunctival injection, chemosis, swelling of the caruncle, and eyelid edema or fullness. The sum of these points defines the clinical activity score (range 0-7).

As another alternative to NOSPECS (put forward a few years before the international convention), Mourits et al proposed a clinical activity score (CAS). The CAS has effectively been used to grade severity of disease previously, and is gaining increasing acceptance among clinicians. (Table 1.3). Nevertheless, it appears that no new methods will be accepted without extensive and convincing clinical trials.
Table 1.3: The 10 items of the clinical activity score (CAS)*

<table>
<thead>
<tr>
<th>Item</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pain 1</td>
<td>Painful, oppressive feeling on or behind the globe, during the last 4 weeks</td>
</tr>
<tr>
<td>Pain 2</td>
<td>Pain on attempted up, side or down gaze, during the last 4 weeks</td>
</tr>
<tr>
<td>Redness 3</td>
<td>Redness of the eyelid(s)</td>
</tr>
<tr>
<td>Redness 4</td>
<td>Diffuse redness of the conjunctiva, covering at least one quadrant</td>
</tr>
<tr>
<td>Swelling 5</td>
<td>Swelling of the eyelid(s)</td>
</tr>
<tr>
<td>Swelling 6</td>
<td>Chemosis</td>
</tr>
<tr>
<td>Swelling 7</td>
<td>Swollen Caruncle</td>
</tr>
<tr>
<td>Swelling 8</td>
<td>Increase of proptosis of ≥ 2mm during a period of 1-3 months</td>
</tr>
<tr>
<td>Impaired function 9</td>
<td>Decrease of eye movements in any direction ≥ 5° during a period of 1-3 months</td>
</tr>
<tr>
<td>Impaired function 10</td>
<td>Decrease of visual acuity of ≥ 1 line(s) on the Snellen chart (using a pinhole) during a period of 1-3 months</td>
</tr>
</tbody>
</table>

*For each item present, 1 point is given. The sum of these points is the CAS. e.g. a CAS of 6 means that six items were present, regardless of which items.

1.2.3 Diagnosis of Thyroid Orbitopathy

In the ophthalmologists office, effective diagnostic evaluation involves the use of diagnostic imaging, detailed history-taking, and identification of a number of diagnostic signs and symptoms.

Computer-assisted Tomography (CT) scans are one of the most common imaging methods used by clinicians. (Magnetic Resonance Imaging (MRI) is another common method). Its primary utility lies in its ability to demonstrate degree of enlargement of the extraocular muscles, which can then be compared to published normals. In most institutions, orbital CT scans are routinely performed using 1 to 2mm thick sections at 2-mm intervals in the axial plane.

On imaging, the extraocular muscles appear to be the primary area of orbital involvement in thyroid eye disease. Some CT studies of thyroid ophthalmopathy patients have not noted orbital fat abnormalities, while others suggest that fat and muscle volume are increased... CT is also useful in thyroid eye disease patients to delineate the etiology of decreased vision... patients with thyroid optic neuropathy display enlarged extraocular muscles that cause a demonstrable compression of the nerve at the orbital apex.
The symptoms of TO are greatly varied, and common to other related diseases. However, as Char explains, particular combinations of symptoms may be used to indicate presence of TO:

The combination of bilateral exophthalmos, lid retraction, stare, and an enlarged thyroid are virtually pathognomic for endocrine exophthalmos. Some ocular signs are relatively specific for thyroid ophthalmopathy. These include proptosis and lid lag or stare, proptosis plus a restrictive extraocular myopathy, or the presence of isolated enlarged vessels over the insertion of the medial or lateral rectus muscles. Conjunctival edema or periorbital edema is also quite common in thyroid exophthalmos and is occasionally observed in other conditions. Almost all endocrine exophthalmos patients, even when asymptomatic, have some degree of extraocular muscle involvement demonstrable by abnormal ultrasonography or intraocular pressure (IOP) tests. The most common extraocular muscles involved in thyroid eye disease, in order of frequency, are the inferior, medial, superior and lateral recti. The oblique muscles are involved less frequently.

Since a large number (greater than 75%) of TO patients will have a history of systemic thyroid disease along with the symptoms described above, diagnosis is often straightforward. However, in cases of euthyroid Graves' disease, clinicians must rely on other clinical signs (such as those described above) and CT scans. This kind of differential diagnosis can be particularly challenging since normal lid and orbital anatomy is variable. Myopia, race, and familial attributes, for instance, can alter the degree of ocular protrusion and overall lid position. As Rootman reports:

Normal exophthalmometry measurements will vary according to age, sex, and race. The mean normal protrusion values are 16.5 mm in white men, 18.5 mm in black men, 15.4 mm in white women, and 17.8 mm in black women.

1.2.4 Treatment of Thyroid Orbitopathy

A large amount of literature has been published on the subject of TO treatment, but controversy remains today as to which is the most appropriate treatment for patients at different stages of disease. It was demonstrated in 1995 by Perros et al that in a significant proportion of TO patients, (64.4% in their series of 59 patients) the disease spontaneously improves with no intervention. It was
their conclusion, therefore, that all efficacy tests of new TO treatments should be carefully controlled “to allow for the natural tendency towards remission.” It should also be noted that the nature of “improvement” in a TO patient must be carefully defined before embarking on a clinical trial. (Patients may get better, but rarely return to the physical and emotional state that they enjoyed before getting sick).12

The majority of TO patients are referred to an ophthalmologist through general practitioners or endocrinologists. The referrals are most often made after significant changes in the eye or vision of the patient have been brought to the attention of the clinician.28 Thus, many referrals take place at a reasonably advanced point in the natural history of the disease, when major rather than minor interventions are required.

The most common therapies used to control severe TO include surgery, steroids and radiotherapy.

Surgical intervention is indicated in some cases of severe infiltrative disease for the correction of the crowded orbital apex syndrome through orbital decompression (in which the apical portion of the orbital walls are removed to release pressure on the optic nerve).11 Various other surgical procedures exist for the oculomotor abnormalities, lid malposition and cosmetic deformity of the condition, but these are only used once the condition has been quiescent for at least 6 months.11

Steroids are the most commonly prescribed treatment for patients with recently developed disease.30 The mechanism of action of these drugs remains uncertain. As Char indicates, their anti-inflammatory and immuno-modulatory actions are probably most important.12 Steroids have been shown to be very effective in treating TO early in the course of disease, but carry with them a host of potential side effects, making their use less palatable. As a result, numerous studies have looked at
dosage and duration of steroid therapy, as well as treatment regimes including both radiotherapy and steroids.\textsuperscript{11,12,17,31,32} Table 1.4 lists the potential complications of corticosteroid therapy.

Supervoltage orbital radiotherapy has a significant beneficial effect in patients with significant acute, active disease and severe or subacute inflammatory soft-tissue signs, with little or no side effects, as compared to the greater risk involved in surgery and in steroid use.\textsuperscript{12,17,31} Although, like steroids, the precise mechanism of action of radiation is not understood, it is believed that radiation causes local immuno-suppression in the eye muscles it targets, thereby decreasing inflammation and rapidly relieving soft tissue signs. The standard measurement unit of radiation dose is the gray (Gy). One Gray equals 100 radiation absorbed doses (rads). Typically, radiotherapy is administered in a number of fractionated daily doses in order to minimize morbidity in other structures. Eligible patients referred for treatment to the British Columbia Cancer Agency currently receive a dose of 20 Gy given in 10 fractions over a 2 week period.

<table>
<thead>
<tr>
<th>Table 1.4: Complications of Corticosteroid Therapy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Pituitary-adrenal suppression</td>
</tr>
<tr>
<td>2. Infection</td>
</tr>
<tr>
<td>3. Cataract</td>
</tr>
<tr>
<td>4. Systemic hypertension</td>
</tr>
<tr>
<td>5. Osteoporosis</td>
</tr>
<tr>
<td>6. Renal lithiasis</td>
</tr>
<tr>
<td>7. Diabetes</td>
</tr>
<tr>
<td>8. Ecchymoses</td>
</tr>
<tr>
<td>9. Phlebitis</td>
</tr>
<tr>
<td>10. Hirsutism</td>
</tr>
<tr>
<td>11. Reactivation of chronic disease (tuberculosis etc.)</td>
</tr>
<tr>
<td>12. Psychosis</td>
</tr>
</tbody>
</table>

\textsuperscript{note: adapted from Char\textsuperscript{12}}

\textbf{1.2.5 Epidemiology of Thyroid Orbitopathy}

Very few Canadian studies on the epidemiology of TO or GH have been conducted in the past 10 years. Therefore, most epidemiological information is supplied by European and American
analyses. It should be noted that the range of estimates quoted in this section are as much attributable to different investigators defining presence of TO in different ways as they are to differences between clinical populations. (For example, if the definition of Orbitopathy is based on extraocular muscle thickness, the incidence may be up to 90%, whereas other clinical measures may place it as low as 10%).

The vast majority (~90%) of TO cases are associated with GH, the diseases most often occurring with 2 years of each other. Kendall-Taylor and Perros report that TO is clinically apparent in about 50% of hyperthyroid patients, with subclinical orbital involvement in up to 90% of patients. Streeten et al, in an effort to quantify incidence of TO, defined presence of disease as an exophthalmic measurement of >19mm. (This cutoff was established after measurement of 105 individuals with no known thyroid disease to establish a normal range of exophthalmos values). By that definition, 21.3% of their cohort of thyrotoxic patients had disease. This study may have underestimated disease however, since no discrimination between the different normal ranges of male and female patients’ proptosis were measured (as described in section 1.2.3).

Another study by Sridama and DeGroot in 1989 determined that TO affects 44% of patients with GH. In this investigation, presence of TO was defined as one of:
1) Exophthalmic measurement of 20 mm or more,
2) Exophthalmic measurement of 18-20 mm with obvious signs of infiltrative orbitopathy (e.g. inflamed extra-ocular muscles, diplopia or chemosis
3) Muscle enlargement detected on Computerized axial tomographic scan

Most TO patients experience mild to moderate disease. Once again, estimates of severe disease range widely, with the most commonly sited figures being 3-5%. (In contrast, Sridama and Degroot found that 14% of their clinical population was severely affected, where severe
ophthalmopathy was defined as exophthalmos greater than 23 mm, extraocular muscle involvement, moderate to severe chemosis or a combination of all three.

TO, like Graves' disease in general, occurs most commonly in females 30-50 years old.\textsuperscript{12} The male-to-female ratio in TO has been reported as approximately 2.5:1. This represents a less slanted ratio than that reported for GH (F:M 4:1).\textsuperscript{12, 37-39} In both TO and GH, the mean age of onset is about 43-45 years, and the disease is generally more severe in men and patients older than 50 years.\textsuperscript{15} There is also evidence to suggest that smoking increases the risk of TO, as does particular ethnic background (i.e. Europeans are at higher risk than Asians).\textsuperscript{38} A study of severity of TO in patients with Graves' Disease for the period 1960-1990 indicated that the severity of TO declined significantly from 57.1% severe cases in 1960 to 35% in 1990. This has been attributed to improved diagnostic screening and more effective treatments having been introduced.\textsuperscript{39, 40}
1.3 Diagnostic Testing

1.3.1 Introduction

The field of diagnostic testing in medical care is a broad one. It includes all aspects of development, conduct, appropriateness and effectiveness of diagnostic tests and screening programs. It delves into the details of test construction and psychometric properties, spilling over, and borrowing heavily from the fields of Statistics and Psychology. It also dips into policy, program evaluation and the economics of health care. For the purposes of this discussion, the whole issue of diagnostic testing will be limited to the most direct issues involved in development and properties of these tests. The issues of implementation, economics and policy will be left students of those fields.

In order to provide a framework for the understanding of diagnostic tests, it is necessary to define early on the relationship between diagnostic tests and screening for disease.

Screening, according to Friis and Sellers, can be defined as the presumptive identification of unrecognized disease or defects by the application of tests, examinations, or other procedures that can be applied rapidly. Sackett defines the objective of screening (or early diagnosis, as he calls it) as “the early detection of presymptomatic disease”. The natural extension of this objective is to control disease through early intervention and treatment. However as the world changes, the use of these tests has extended beyond simply the benefit of the patient, to such things as life insurance, immigration control (denying entry to sick individuals) and potential employee screening. Screening may also be used to determine baseline attributes of particular tests in order to characterize healthy patients. For the purposes of this discussion, it will be assumed that the motivation behind screening is to improve the health of the patient.

Diagnosis is the process of confirming an actual case of disease. Once diagnosis is made, interventions may be attempted to reduce or eliminate the affect of the condition on a person’s health.
A diagnostic test must necessarily still be performed following a positive screening test. However, diagnostic tests are often used without the administration of a screening test. Thus, a screening test may be considered a subtype of diagnostic test, which is simply administered earlier on in the natural history of disease, and must therefore be confirmed at a later date by a second diagnostic test and clinical evaluation. This is supported by Sackett's treatment of the issue of screening as simply one of "early diagnosis." The discussion that follows, then covers the diagnostic test in general, including a section on special considerations of screening tests.

1.3.2 Properties of diagnostic tests

The perfect diagnostic test would tell an investigator with 100% certainty whether or not an individual has a particular condition. Its very purpose is to diagnose, thereby allowing a clinician to proceed to the next step of attempting to control the disease. However, in reality, no diagnostic test offers perfect and indisputably correct results. Virtually every test will produce at least some false-positive and false-negative results. Thus test developers, while keeping the "perfect test" in mind, must strive for the less lofty goal of instruments that provide clinically useful information while minimizing error as much as possible.
1.3.2.1 Terminology and Two-By-Two Tables

Having accepted that error must accompany all diagnosis, we are in a position to discuss the results of a diagnostic test summarized by a 2 X 2 table:

Table 1.5: Diagnostic Test 2 X 2 table

<table>
<thead>
<tr>
<th>Disease: (^1)</th>
<th>Positive: ((D^+))</th>
<th>Negative: ((D^-))</th>
<th>Totals:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test: (T^+)</td>
<td>(a)</td>
<td>(b)</td>
<td>(a+b)</td>
</tr>
<tr>
<td>(T^-)</td>
<td>(c)</td>
<td>(d)</td>
<td>(c+d)</td>
</tr>
<tr>
<td>(\text{Totals}):</td>
<td>(a+c)</td>
<td>(b+d)</td>
<td>(a+b+c+d)</td>
</tr>
</tbody>
</table>

\(^1\)as measured by Gold Standard

Table 1.6: Terminology

<table>
<thead>
<tr>
<th>Epidemiological Term</th>
<th>Probability</th>
<th>2 X 2 Algebra</th>
<th>Alternate Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity:</td>
<td>(P(T^+</td>
<td>D^+))</td>
<td>(\frac{a}{a+c})</td>
</tr>
<tr>
<td>Specificity:</td>
<td>(P(T^-</td>
<td>D^-))</td>
<td>(\frac{d}{b+d})</td>
</tr>
<tr>
<td>False Positive:</td>
<td>(P(T^+</td>
<td>D^-))</td>
<td>(\frac{b}{b+d})</td>
</tr>
<tr>
<td>False Negative:</td>
<td>(P(T^-</td>
<td>D^+))</td>
<td>(\frac{c}{a+c})</td>
</tr>
<tr>
<td>Positive Predictive Value (PPV):</td>
<td>(P(D^+</td>
<td>T^+))</td>
<td>(\frac{a}{a+b})</td>
</tr>
<tr>
<td>Negative Predictive Value (NPV):</td>
<td>(P(D^-</td>
<td>T^-))</td>
<td>(\frac{d}{c+d})</td>
</tr>
<tr>
<td>Post-test Likelihood (PTL(-))</td>
<td>(P(D^+</td>
<td>T^-))</td>
<td>(\frac{c}{c+d})</td>
</tr>
<tr>
<td>Pre-test Likelihood</td>
<td>(P(D^+))</td>
<td>(\frac{a+c}{N})</td>
<td>Prevalence</td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td>(\frac{a+d}{N})</td>
<td>Comparison of new test to gold standard</td>
</tr>
<tr>
<td>Likelihood Ratio (LR+)</td>
<td>sensitivity (1-specificity)</td>
<td>True positive rate/False positive rate</td>
<td></td>
</tr>
<tr>
<td>Likelihood Ratio (LR-)</td>
<td>(1-sensitivity) specificity</td>
<td>False negative rate/True negative rate</td>
<td></td>
</tr>
</tbody>
</table>
Sensitivity is a common measure of the utility of a diagnostic test. It produces the probability that an individual will produce a positive test, given that they have the disease. Sensitivity is also called the True Positive rate. Thus, if a test is able to correctly classify all individuals who truly have disease as positive, it will have a sensitivity of 100%. The false negative rate (i.e. the probability of scoring a negative test result when disease is actually present) is given by 1 – sensitivity.

Specificity is usually reported along with sensitivity. It provides the probability that an individual will produce a negative test, given that they do not have the disease. Specificity is also referred to as the True Negative rate. Thus, if a test is able to correctly classify all individuals who do not have disease as negative, it will have a specificity of 100%. The false positive rate (i.e. the probability of scoring a positive test result when disease is actually absent) is given by 1 – specificity.

The Positive Predictive Value (PPV), (also known as the Post-Test Likelihood (PTL(+)), measures the probability that a person has disease, given that they received a positive test. The Negative Predictive Value (NPV) measures the probability that a person does not have disease, given that they received a negative test. The Post-Test Likelihood (PTL-) represents the complement of NPV, providing the probability that a person actually has disease, given that they received a negative test. The predictive value measures and their complements represent the clinically important quantities of a diagnostic test, and are used in clinical decision making (section 1.3.3.3). As Sackett points out:

> When we use diagnostic tests clinically, we do not know who actually has or does have the target disease; if we did, we would not need the diagnostic test! Our clinical concern is not a vertical one of sensitivity and specificity, but a horizontal one of the meaning of positive and negative test results.

The pre-test likelihood refers to the probability that an individual in the study population is positive for disease. In general terms, this would be identified as the prevalence of the disease in the sample population. This quantity, besides being interesting and useful in its own right, has a significant
impact on the predictive value of a test. Specifically, when the prevalence of a disease within a population falls, the PPV falls and the NPV increases. This is not particularly surprising when one examines the 2 X 2 table. A low prevalence will result in the quantities $b$ and $d$ being far greater than the quantities $a$ and $c$. Thus, when one calculates the PPV in a low prevalence situation, the resulting probability will necessarily be small since $a$ will be a good deal less than $b$. (unless the test has 100% specificity). Put another way, if a test produces any false positives at all, it’s ability to discriminate those with disease from those without disease becomes increasingly limited as the prevalence of the disease decreases. Conversely, its NPV will increase greatly since hardly anyone has disease anyway. (More as a consequence of prevalence then due to the quality of the test).

The accuracy of a test compares its performance to that of the gold standard, and as such is a measure of the validity of the test. This attribute is discussed further in section 1.4.

Finally, the likelihood ratios provide an alternate method of summarizing the utility of a diagnostic test. They are discussed in more detail in section 1.3.3.

### 1.3.2.2 ROC curves

Sensitivity and specificity are considered to be stable properties of diagnostic tests and, as a result, are unaffected by the prevalence of a disease. They do, however, have a direct effect on each other. As the sensitivity of instrument is increased, the false negative rate decreases and the false positive rate tends to increase. To illustrate this, consider the example of fasting blood glucose levels as a diagnostic test for diabetes. Fasting blood glucose levels may approximate a normal distribution, with a mean of 100mg/dL. A particular cut-off level, therefore, must be established in order to label individuals as having disease or not having disease. If that level is set very low (say at 110mg/dL), virtually all patients with disease will most likely be captured, producing a very high sensitivity for the test. However, at this level, a large number of patients will be labeled positive for disease when in fact
they have none. Not only does this produce unnecessary anxiety for the patients and cost to the system; it lowers the specificity of the test. Conversely, if the level is set very high (say at 140mg/dL), virtually no individuals will be diagnosed with disease when in they don’t have it, producing a very high specificity. The trade-off, once again, will be that a great number of positive patients will be missed.

The optimal cut-off point for a particular diagnostic test changes depending on the objectives and concerns of the investigator. A useful method for determining this point is on a Receiver-Operator Characteristic (ROC) curve (fig 1.1). On this curve, the true positive rate (sensitivity) is plotted against the false positive rate (1-specificity) for different cut-offs of a particular test. The resulting diagram allows the investigator to identify the most helpful cut-off value for the test.

**Figure 1.1: ROC Curve**
1.3.3 Diagnostic Decision Analysis

Clinicians are frequently faced with the challenge of making clinical decisions based on diagnostic tests. They must therefore develop the ability to translate the results of a test into a meaningful clinical context. This is process is referred to as diagnostic decision analysis. Schechter and Sheps\textsuperscript{43} summarize the principles of diagnostic decision analysis as follows:

| Principle 1: | In the diagnostic context, patients do not have disease, only probability of disease |
| Principle 2: | Diagnostic tests are merely revisions of probabilities |
| Principle 3: | Test interpretation should precede test ordering |
| Principle 4: | In general, if the revisions in probabilities caused by a diagnostic test do not entail a change in subsequent management, use of the test should be reconsidered. |

The first principle refers to the idea that a clinician who is preparing to order a particular test does so because disease is suspected in the patient. If the clinician is certain that disease is present or absent, no test would be necessary. Thus, a patient presents to his/her clinician with a probability of disease.

Given principle 1, it follows that the results of a diagnostic test will revise the patients initial probability of disease to a new probability. If the test is a useful one, the revision will aid in clinical decision making. This is termed a diagnostic tree, a generic one of which is summarized in fig 1.2a.\textsuperscript{43} In the general case, a patient enters a clinic with a particular pre-test likelihood for disease (P(D+)). After being tested, this probability is revised depending on the results of the test. In the case of a positive test, the patient’s revised probability is equal to PTL(+). In the case of a negative test, the patient’s revised probability is equal to PTL(-). Fig 1.2b provides a diagnostic tree for a particular test with sensitivity and specificity of 0.65 and 0.90 respectively in a patient with a pretest probability of disease of 0.20.
Fig 1.2a: Diagnostic tree for a diagnostic test. The patient enters the test at the left, with a pretest probability of the target disease. If the test result is positive, the probability rises to $\text{PTL}(+)\text{,}$ the post-test probability of a positive test. If the result is negative, the probability falls to $\text{PTL}(-)$, the post-test probability of a negative result.

Fig 1.2b: Diagnostic tree for a test with a sensitivity of 0.65 and a specificity of 0.90 in a patient with a pretest probability of the target disease of 0.20.

The likelihood ratio (LR) provides an alternate manner of estimating a diagnostic test's clinical utility:

$\text{LR}(+)\ldots$ is a quantity greater than or equal to 1.0, and the magnitude by which it exceeds 1.0 is a measure of the test's ability to revise probabilities upward when the test result is positive. An $\text{LR}(+)\text{ of 2.0 to 5.0 should be considered as poor to fair, while one exceeding 10.0 might be considered good. Conversely, LR}(-)$ is a quantity less than or equal to 1.0 and the magnitude by which it falls below 1.0 is a measure of the test’s ability to revise probabilities downward when the test result is negative. An $\text{LR}(-)\text{ of 0.5 - 0.2 should be considered poor to fair, while one below 0.1 might be considered good.}$
The third principle of diagnostic test decision analysis concerns itself with test interpretation. Specifically, the interpretation of the test result should be clear before the test is ever ordered. Since the decision trees in Fig. 1.2 may be constructed before a test is ordered, the clinical meaning of a positive or negative result in terms of probability of disease may be determined ahead of time. If it is discovered that a test result will provide no helpful information, than no test should be ordered. For example, consider the test of radionuclide angiography (RNA) for the detection of coronary artery disease (CAD), with a 0.87 sensitivity and a 0.54 specificity. Fig 1.3. (taken from Schechter & Sheps43) provides three different decision trees based on patients with different pre-test probabilities. As can be seen from the figure, the revision of probabilities in the first two cases are of minor use, and probably would have no influence in the continuing care of the patient. However, in the case of the third patient, a negative test would reduce the probability of disease significantly, while a positive test would provide some support for the presence of disease. Thus, an understanding of the results of the test and its implications for particular patients is of major importance.

Finally, principle 4 simply reiterates the fact that if a test provides no useful revision to the probability of disease, the test should not be used in clinical practice.
1.3.4 Evaluation of diagnostic tests

It is clear that the type and quality of diagnostic tests available to clinicians vary a great deal. It follows naturally, then, that a significant number of tests being considered for use, or already in use, may fall well below what would be considered the acceptable level to the average clinician. A framework for the evaluation of tests has been advanced by Sackett et al and is summarized in table 1.8 and the following discussion.

Table 1.8: Eight guides for deciding the clinical usefulness of a diagnostic test

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Has there been an independent, “blind” comparison with a “gold standard” of diagnosis?</td>
</tr>
<tr>
<td>2.</td>
<td>Has the diagnostic test been evaluated in a patient sample that included an appropriate spectrum of mild and severe, treated and untreated disease, plus individuals with different but commonly confused disorders?</td>
</tr>
<tr>
<td>3.</td>
<td>Was the setting for this evaluation, as well as the filter through which study patients passed, adequately described?</td>
</tr>
<tr>
<td>4.</td>
<td>Have the reproducibility of the test results (precision) and its interpretation (observer variation) been determined?</td>
</tr>
<tr>
<td>5.</td>
<td>Has the term <em>normal</em> been defined sensibly as it applies to this test?</td>
</tr>
<tr>
<td>6.</td>
<td>If the test is advocated as part of a cluster or sequence of tests, has its individual contribution to the overall validity of the cluster or sequence been determined?</td>
</tr>
<tr>
<td>7.</td>
<td>Have the tactics for carrying out the test been described in sufficient detail to permit their exact replication?</td>
</tr>
<tr>
<td>8.</td>
<td>Has the utility of the test been determined?</td>
</tr>
</tbody>
</table>

1. Has there been an independent, “blind” comparison with a “gold standard” of diagnosis?

A common test of the validity of a diagnostic test involves a *blinded* comparison with a *gold standard* of diagnosis. This is discussed further in section 1.4. A *blinded* study indicates that the investigators interpreting the results of the test do not know whether the given patient is positive for disease. Usually, the determination of “truth” concerning presence or absence of disease is decided by the most acceptable available reference test, dubbed the *gold standard*. In making this comparison, other issues need to be considered. Firstly, the gold standard test should be trustworthy and valid or else comparison may be meaningless. In fact, if the gold standard is not acceptable, the question of
whether the diagnostic data are worth capturing at all should be considered. Secondly, given that the
gold standard is acceptable, how useful is the new diagnostic test, and what advantages does it have
over use of the gold standard? (For example, is it less risky, less uncomfortable or embarrassing for the
patient, less costly, or applicable earlier in the course of illness.42)

2. Has the diagnostic test been evaluated in a patient sample that included an appropriate
spectrum of mild and severe, treated and untreated disease, plus individuals with different
but commonly confused disorders?

A diagnostic test should successfully be able to determine presence or absence of disease.
Furthermore, a good test should also be able to distinguish between similar conditions and the target
condition particularly in situations where prognoses or therapies differ sharply.

3. Was the setting for this evaluation, as well as the filter through which study patients passed,
adequately described?

Considering the great influence of prevalence of disease on predictive value and post-test
likelihood, and the ever-present need to control bias, it is important for investigators to outline carefully
the setting and selection criteria in a test’s evaluation. This must include a full characterization of both
the cases and controls in a test population.

4. Have the reproducibility of the test results (precision) and its interpretation (observer
variation) been determined?

The precision of a test refers to a kind of reliability (discussed further in section 1.4.2.1)
Simply put, the same test applied to the same, unchanged patient must produce the same result. The
developers of a test should report how well the test fulfills this attribute. Test developers should also
be concerned with making interpretation of test results as simple as possible so as to limit observer
variation, particularly in the case of complicated procedures.
5. Has the term *normal* been defined sensibly as it applies to this test?

Any use of the term *normal* must be defined by the authors of a test. Furthermore, the evaluator of the test should be content that the definition used is appropriate for the particular question addressed by the study. There are a number of definitions of *normal* in different situations. Normal may refer to the normal distribution, implying that test results follow the bell-shaped curve of that probability distribution, or it may refer to some sort of a reference population who is considered free of disease. For a more detailed discussion, see Sackett\(^2\)

6. If the test is advocated as part of a cluster or sequence of tests, has its individual contribution to the overall validity of the cluster or sequence been determined?

The measurement of the validity of a test should be conducted in a manner that reflects actual clinical use of the test. If the test is to be used along with a number of other tests to identify a condition, its utility should be measured based on the patients it will be used on in clinical practice.

7. Have the tactics for carrying out the test been described in sufficient detail to permit their exact replication?

This point includes a detailed description of the target population, concerns, cautions and caveats of the test. It should also include a description of the subjective affect of the test on patients and how this may in turn affect results.

8. Has the utility of the test been determined?

Sackett et al state the following in their text:

> The ultimate criterion for a diagnostic test or any other clinical maneuver is whether the patient is better off for it... In addition to telling you what happened to patients correctly classified by the diagnostic test, its advocates should describe the fate of the false positives and the false negatives. Moreover, when the execution of a test requires a delay in the initiation of definitive therapy (while the procedure is being rescheduled, the test is incubating, or the slides are waiting to be read) the consequences of this delay should be described.\(^2\)
1.3.5 Subjective Diagnostic Tests

Once again, a specific definition of the quantities under discussion becomes necessary: *objective* parameters are considered to be those quantities free of the effects of opinion and its associated bias. (such as measurement of a level of a chemical in the blood, height or visual acuity). Subjective parameters, on the other hand, are understood to be partly or wholly influenced by the opinion of the rater (such as quality of life questionnaires, symptom reports, self-administered instruments).

The explicit delineation of these definitions immediately provokes a discussion of how truly *objective* measurements are. Certainly, an argument can be made that virtually every objective measurement has a subjective influence somewhere along the line. This argument may be supported by simple thought experiments (e.g. the apparently objective measurement of visual acuity may be altered depending on how the tester carries out the test: “Can you read the first letter?” vs. “Can you see the E?”). Often the source of the problem in these cases lies with lack of conformity to particular diagnostic criteria among clinicians, lack of understanding concerning what marks presence or absence of disease, or simple difference in measurement techniques and associated error leading to different diagnoses. Nevertheless, objective standards for medicine and science have long been regarded as the best and “cleanest” method of measurement.

In recent times, the beginnings of a paradigm shift in the delivery of health care has made itself apparent in the literature and in the training of medical students. These changes reflect among other things, an increasing trend toward the inclusion and use of subjective data in clinical management of disease, and the growing demand for evidence-based data to support efficacy of particular interventions and natural history of disease. In part, the catalyst for this change has been the growing technological abilities of society: affordable computer systems, increased database use, and convenient statistical
software that turns once arduous calculations into simple point and click solutions. Thus, through integration of knowledge from the fields of Psychology, Statistics, Sociology and Epidemiology, techniques for the quantification of subjective data into meaningful and useful outcomes is becoming increasingly accepted. In the current paradigm, objective and subjective measures together present the complete picture of a patient's health status.

1.3.5.1 Hard and Soft Subjective Rating

The majority of patient-completed diagnostic tests fall into one of two categories. They can refer to particular behaviours, feelings and emotions, (the soft category) or they can deal directly with particular tasks or symptoms (the hard category). What separates these two concepts is the degree of inherent subjectivity of the target construct (the term construct is defined in section 1.4).

The soft group of subjective tests would include most psychological tests (such as Beck's Depression Inventory, the Psychiatric Diagnostic Screening Questionnaire (PDSQ) and anxiety tests). In these tests, the condition of interest is inextricably tied in with the emotions of the patient, and the goal of the test is to measure these emotional and mental effects.

The hard group of tests are more concerned with conditions that demonstrate symptoms or limitations beyond the mental and emotional realm, and that can be looked at and asked about by a patient. (e.g. blurriness of vision, itching and pain). They may also include how well a patient can perform a particular task (e.g. climbing stairs). In these cases, responders may indicate whether or not they can do something, rather than how they feel about something, thereby making the test "harder." Examples of this include the Foot Disability Questionnaire, The Prenatal Alcohol Use Interview, the HIV-risk screening instrument and functional assessment tests (see section 1.3.5.2) such as the VF-14. In addition to these tests, self-administered symptom indices are enjoying increased acceptance as
hard subjective tests. These tests simply contain a checklist of symptoms to which the patient can respond.\textsuperscript{49,50}

There is, necessarily, a great deal of overlap between these groups, as well as a host of tests that are difficult to pigeon hole appropriately, since they combine elements of both. Examples of this are quality of life tests (see section 1.3.5.2) like the SF-36\textsuperscript{51}, National Eye Institute’s Visual Function Questionnaire\textsuperscript{52} and the University of Florence’s Binge Eating Scale.\textsuperscript{53}

**1.3.6 Screening Programs**

An implicit assumption underlying the concept of screening is that early detection, before the development of symptoms, will lead to a more favorable prognosis because treatment begun before the disease becomes clinically manifest will be more effective than later treatment.\textsuperscript{54} Thus, screening also implicitly assumes an orderly natural history of disease summarized by Figure 2 and described below (Adapted from Sackett\textsuperscript{42}). This order may be broken down into four stages:

![Figure 1.4: Natural History of Disease](image)

---

30
Biologic Onset: The disease initiates through natural interaction between human and environment. However, at this stage the presence of disease cannot be detected. The temporal location of this stage in relation to the remaining stages varies greatly depending on the disease.

Early Diagnosis Possible: At this point, although the individual is largely or completely free of symptoms, the mechanisms of disease produce sufficient change that early diagnosis is possible (given the correct test.)

Usual Clinical Diagnosis: The stage at which disease progression produces symptoms and the affected individual seeks clinical help.

Outcome: The final resolution of the disease resulting in recovery, permanent disability or death.

Coupled with these stages is the concept of the “critical point” as described by Hutchison. The critical point in the natural history of a disease is the point before which therapy is either more effective or easier to apply than afterward. Thus a disease may have several critical points (as in pulmonary tuberculosis) or none (as in some cancers). It follows, then, that the location and number of critical points in a disease is highly important in determining the value and use of screening tests. If no particular benefit will be provided to the patient by early diagnosis, there is not much use in performing the test. (Unless the ultimate goal was not in fact to benefit the patient, as explained above). However, if a critical point lies between stages 2 and 3, and therapies exist for early intervention, screening becomes a useful tool to aid in improving outcomes. This and related concepts is covered further in section 1.3.2.

Screening programs may be targeted at the general population (mass/population screening) or at particular high-risk individuals (selective screening). They may also be restricted to the detection of a single disease or for a number of diseases (multiphasic screening). Further, they may be used in clinical settings to help identify patients who need to be referred quickly to the next level of intervention, while filtering out patients who have clinically similar but unrelated conditions (as in the
case of the Ottawa Ankle Rule\textsuperscript{46} and the helical CT screening rule\textsuperscript{47}—This is truly a type of selective screening).

A number of authors\textsuperscript{41,42,54,58} provide discussions on the use and evaluation of screening programs. In this context, a screening program refers to the entire process of identifying and capturing a target population and applying a screening test to them. The major points made by these authors are summarized in Table 1.9:

Table 1.9: Seven guides in the evaluation of a screening program

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Does the current burden of suffering warrant screening?</td>
</tr>
<tr>
<td>2.</td>
<td>Are efficacious treatments available?</td>
</tr>
<tr>
<td>3.</td>
<td>Will those who had a positive screening comply with subsequent advice and interventions?</td>
</tr>
<tr>
<td>4.</td>
<td>Is there a good screening test?</td>
</tr>
<tr>
<td>5.</td>
<td>Can the health system cope with the screening program? (money and time)</td>
</tr>
<tr>
<td>6.</td>
<td>Does the program reach those who could benefit from it?</td>
</tr>
<tr>
<td>7.</td>
<td>Has the program’s effectiveness been demonstrated in a randomized trial?</td>
</tr>
</tbody>
</table>

1. **Does the current burden of suffering warrant screening?**

   Does the untreated disorder substantially impact the present (or future) physical, social, emotional, and/or intellectual function of the person (and/or his family)? If the disorder is of trivial consequence to affected persons and their families, a community-wide screening program is not indicated.\textsuperscript{58}

2. **Are efficacious treatments available?**

   This question gives rise to two issues: Firstly, as described above, will treatment of preclinical disease be more effective/useful than treatment begun after the development of symptoms? Secondly, if disease is diagnosed, are there available treatments that have had their efficacy demonstrated in a rigorously controlled study (ideally, a randomized control trial\textsuperscript{54})?

   The first issue may be addressed with a full understanding of the natural history of the disease followed by a careful reevaluation of the objectives of the utility of the screening test. Obviously, if early diagnosis is does not help the patient, its use should be reconsidered.
The second issue addresses the fundamental reasons for screening. If treatment does not lead to improved clinical outcomes, or is unavailable for patients who are found to be positive for disease, what is the purpose of identifying them? As has been explained earlier, there may well be other perfectly acceptable reasons. However, these should be absolutely clear from the outset.

3. **Will those who had a positive screening comply with subsequent advice and interventions?**

   If patients will not take their medicine, all the foregoing screening and diagnosis, however elegantly they were conceived and executed, are nullified.\(^42\)

Before even considering a test, the compliance of the target population should be considered. If the population *must* be treated no matter how they feel about it, it is important for clinicians to work closely with their patient groups in order to promote best possible compliance.

4. **Is there a good screening test?**

   A screening program can be effective only if it starts with a good test that is both accurate and feasible to use in screening situations.\(^58\) This includes all the determinants discussed in earlier sections (sensitivity, specificity, PPV etc.) as well as the following table abstracted from Friis and Sellers:\(^41\)

### Table 1.10: Characteristics of a Good Screening Test

<table>
<thead>
<tr>
<th>Simple</th>
<th>The test should be easy to learn and perform</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rapid</td>
<td>The test should not take long to administer, and the results should be available soon</td>
</tr>
<tr>
<td>Inexpensive</td>
<td>The lower the cost of a screening test, the more likely it is that the overall program will be cost beneficial.</td>
</tr>
<tr>
<td>Safe</td>
<td>The screening test should not carry potential harm to screenees</td>
</tr>
<tr>
<td>Acceptable</td>
<td>The test should be acceptable to the target group</td>
</tr>
</tbody>
</table>

5. **Can the health system cope with the screening program?**

   Clinical work begins with the administration of a screening test. After administration, follow-up is necessary in the form of definitive diagnostic evaluations of positive individuals, appropriate therapy prescribed, monitoring of compliance, and future follow-up appointments. If the health system is not able to absorb and appropriately handle this increased demand (both of time and money), all previous
screening efforts are wasted and all that remains are citizens who have been told there is something wrong with them.\textsuperscript{58}

6. **Does the program reach those who could benefit from it?**

Screening programs aimed at the general population (rather than specific subgroups in the community) are particularly susceptible to the "inverse care law"—those in greatest need tend to be those least likely to be screened.\textsuperscript{58}

It is necessary for investigators to work with policy makers and each other to ensure that those who need testing are, in fact, receiving it.

7. **Has the program's effectiveness been demonstrated in a randomized trial?**

Finally, just like a diagnostic test, a screening program should be evaluated through a controlled experiment to determine if it does more harm than good. For example, those in the case group would be encouraged to receive definitive diagnosis of their conditions should they test positive, while those in the control group would not undergo screening, and would receive diagnosis through routine check-up.

An experiment of this kind should include a population that is representative of the target population, should be randomized, and should be able to detect a *clinically* and *socially* significant improvement in outcomes (versus simply a *statistically* significant improvement.) Persons not wishing to participate after randomization, dropping out, or failing to comply with therapeutic intervention should be included in the analysis. (i.e. an Intent-To-Treat analysis\textsuperscript{59}) This represents the best type of an analysis for an *effectiveness* vs. an *efficacy* trial. (An efficacy trial will test how well a particular intervention works under laboratory-controlled ideal circumstances (perfect compliance, dosage etc). An effectiveness trial determines how well the intervention works in a real-world setting.)
1.3.7 Thyroid Orbitopathy and Screening

At present, no screening test is used for early diagnosis of thyroid orbitopathy. As a result, patients are generally referred to ophthalmologists at a somewhat advanced stage in the natural history of their disease. Most often, the moderately to severely affected individuals require major intervention (such as surgery or radiotherapy) in order to treat their disease. While these treatments are quite effective, they usually do not restore a patient’s physical appearance completely. Thus, the disease is costly both in its severe impact on quality of life, and in its affect on the health care system.

The introduction of a screening test for TO would pave the way for experimental early intervention in the disease. One possibility offered is the aggressive use of immuno-suppressive corticosteroids at TO diagnosis in order to prevent further development of the disease.
1.4 Instrument Development and Design

1.4.1 Introduction

An instrument may be defined as a measurement tool. In this sense, the term tool may refer to a device such as a ruler or speedometer, or may simply refer to a set of questions. For the purposes of clarity, it is important that the relationship between instruments designed for general health measurement scales and instruments designed for screening be discussed. In fact, it will be argued here that screening instruments (or diagnostic tests) are actually a particular kind of general measurement scale, and therefore represent a subset of this assessment modality.

Measurement scales in general, and health measurement scales in particular, are intended to provide a numerical score based on subjective data in an effort to quantify a particular construct. In this sense, the term construct refers to a psychological attribute such as anxiety, visual function or intelligence. Since these “constructs” are often somewhat abstract, they are frequently measured through the use of surrogate physical manifestations. Streiner and Norman provide a helpful example here:

> We cannot “see” anxiety; all we observe are behaviours which, according to our theory of anxiety, are the results of it. We would attribute the sweaty palms, Tachycardia, pacing back and forth, and difficulty in concentrating experienced by a student just prior to writing an exam, to his or her anxiety. These proposed underlying factors (such as anxiety or intelligence) are referred to... as constructs. A construct can be thought of as a “mini-theory” to explain the relationships among various behaviours or attitudes.\(^\text{60}\)

A construct is often considered to consist of a number of domains. Thus, an effective instrument will be able to successfully quantify a construct of interest by asking questions concerning all the domains that make up that particular construct. Conclusions may then be drawn about the construct in the population under study. For example, the construct “visual function” may consist of
the domains “central vision”, “peripheral vision”, “night vision” and “depth perception”. Each of these components are elements of visual function, but play different roles and are of varying levels of importance to the central construct. The purpose of a health measurement scale, then, is to measure the level of a particular construct in a population of interest. This information may then be used for such things as clinical trials, policy making, treatment considerations or screening.

In the medical field, a screening tool or diagnostic test is a type of health measurement scale intended to indicate presence or absence of a disease of interest. The construct being tapped in these cases is the disease of interest. Methods of gathering data, as with measurement scales, may be objective or subjective. Further, since the data in diagnostic tests is often dichotomized (disease/no disease), additional descriptive statistics may be used such as sensitivity and specificity (see section 1.3). While there is great variation in the type, mechanism of action and sensitivity of these tools, they share the common purpose of identifying disease early in order to aid clinical decision making. They may, like a general scale, provide information concerning severity or progress of a condition, or may simply indicate its presence or absence.

The remaining discussion in this section concerns itself with design and development of subjective instruments and questionnaires. The general principles, however, may be extended to all kinds of health measurement scales.

1.4.2 Desirable properties of an instrument

There is general agreement concerning the particular attributes that an effective instrument should demonstrate. Among the most important of these concepts are reliability, validity and responsiveness.
1.4.2.1 Reliability

Reliability refers to the fundamental capacity of an instrument to measure something in a reproducible fashion. Put another way, it is an index of the extent to which measurements of individuals obtained under different circumstances yield similar results. A more formal definition of this quantity exists in measurement theory, and is discussed at length by Streiner and Norman. However, for the purposes of this paper, discussion will be limited to broader definitions of the concept that include internal consistency and stability. (Yet another perspective of these parameters is provided by Fig. 1.3).

Measures of internal consistency are based on a single administration of the instrument. The idea is that if a number of questions are addressing the same underlying dimension or domain, then it is reasonable to expect that scores on each item would be correlated with scores on all the other items. This statistic is calculated through the use of methods such as Cronbach's alpha or Kuder—Richardson. However since they based on a single administration, it does not take into account day to day variation. This sort of variation is captured by stability measures.

Stability may be measured through comparing degree of agreement of different administrations of a test to the same observer separated by a period of time (test-retest reliability) or through degree of agreement between different observers on the same administration of the test. (inter-observer reliability). Acceptable measures of stability (and of internal consistency) are related to the objective of the test.
1.4.2.2 Validity

The validity of an instrument may be defined by the question “does the instrument measure what it is intended to measure?” (e.g. Is the speedometer actually measuring the velocity of the car?) Although various types of validity are cited in the literature, they all share the common goal of evaluating the degree of confidence we can place in the tools we use. The most traditional methods adopted to this end should be mentioned—they are: content, criterion and construct validity.

Content validity refers to the adequate coverage of the particular domain under investigation. In the case of a questionnaire, this would translate into having asked a sufficient number of appropriate and relevant questions such that the domain is successfully and accurately covered. The higher the content validity of an instrument, the more confident we can be in drawing conclusions from its results, and the broader are the inferences that we can validly draw about the person (being measured) under a
variety of conditions and in different situations. As an example of content validity, consider a mathematics exam in high school. The ultimate goal of providing the exam is to verify a student's understanding of the material taught in the course. Therefore, the exam should be reflective of the material taught and should adequately cover all important concepts covered. If it falls short of these basic goals, it may not be measuring what it is intended to measure.

Criterion validity may be defined as the extent to which a measure corresponds to an accurate or previously validated measure of the same concept. This form of validity is most often tested through the comparison of the instrument of interest to some gold standard. In this case, gold standard is defined as the best available measure for the domain of interest. For example, a new tool for the measurement of depression may be administered to a sample of patients along with the Beck Depression Inventory (an accepted depression measurement instrument). The performance of the instrument of interest may then be evaluated compared to the accepted measure.

Construct validity is defined as a process in which validity is evaluated as the extent to which a measure correlates with variables in a manner consistent with theory. Thus, having convinced himself/herself that an instrument is successfully measuring a particular construct, the investigator would form some sort of hypothesis based on how patients with disease will respond to the instrument. For instance, to return to the example of visual function, an investigator may hypothesize that individuals with keratoconus, (a disease of the cornea) have poorer visual function than individuals with cataract despite having identical visual acuities. Therefore, the keratoconus patients would be expected to perform more poorly than the cataract patients on a visual function test. If this turns out to be the case, support for construct validity has been achieved. However if no difference is measured the problem could be that the instrument is good, but the theory is wrong, or that the theory is good but the instrument is flawed, or that both are flawed.
1.4.2.3 Responsiveness

Responsiveness refers to the capability of an instrument to detect changes over time. Thus, if a condition deteriorates or improves over a particular time period, an effective instrument should be able to detect and reflect that difference. These changes may be quantified through statistical tests (such as an F-test) or various measures of the strength of effect, expressed as a ratio of the difference between groups to the variability within groups.
1.4.3 Steps in instrument development

The basic steps in the development of a questionnaire are the following:

Searching the literature, devising the items, scaling responses, selecting the items, scaling the instrument and validating the instrument. (The following discussion, unless otherwise stated, is adapted from 60—Other issues such as methods of administration and ethical issues are covered in more detail in this reference).

1.4.3.1 Searching the literature

Before launching a development project, it is necessary to conduct a thorough literature search in order to confirm that no appropriate instrument already exists. In particular, large searchable databases such as Medline provide a useful beginning to such an exploration. Alternatively, particular textbooks or articles on related subjects may provide useful reviews or citations which inform the investigator as to the current state of the research in the field. A final potential source for instruments lies in a publication entitled *Measuring Health: a guide to rating scales and questionnaires* 65 which provides a critical review of the more popular existing tools in the health care field.

1.4.3.2 Devising the items

This represents the first and most important step in questionnaire construction. As Streiner and Norman explain:

> ...no amount of statistical manipulation after the fact can compensate for poorly chosen questions; those that are badly worded, ambiguous, irrelevant, or—even worse—not present.60

A common first step to producing items is to look at what others have done in the past. It is logical to conclude that if someone else has gone through the tortuous process of scale development, that often the wording and style of the items they have included have been mulled over and tested at
length, and therefore may be well worth using. Streiner and Norman quote a passage from Goldberg in their text:

Items devised around the turn of the century may have worked their way via Woodworth’s Personal Data Sheet, to Thurstone and Thurstone’s Personality Schedule, hence to Bernreuter’s Personality Inventory, and later to the Minnesota Multiphasic Personality Inventory, where they were borrowed for the California Personality Inventory and then injected into the Omnibus Personality Inventory—only to serve as a source of items for the new Academic Behavior Inventory.60

Commonly, investigators will find a limited source of questions from other instruments to aid them in construction—if they were able to gather all questions from other instruments, it is likely that they would have no need for a new instrument. In this case, the best sources for items include patients who have suffered or are suffering from the condition of interest, and clinicians or experts in the field, who deal with these patients on a daily basis.

The most common way of gathering information from patients is to hold focus groups. In these sessions, a small group of patients (up to about 12) meet with a facilitator to discuss the condition and talk openly about personal experiences relevant to the investigation. The session may be recorded or include an observer who takes notes for later analysis.

In depth interviews with clinicians offer another source of items. The investigator meets with experts in the field, and discusses pertinent items that may be included on the form. Usually the presence of several experts, or a series of meetings provides more successful and exhaustive coverage of the items. Additionally, the investigator may seek expert opinion in journal articles and textbooks written on the subject of interest.

1.4.3.3 Scaling Responses

Having produced a complete set of items that adequately and accurately reflects the domain of interest, the investigator must decide exactly how to present those items on the form. Questions may
be categorical, open-ended or closed-ended, offer dichotomous answer choices (such as yes/no or true/false) or require more specific response as in ordinal or continuous scales. Streiner and Norman provide a very useful chapter on this subject in their text, with an extensive discussion on scaling methods including Guttman and Thurstone methods. These are not directly relevant to the discussion, and therefore are not included.

Categorical judgment refer to questions that may simply be answered with a check mark, or as a yes—no response. Naturally, in cases such as “ethnic background” or “gender” the format of such questions are trivial. However there are a number of cases where apparently dichotomous items would be better scaled on a more continuous scale. For example, the question “Is your vision blurry” on the one hand may be responded to with a simple yes—no response key, or may be addressed with a more continuous response that includes no, a little, a moderate amount, a great deal. By limiting the response to either all positive or all negative, as in the first case, one loses a good deal of information that would be revealed by the second format. Secondly, different people may have different ideas of the meaning of the questions to them—thus what constitutes a wholly positive or negative response may differ from person to person. Thirdly, there is a loss of efficiency of the instrument when dichotomous rather than continuous outcomes are measured. In this context, efficiency refers to the number of subjects required in order to show an effect. This effect has been well illustrated. One potential objection to the continuous scale is that the researcher may only be interested in yes/no answers, and therefore the extra effort is unnecessary. A response to this is framed nicely by Streiner:

This argument confuses measurement with decision-making; the decision can always be made after the fact by establishing a cutoff point on the response continuum, but information lost from the original responses cannot be recaptured.
A great number of methods for scaling continuous responses have been developed. Major methods include visual analogue scales (VAS), Likert scales and adjectival scales.

The VAS and adjectival scales are similar in design. Both involve a line of fixed length, with anchors on either side describing a condition such as "no pain" on one end, and "pain as bad as it could be" at the other. The respondent then places an X or a vertical line on the scale where they feel they belong. The adjectival scale uses the same principal, however includes additional description along the length of the fixed line to aid respondents to hone in on their particular health state. Adjectival scales may also have discrete boxes representing particular responses (see fig 3). The Likert scale is similar to the adjectival scale, however answers are placed on an agree-disagree continuum.

There are two final interesting points on which to conclude. First of all, it is worthwhile noting that in the case of continuous ratings, there is good evidence that individuals are unable to discriminate much beyond seven levels of response. Theoretically, reliability will continue to increase with increasing number of categories. However, there is a sufficiently small improvement in reliability once numbers begin to exceed seven categories that investigators are content to limit the response choices to that number. Secondly, although from a purely statistical standpoint these scales are truly ordinal, and not continuous; the evidence suggests that they may successfully be analyzed as interval scales without the introduction of any great amount of bias.60

1.4.3.4 Selecting the items

In order to retain the best possible questions, items should be examined carefully for clarity and face validity, and should be homogenous about a particular domain without being redundant. Potential problems include ambiguous terms, double-barreled questions, use of jargon, value laden words, negatively-worded items and lengthy questions. Furthermore, the instrument should be tailored to the age and educational background of those who will be completing it.
Ambiguous terms such as "often" "trivial" and "too much" will be interpreted differently by different responders and should therefore be avoided. Clear wording provides the least opportunity for bias.

Double barreled questions are those items that ask two or more questions at the same time. An example of this would be a question such as "Do you feel happy and healthy?" The question does not leave room for those who feel just happy and not healthy or vice versa, and thus will cause problems.

Jargon, as Streiner and Norman point out, can slip into a scale or questionnaire quite insidiously. Since investigators use technical vocabulary regularly, it is sometimes forgotten that some words are not part of every day vocabulary. Items containing jargon should be reworded or removed.

Value laden wording will bias the response to a question. A question such as "should our dedicated and overworked graduate students be paid more?" Implies an answer, and leads to responses that may agree with the sentiment of the question rather than the answer of interest.

As a rule, negatively-worded questions should be avoided. It is better to phrase an item as "I feel ill most of the time" than "I rarely feel well". Also, negative words such as "not" or "never" are best reworded.

Last, but not least, questions should be kept as short as possible, while still maintaining maximum comprehensibility. The ideal length of a question falls between 10-20 characters, while longer questions tend to have lower associated reliabilities.

The homogeneity of a questionnaire may be tested through item to item, and item to total correlations. This concept forms the basis of the Cronbach’s alpha test for internal consistency described in section 1.4.2. While items that are assessing the same domain should be moderately correlated, it is less useful to include two items that are asking exactly the same question (or that are
correlated with an $r$ greater than 0.9). That is, unless the question is of particular importance and is included as a checkpoint by the investigator.

1.4.3.5 Scaling the instrument

In the final stages of questionnaire development, the investigator needs to consider the method by which to summarize the data collected by the instrument. Given that the more usual and desirable characteristic of a questionnaire is that it have a number of items examining a particular domain, it logically becomes necessary to produce a meaningful score from the items that may be used in clinical practice.

The easiest method of doing this is simply to add up the scores on the individual items. This method is easy to work with, and makes few assumptions; the only implicit assumption being that the items are equally important in contributing to the total score. Therefore, it is not surprising to find that this method has garnered a fair bit of use in the health care field.

Summing (or averaging) the individual items suffers from one drawback: There may be items which are far more important than other items on the questionnaire. Thus it may be useful to attempt to apply some sort of weighting to the items in order to reflect their varying importance. This may be applied conceptually, based on the investigators understanding of the disease, or it may be determined using multiple regression. However, as it turns out, weighting more often than not provides minimal improvement to a scale, as is therefore often not worth doing. Streiner and Norman provide a useful discussion of precisely why this is. To summarize briefly, when a scale is made up of a large number (~40) items contributing to a single score, the correspondingly minute amount that each item contributes to the scale is altered hardly at all by the inclusion of weighting. Weighting may have an effect when 20 or fewer items are used, and this may be explored further by multiple regression. There
are two other methods of providing weighting to an instrument that are often included unintentionally. The first is asking more questions about one domain than another, thereby boosting the score of the first. For an example, consider a questionnaire measuring the construct of visual function once again. If 10 questions concern themselves with central vision, and 2 questions tap into peripheral vision, an obvious preference is provided for the first domain. This may be controlled through devising subscales which in effect divide a particular domain score by the number of questions asked, thereby having each domain providing a proportional score to the total. The second situation involves including items that are highly correlated. This may be eliminated through pilot testing the instrument, and eliminating those items that are asking redundant information.

1.4.3.6 Validating the instrument

Having constructed the instrument, the final step before its clinical use is to validate it on a population of interest. This simply involves distributing the questionnaire to the intended recipients, collecting, entering and analyzing the data. From this study population, one may fine tune any ambiguities of the instrument, determine how long it takes to complete and how well patients are able to respond to it. This also provides data from which one can measure the internal consistency of the instrument. Validity of the instrument may also be tested as described above in section 1.4.2.

It is worth reiterating that often the continuous data collected by an instrument will be later dichotomized for the sake of clinical utility. As such, cut points may be necessary above which an individual may determined to be considered positive or negative for a particular condition (i.e. screening). Once this cut point is established (usually through the use of a gold standard), it is possible to measure the various psychometric properties of the instrument (see section 1.3.2).
1.5 Study Rationale

In current practise, a patient diagnosed with GH is referred by a general practicioner to an endocrinologist. There, treatment options are discussed following which the patient is observed and treated with the goal of achieving euthyroidism. If eye problems manifest themselves, the clinician will refer the patient to an ophthalmologist for treatment. In the present setting, there is great variation in the amount of knowledge individual clinicians have regarding TO. This often results in patients being referred rather late in the development of disease, requiring ophthalmologist to use the most invasive and expensive interventions.

It is the feeling of the VGH UBC Thyroid Orbitopathy Research Group, based on the management since 1976 of over 1800 cases, that early intervention in patients with developing TO would reduce the number of severe cases of this disease. The identification of these individuals should be possible through focussing on particular symptoms of the disease. If these risk factors can indeed be used to predict the course of disease, potential prophylactic interventions may be introduced early in the progression of the condition which would aid in control and prevention of the more serious cases. This would lessen the impact of TO on the health of these individuals, and substantially decrease the cost of care of affected patients.

1.6 Hypothesis

The ultimate purpose of this study is to evaluate the hypothesis that patient’s self-assessment of presence or absence of particular early symptoms of TO may be used as an effective early screening test for the disease. The study is labeled a pilot study since it will be both to identify the particular symptoms that are associated with disease as well as to create a rule from these symptoms which may be used as a screening instrument. Further validation of the rule on a new cohort of patients will be required.
2.0: **Study Design**

2.1 **Introduction**

The development of the Thyroid Orbitopathy Indicator (TOI) was initiated in September of 1998. At that time, the project was divided into two distinct phases, characterized by the timeline provided in Appendix 1.

In the first phase, the content of the questionnaire was assembled and developed through literature review, expert consultation and focus groups (section 2.2). The end result of this phase was the production of a complete and reasonably brief questionnaire which could be easily understood and filled out by patients.

In January of 1999, the second phase was initiated. In this stage, patients newly diagnosed with Graves’ Hyperthyroidism were referred to the Vancouver General Hospital Eye Care Centre. There, they received a copy of the TOI to complete, following which they received a standardized clinical exam administered by one of the clinic ophthalmologists. (Section 2.3) The data from the forms was then analyzed. (Section 2.4)

The project was funded by Drs. Rootman and Dolman, and was approved by the VGH and UBC board of ethics.
2.2 Questionnaire Design and Development

The development of the content of the Thyroid Orbitopathy Indicator (TOI) followed the general strategy outlined in section 1.4.

2.2.1 Devising the items

Items included in the TOI were researched and assembled from textbooks, journal articles and ophthalmologists. These sources represent a collection of opinions of a number of experts in the field, and therefore provided a near exhaustive source of items to be included.

2.2.1.1 Literature Search

The scientific literature was examined through the U.S. National Library of Medicine’s Medline. A literature search incorporating articles published from 1966 to 1998 was conducted, and pertinent articles relating to the clinical manifestations of Thyroid Orbitopathy were examined (e.g. 26, 27, 28).

Textbooks provided a second major source of items. The concise nature of textbook discussion made abstracting items reasonably straightforward. Texts used included Char, Falk, DeGroot, Pope and Rootman.

2.2.1.2 Expert Consultation

Following the literature search, and the assembly of a draft of items to be included, a number of meetings involving two clinical ophthalmologists with expertise in TO (Dr. Jack Rootman and Dr. Peter Dolman), were convened. Common symptoms were discussed and recorded, producing a list of 67 items. Further discussion resulted in the removal of some redundant or unimportant items, resulting in a list containing 53 possible questions, including non symptom-related questions. These included questions on stress, quality of life, family history of disease and social implications of the condition. In the final draft of the questionnaire, modified versions of these items were also included for use.
However, the majority of this discussion will be focused on the symptom questions alone, as these were what was used to construct the scale.

### 2.2.2 Scaling Responses and Selecting the items

The basic format of the questions chosen for the TOI was based on the organization of another Ophthalmological instrument, The VF-14. This popular and accepted visual function assessment form was originally designed by Steinberg and colleagues for cataract patients.\(^{48}\) It has, since then, been validated in other areas of Ophthalmology.\(^{70-72}\) The format of the questions is closed-ended and ordinal. The respondent indicates how much difficulty s/he experiences performing particular activities of daily living by checking the appropriate response on the form (no difficulty, a little bit, a moderate amount, or a great deal of difficulty). This format was applied to each of the symptom questions on the TOI, to provide a consistent method of responding to the items. The wording of these items was further refined in order to avoid double-barreled and vague questions.

The final polishing of the TOI draft was made possible through focus groups organized with the collaboration of the Vancouver Chapter of the Thyroid Foundation of Canada.

#### 2.2.2.1 Focus Groups

Having chosen the format of the questions, drafts of the questionnaire were presented to a series of three focus groups for discussion. All three groups were composed of a mixture of patients with Graves' Hyperthyroidism (GH) and Thyroid Orbitopathy (TO).

The first two groups were held at Vancouver General Hospital, and consisted of members of the Thyroid Foundation of Canada. The groups were composed primarily of white women of 45 years of age or more. The first group had 12 individuals in attendance and was held on January 21, 1999. The second was conducted on March 2, 1999 and had 6 individuals in attendance. In each session, the most current draft of the TOI was distributed to the Group members, read, and completed over the
first 10 minutes of the meeting. Each item was then discussed one at a time, and explored for clarity, brevity and ease of response. Notes were made by the study coordinator, and the draft was updated accordingly. At the conclusion of the item-by-item examination, a discussion was conducted concerning the general utility and feel of the questionnaire. This also provided an opportunity for Group members to bring up issues such as the inclusion of additional items that were not present, and the eventual clinical use of the instrument.

The third focus group consisted of a collection of individuals assembled through an advertisement in the Thyroid Foundation of Canada newsletter, distributed nationally twice a year. Readers were invited to contact the study coordinator through regular postage or email with contact information. Those individuals with access to email were contacted by the study coordinator, and an on-line focus group was established, consisting of ten individuals from across the country. They were each given copies of the survey and urged to complete it and discuss it on-line.
2.2.3 Final Draft

The final draft of the TOI was completed on March 3, 1999. A copy of the form is included in Appendix 2. The final form had a total of 49 items on it covering symptoms, demographic information (birth date, gender, place of birth), behavioural questions (stress level, smoking behaviour), family and personal disease history questions.

There were a total of 29 symptom questions broken down into distinct sections, comprised of between 4 and 8 questions each, relating to the 5 major domains that appear to be affected by the disease: exposure, inflammatory signs, proptosis, vision effects and strabismus. The final questionnaire was four pages in length.
2.3 Inception Cohort Recruitment Design Architecture

2.3.1 Patient Selection and Referral

Initially, six endocrinologists were invited to participate in the study. They were asked to refer individuals over eighteen years of age who were newly diagnosed with GH (i.e. had no previous history). To assist the clinicians, the study coordinator provided each of the doctors with a number of copies of an information sheet containing a brief description of the study, and a space for interested parties to write in name and phone number. (Appendix 3). The research coordinator subsequently contacted the doctor’s offices at regular intervals (weekly or fortnightly), collecting these introduction sheets, and entering the pertinent information on to the study data base. At point of entry, each patient was assigned a study number. Patients were then contacted, reminded about the study and invited to set an appointment at ~2 months after their initial endocrinologist appointments. Appropriate times were arranged with the clinical secretary at the Eye Care Centre, and booked in advance as “study patient” times. All appointment times were entered into the database.

2.3.2 Questionnaire Administration

On arrival at the hospital, patients were greeted and given a questionnaire package consisting of a consent form (Appendix 4) and the Thyroid Orbitopathy Indicator (TOI). (Several other questionnaires were also included. These are discussed in section 4.4). In the interests of confidentiality, forms were labeled only with the patient study number. A clinical examination form was provided for the doctor’s use. Once the package was completed, patients were admitted to see the doctor for a standard Thyroid Orbitopathy physical exam. During the course of this examination, the doctor did not discuss history or symptomology with the patient. (The questionnaire was completed before the clinical assessment to avoid any effects that the exam may have had on answers to symptom
questions). The completed information package was then entered into the database. Consent forms and other hard copies were filed in a secure office.

Two of the patients in the sample were able to communicate in English but were unable to read. In these cases the study coordinator provided assistance by reading out the questions and recording the answers on their behalf.

2.3.3 Clinical Evaluation

Clinical evaluation was conducted by one of two clinicians involved in the project: Dr. Jack Rootman or Dr. Peter Dolman. The method of clinical evaluation used in this study follows the methodology described by Rootman. A copy of the evaluation form is included in Appendix 5.

The examination was divided into five major categories: general, psychophysical, orbital, ocular movement and ocular assessment.

General assessment of the patient included observations concerning facial contours, and lateral and vertical symmetry of facial, lid, orbital, and ocular structures. Orbital and periorbital structures were palpated, as were the preauricular and cervical nodes. Lids and conjunctiva were assessed for position and alterations in structure and degree of injection and were graded on a 2 point scale. Preseptal, pretarsal, conjunctival edema and degree of chemosis were assessed and graded on a 4 point scale. The interpalpebral fissure, upper and lower lid retraction, Margin Reflex Distance lid lag and degree of scleral show were also measured.

Psychophysical examination included a study of the best corrected visual acuity, colour vision assessment and pupillary examination. The latter included assessment of size, symmetry, light and near reaction as well as a check for afferent pupillary defects.

Orbital examination assessed degree of horizontal and vertical displacement of the globe. Degree of proptosis was measured using a Hertel exophthalmometer, with the patient regarding the
axial direction. Ocular movement was evaluated by recording ductions in the four cardinal positions by degrees using a Krimsky method. (This method has been previously validated using objective perimetry, and was demonstrated this method to be accurate to within 5% when compared to quantitative parametric methods).

Finally, ocular examination included measurement of the intraocular pressure in the primary position and upgaze and biomicroscopy of the cornea, conjunctiva and fornices. Additionally, the fundus, optic nerve head, retinal blood vessels and choroid were examined.

Degree of chemosis and upper and lower lid edema for each eye were scored on a 4 point scale (0,1,2,3). Conjunctival injection, lid injection and lid edema, pain at rest and movement pain were scored on a 2 point scale (0,1). Increasing values indicated a more serious condition. A final Inflammatory Score was then calculated for the worst eye, providing a score of up to 10.

At the conclusion of the clinical examination, clinicians used the information gathered to classify patients as having either no disease, or mild, moderate or severe disease. This diagnosis was indicated and graded based on system that summarized vision (normal, abnormal), Inflammatory Score (0-8, mild, moderate, severe), Strabismus (absent, intermittent, constant) and Appearance (normal, mild, moderate, severe). If any of the latter three categories were not normal, disease was considered to be present, and was indicated as such under a heading labeled diagnosis. For the purposes of data analysis the outcome indicated under diagnosis was dichotomized into presence or absence of disease.
2.3.4 Database Organization

Data from clinical forms and from questionnaires were entered onto Microsoft Access electronic forms designed for the study. Corresponding spreadsheets were created summarizing each form individually, cross-referenced by study number. Data back-ups were made on a regular basis, and stored on floppy disk.
2.4 **Statistical Analysis**

Data analysis was conducted using SPSS statistical software (Version 8.0), Microsoft Access, Microsoft Excel and a Texas Instrument hand calculator.

2.4.1 **Sample Size Considerations**

In order to ensure that the proposed analysis was feasible, endocrinologists were interviewed by the study coordinator in order to determine what volume of patients were expected at each practice. Estimates ranged from 2-8 new patients a month depending on the practice. In order to provide an appropriate conservative measure of new patients, the minimum estimate of 2 per month per endocrinologist was used. This in turn produced an expected 12 patients per month invited to attend. Since it was expected that not every patient may wish to take part in the study, a rough participation rate of 75% was expected, yielding 9 patients per month, or 108 per year. For the purposes of power, it was decided that as many patients as possible would be recruited within the time-frame of the study.

2.4.2 **Questions used**

For the purposes of analysis, only symptom questions were examined. These were chosen above all others to maintain greatest simplicity of the rule, to focus on hard subjective outcomes and to agree most closely with the original hypothesis. (i.e. that patients are aware of early signs and symptoms of disease).

As explained in section 2.2.2., the symptom questions of the TOI were responded to based on a four point scale (no symptom, a little, a moderate amount, a great deal). For the purpose of analysis, this scale was dichotomized into a two point scale (symptom present or symptom absent). A listing of the original format of the questions is provided in Appendix 2.
2.4.3 Chi-squared tests

The chi-square test is a statistical test used for the comparison of independent proportions in a contingency table. It tests the null hypothesis that the column variable and the row variable are independent. To put it another way, the chi-square test is used to determine whether a row variable is mathematically associated with a column variable. In the case of a 2X2 table for symptom and disease, this translates to testing whether the presence (or absence) of a symptom is associated with the presence (or absence) of disease. The test is conducted in the following manner: Consider a 2X2 table summarizing the results of a diagnostic test such as the following:

Table 2.1 Chi-Square Test

<table>
<thead>
<tr>
<th>Test:</th>
<th>Disease:</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive: (D+)</td>
<td>Negative: (D-)</td>
<td>Totals:</td>
</tr>
<tr>
<td>Positive: (T+)</td>
<td>a</td>
<td>b</td>
<td>a+b</td>
</tr>
<tr>
<td>Negative: (T-)</td>
<td>c</td>
<td>d</td>
<td>c+d</td>
</tr>
<tr>
<td>Totals:</td>
<td>a+c</td>
<td>b+d</td>
<td>a+b+c+d (N)</td>
</tr>
</tbody>
</table>

Having collected sufficient data, an investigator may construct a table of observed values. In the first step of a chi-square test, a table of expected values are generated. These are produced by applying the statistical principle that the probability of 2 independent events occurring is equal to the product of their individual probabilities. That is,

\[ P(A \cap B) = P(A) \times P(B) \]

Thus, a table of expected values would look like the following:
Table 2.2 Expected Values in the Chi-Square Test

| Test | Disease: | | | Totals: |
|------|---------|------------------|------------------|
| Positive: (T+) | (D+) | (D-) | Totals: |
| (a+b)(a+c)/N | (a+b)(b+d)/N | a+b |
| Negative: (T-) | (c+d)(a+c)/N | (c+d)(b+d)/N | c+d |
| Totals: | a+c | b+d | a+b+c+d (N) |

As can be seen from Table 2.2, the row and column totals of expected counts are the same as those in the observed table. Only the body of the table is changed to reflect expected counts.

The final step of the chi-square test is to generate a test statistic. To perform the test for the data in a contingency table with r rows and c columns the following sum is calculated:

\[
\chi^2 = \sum_{i=1}^{rc} \frac{(O_i - E_i - 0.5)^2}{E_i}
\]

Where \( O \) represents the observed frequency

\( E \) represents the expected frequency

\( r \) represents the number of rows in the contingency table

\( c \) represents the number of columns in the contingency table

\( i \) represents the next cell on the body of a contingency table

(NB: the Yates correction factor (-0.5) is applied to chi-square tests involving 2x2 tables. It is generally not necessary for tables with greater than 1 degree of freedom).

The probability distribution of this sum is approximated by a chi-square (\( \chi^2 \)) distribution with (r-1)(c-1) degrees of freedom. For instance, a 2 X 2 table has (2-1)(2-1) = 1 degree of freedom; a 3 x 4 table has (3-1)(4-1) = 6 degrees of freedom. To ensure that the sample size is large enough to make this approximation valid, no cell should have an expected count less than 1, and no more than 20% of the cells should have an expected count less than 5. (In these cases, the non-parametric Fisher's Exact Test is usually used).
The value for a particular table derived from the sum calculated above is then checked against a table of areas for the chi-square distribution. From this table (or from a statistical program) the significance (p-value) of the results may be evaluated, and the null hypothesis (of no association between rows and columns) may be alternatively rejected or not rejected. For the purposes of this analysis, a p-value of less than 0.05 was considered significant.

2.4.4 Development of the Decision Rule

Having determined which symptoms are associated with disease using the chi-square test, the final step in the data analysis called for the design of a clinical decision rule. The objective of this rule was to combine the disease-associated symptom questions in a manner that provided the simplest, shortest manner of capturing all patients with clinically manifest disease and a minimum of patients without disease. The guideline was put forth in this manner in order to create a rule that would maximize sensitivity (as opposed to both sensitivity and specificity), since the clinicians were willing to bear the expense and time of examining false positives in order that no true positives were missed.

Once the rule was determined, its measures of diagnostic accuracy were tested through the construction of a 2 x 2 table. (As illustrated in section 1.3). These results are reported in section 3.0.
3.0: Results

3.1 Participation rate

Of the original 6 endocrinologists initially agreeing to participate in the study, only one (Dr. G. E. Wilkins) referred patients to participate in the study. This resulted in substantially less patients participating then expected, at an average rate of 4.5 patients per month. Various incentives were attempted to encourage more active participation from the other endocrinologists, including regular visits by the study coordinator and phone calls. In direct interviews with these clinicians, in which the lack of participation was brought up, the endocrinologists claimed variously that they had seen no patients, found that no-one was interested or were forgetting to ask.

After 4 months of study had passed, and the poor enrolment rate had been noted, the criteria for patients was loosened to include any patients diagnosed with GH in the previous 6 months. The first patient enrolled in the study and was referred in late January, and seen at the Eye Care Centre on April 7, 1999. At the time of analysis, 66 patients had been referred by Dr. Wilkins. 50 of these 66 patients were eventually seen at the Eye Care Centre, a participation rate of 76%. The major reason for not participating was patient’s unable to find time to make an appointment. Other reasons included missed appointments with no new time rescheduled and lack of interest by patients.
### Descriptives of study population

Descriptive data from the cohort of study patients is summarized in the following tables.

Further discussion of these results may be found in section 4.

**Table 3.1: Ethnic Background**

<table>
<thead>
<tr>
<th>Ethnic Background</th>
<th>Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>22</td>
</tr>
<tr>
<td>Chinese</td>
<td>22</td>
</tr>
<tr>
<td>East Indian</td>
<td>1</td>
</tr>
<tr>
<td>Other</td>
<td>5</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>50</td>
</tr>
</tbody>
</table>

*other may be unknown or unreported

**Table 3.2: Gender Distribution**

<table>
<thead>
<tr>
<th>Gender</th>
<th>Cohort #</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>37 (74%)</td>
</tr>
<tr>
<td>Male</td>
<td>13 (26%)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>50</td>
</tr>
</tbody>
</table>

**Table 3.3: Age Distribution**

<table>
<thead>
<tr>
<th>Age</th>
<th>Cohort #</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average (SD)</td>
<td>39.7 (12.3)</td>
</tr>
<tr>
<td>Minimum</td>
<td>20</td>
</tr>
<tr>
<td>Maximum</td>
<td>68</td>
</tr>
</tbody>
</table>

**Table 3.4: Smoking Behaviour**

<table>
<thead>
<tr>
<th>Smoker</th>
<th>Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoker</td>
<td>Current</td>
</tr>
<tr>
<td>Yes</td>
<td>7 (14%)</td>
</tr>
<tr>
<td>No</td>
<td>43 (86%)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>50</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Smoker</th>
<th>Ever Smoked</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>17 (34%)</td>
</tr>
<tr>
<td>No</td>
<td>23 (46%)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>50</td>
</tr>
</tbody>
</table>

**Table 3.5: Distribution of disease**

<table>
<thead>
<tr>
<th>Disease</th>
<th>Cohort #</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present</td>
<td>12 (24%)</td>
</tr>
<tr>
<td>Absent</td>
<td>37 (76%)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>49</td>
</tr>
</tbody>
</table>

*One patient was missing clinical examination data

**Table 3.6: Disease Status**

<table>
<thead>
<tr>
<th>Disease Status</th>
<th>Cohort #</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absent</td>
<td>37</td>
</tr>
<tr>
<td>Mild</td>
<td>10</td>
</tr>
<tr>
<td>Moderate</td>
<td>1</td>
</tr>
<tr>
<td>Severe</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>49</td>
</tr>
</tbody>
</table>
3.3 Significant Symptoms

Using the chi-square procedure described in section 2.4.3, symptoms significantly associated with presence of disease were identified. These are summarized on the following table:

Table 3.7: Significant Symptoms

<table>
<thead>
<tr>
<th>Item #</th>
<th>Question</th>
<th>p-value</th>
<th>Odds Ratio</th>
<th>95 % C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>Eyes watering more than normal?</td>
<td>0.055</td>
<td>3.69</td>
<td>0.93</td>
</tr>
<tr>
<td>12</td>
<td>Redness in your eyes or eyelids?</td>
<td>0.053</td>
<td>3.78</td>
<td>0.97</td>
</tr>
<tr>
<td>15</td>
<td>Swelling or feeling of fullness in one or both of your upper eyelids?</td>
<td>0.036</td>
<td>4.36</td>
<td>1.11</td>
</tr>
<tr>
<td>17</td>
<td>Bags under the eyes?</td>
<td>0.007</td>
<td>8.21</td>
<td>1.57</td>
</tr>
<tr>
<td>24</td>
<td>Do your eyes seem to be open too wide?</td>
<td>0.015</td>
<td>8.10</td>
<td>1.56</td>
</tr>
<tr>
<td>26</td>
<td>Is your vision blurry (even with glasses/contacts?)</td>
<td>0.023</td>
<td>5.08</td>
<td>1.27</td>
</tr>
</tbody>
</table>

Items 9 and 12 were not significant by the definitions laid down at the outset of the study. However, the p-values of these items approached significance, while the remaining items were nowhere close. They were therefore included in rule construction. (A complete list of items and their associated p-values and confidence intervals in given in Appendix 6).
3.4 Final Rule Construction

The final rule was arrived at through trial and error, based on the criteria set forth in section 2.4.4. A practical expression of these criteria was: To determine the simplest possible rule that would capture all of the diseased individuals and a minimum of the non-diseased. As indicated in table 3.5, there were 12 patients with disease, and 37 without disease. The best rule would therefore designate all 12 of these individuals as disease positive and a maximum number of the 37 remaining patients as disease negative. Some examples of rules attempted that captured all patients positive for TO included the following:

- Positive for item 15 or item 17
- Positive for item 15 or item 17 and one of item 9, 15 or 26
- Positive for one of 17, 24 and 26 and one of 15 or 17

The final rule, producing the best sensitivity and specificity was the following:

Positive for one of items 15 and 17 and one of items 12, 24 or 26

This produced a 2 x 2 table as follows:

Table 3.8: Final Rule

<table>
<thead>
<tr>
<th>Rule:</th>
<th>Disease:</th>
<th>Positive: (D+)</th>
<th>Negative: (D-)</th>
<th>Totals:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive: (T+)</td>
<td>12</td>
<td>8</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>Negative: (T-)</td>
<td>0</td>
<td>29</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>Totals:</td>
<td>12</td>
<td>37</td>
<td>49</td>
<td></td>
</tr>
</tbody>
</table>
The corresponding measures of diagnostic accuracy were as follows:

**Table 3.9: Measures of Diagnostic Accuracy**

<table>
<thead>
<tr>
<th>Epidemiological Term</th>
<th>Probability</th>
<th>2 X 2 Algebra</th>
<th>Value</th>
<th>95% C.L.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity:</td>
<td>( P(T+ \mid D+) )</td>
<td>( \frac{a}{a+c} )</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Specificity:</td>
<td>( P(T- \mid D-) )</td>
<td>( \frac{d}{b+d} )</td>
<td>0.78</td>
<td>0.65 - 0.91</td>
</tr>
<tr>
<td>False Positive:</td>
<td>( P(T+ \mid D-) )</td>
<td>( \frac{b}{b+d} )</td>
<td>0.22</td>
<td>0.09 - 0.35</td>
</tr>
<tr>
<td>False Negative:</td>
<td>( P(T- \mid D+) )</td>
<td>( \frac{c}{a+c} )</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>Positive Predictive Value (PPV):</td>
<td>( P(D+ \mid T+) )</td>
<td>( \frac{a}{a+b} )</td>
<td>0.60</td>
<td>0.49 - 0.71</td>
</tr>
<tr>
<td>Negative Predictive Value (NPV):</td>
<td>( P(D- \mid T-) )</td>
<td>( \frac{d}{c+d} )</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Post-test Likelihood (PTL(-))</td>
<td>( P(D+ \mid T-) )</td>
<td>( \frac{c}{c+d} )</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>Pre-test Likelihood</td>
<td>( P(D+) )</td>
<td>( \frac{a+c}{N} )</td>
<td>0.25</td>
<td>0.19 - 0.31</td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td>( \frac{a+d}{N} )</td>
<td>0.84</td>
<td>0.79 - 0.89</td>
</tr>
<tr>
<td>Likelihood Ratio (LR+)</td>
<td>sensitivity</td>
<td>( \frac{a}{a+c} )</td>
<td>4.54</td>
<td></td>
</tr>
<tr>
<td>Likelihood Ratio (LR-)</td>
<td>(1-sensitivity)</td>
<td>( \frac{d}{b+d} )</td>
<td>0.00</td>
<td></td>
</tr>
</tbody>
</table>

Note that only five of the six symptoms were necessary in the construction of the simplest rule.
3.4.1 Limitations of the data

External validity is of fundamental importance in the construction of clinical decision rules. If the rule cannot be generalized to other similar groups of patients, it is of absolutely no value. Thus, it must be noted that the measures of diagnostic accuracy provided in table 3.9 above do not accurately reflect the true performance of this rule. The reason for this is simply that the parameters above and the original rule were both designed using the same data. Furthermore, the “best possible“ rule was developed using this data. This must mean that the corresponding measures of test utility will be very good. (If they weren’t, no useful rule could have been developed). The purpose of calculating these quantities at this point was simply to establish which rule would produce the best possible screening test using an “ideal” population. (Also of concern is the small sample size (n=49) used to construct the rule. This is discussed further in section 3.7).

In order to determine true measures of test utility, the rule must be administered to a new population of patients. In this new experiment, the validation stage of the study, the experimental protocol described above should be matched as exactly as possible. This includes use of the identical questionnaire package, diagnostic rules and recruitment strategy. The rule may then be tested on the final data, and its true measures of diagnostic accuracy (sensitivity, positive predictive value etc.) be ascertained.
4.0: Concluding Remarks

4.1 Conclusions

In recent decades, a shift in the paradigm of medical care has been initiated. A strong movement toward complementing traditional objective measures and quantification with more subjective and qualitative measures has arisen. This is demonstrated by the increasing use of subjective self-administered questionnaires appearing all over the health care system from Ophthalmology (The VF-14 and the NEI-VFQ) to arthritis and cancer. Furthermore, greater interest is now taken in a construct referred to as Quality of Life. It has gradually become the governing purpose in the practice of medicine in the first world to maximize the quality rather than the quantity of life of patients. (Evidence for this paradigm shift is provided by the increasing numbers of studies that incorporate Quality of Life as a primary or secondary outcome\textsuperscript{71, 74, 75}).

This analysis concerns patients afflicted with a condition called Thyroid Orbitopathy (TO). TO is not a fatal disease, however its wide range of intrusive symptoms often lead to a profound deleterious effect on the lives of affected patients. Therefore, clinicians who treat this condition on a daily basis must constantly seek new and improved methods of treating their patients in the constant battle to maintain and improve quality of life.

The objective of this study was to design and develop a screening instrument for the detection of (TO). The purpose of the instrument is to provide a rapid diagnostic test for TO to endocrinologists and general practitioners for use on patients diagnosed with Graves' Hyperthyroidism. This will serve to capture those patients with TO early in the disease process, and have them referred to ophthalmologists where their condition can be treated before it progresses to a severe stage.
The preceding analysis has summarized the methodology and results of a pilot study intended to develop this instrument, currently named the Thyroid Orbitopathy Indicator (T.O.I.). From this instrument, a set of symptoms were identified that appear to yield a useful screening rule for TO. This rule ("The Vancouver Rule") should be validated in a second GH cohort trial in order to confirm its clinical utility. The new trial should adopt the same methodology as described in this report. (As described in section 3.4.1 above).
4.2 Sources of Error

The introduction of bias into the study was possible at a variety of points.

Firstly, at the recruitment stage, it is likely that patients who were more interested or concerned about their eye conditions may have expressed more willingness to participate in the study, and to find time to visit the hospital. As a result, it is possible that the rate of disease in this population (25%) may be somewhat greater than the true rate in patients newly diagnosed with GH. However, it is unlikely that this bias had much influence on the development of the screening rule, since the lack of participation of non-symptomatic patients would not have affected the significance of particular symptoms. On the other hand, if this estimate of the rate of disease is significantly elevated, the positive and negative predictive values will be negatively affected (see section 1.3.2.1).

Secondly, the clinical examination was conducted by two different doctors. This may have resulted in inconsistencies in diagnosis. That is, in a given situation one of the doctors may have considered mild disease to be present, whereas the second doctor given the same patient would consider it to be absent. This kind of bias is believed to be minimal due to the standardized methodology used for diagnosis of disease, and the fact that the two clinicians involved in this evaluation have worked together for some time. Furthermore, a blinded evaluation of inter-observer variance in measurement between the two clinicians was conducted previously. It was determined that the clinician’s appraisals were within 5% of each other in all measurements.

Thirdly, all patients were referred from one endocrinologist’s office. This may have limited the study to a particular demographic of people, affecting the external validity of the test. However, the referring endocrinologist, Dr. Wilkins, is recognized as having the largest and most diverse thyroid-treatment practice in British Columbia. It is therefore believed that the patients in the cohort are reasonably representative of the population of GH patients in Vancouver.
Fourthly, there is a danger that not all patients referred were actually Graves' Hyperthyroidism patients. Once again, since Dr. Wilkins is a recognized expert in the field, it is believed that all patients referred by him suffered from clinically significant disease as determined by RAI and free T4 measurements. (See section 1.1.1) Nevertheless, since these patient records were never examined and confirmed by the study coordinator it remains a potential source of bias.

Other possible sources of random error include measurement error, errors in data entry and coding and errors in calculation.

4.3 Limitations of the Study

Sample size represents the most significant limitation to the study. The clinical decision rule presented above is based on a sample of only 49 participants. Furthermore, a mere 12 of these patients were positive for disease. While it is impressive that this small amount of data provided sufficient power to produce a clinical decision rule, caution in interpretation must nevertheless be exercised. In particular, it is quite conceivable that the rule described above is in fact a random “quirk” of the data, and that the results may not be duplicated on further experimentation. Fortunately, the essential performance of the validation study (as described in section 3.4.1) will effectively serve to confirm or deny the rule’s clinical usefulness.
4.4 Current and Future Studies

This analysis represents the first Canadian study on an inception cohort of TO. Continued studies of this population are expected to yield further information concerning the natural history of this disease which will in turn aid clinicians in effective management of the condition. At this stage, the data collected in this analysis concerned with incidence of mild, moderate and severe disease, gender ratio, symptomology, stress level, onset and smoking behaviour are all of major interest to clinicians practicing in this field. (Section 3.2), and have a good deal of clinical relevance.

Current studies being conducted involve a natural history and quality of life study of patients with TO. Patients who took part in this study are being asked to return at regular intervals for follow-up evaluations by clinicians. As explained in section 4.3, it is hoped that some useful information concerning the natural history and prognosis of disease will be determined. For example, it is possible that particular symptoms or signs of the disease may be related to the development of mild, moderate or severe conditions.

Two quality of life studies are being conducted with two different study groups. The first is comprised of members of the inception cohort, where patients complete a copy of the Medical Outcome Studies’ SF-36 quality of life form. Ultimately, this group will yield two further groups; those with TO and GH and those with just GH. This will allow the investigator to measure the impact of TO on quality of life in a case-control analysis. The second group under analysis is comprised of new patients referred to the UBC VGH Thyroid Orbitopathy Clinic. Participating patients complete the SF-36 on initial visit to the Clinic, allowing a cross-sectional analysis of quality of life of patients suffering from TO.

Future studies include the validation study described in section 3.4.1 above as well as studies exploring risk factors for TO other than symptoms. This would include further examination of factor
such as age, gender, onset speed of disease, stress level and genetic predisposition to TO. Furthermore, if the Vancouver Rule proves a valid clinical tool, patients may be captured at an earlier stage of disease, and studies examining potential new treatments that may help prevent development of severe TO may be initiated.
5.0: References


22. Frueh B. Why the NOSPECS classification of Graves' eye disease should be abandoned with suggestions for the characterization of this disease. Thyroid 1992; 2:85-88.


Appendix 1
Timeline

Sept 15, 1998  Questionnaire development phase. Clinician interviews, patient focus groups and literature search conducted. Ethics approval gained.

Jan 19, 1999:  Patient recruitment begins. Informed consent gained for study. 8 week follow-up appointment made.

Apr 1, 1999-  Apr 31, 2000:  Data collection from cohort.

May 1, 2000-  June 1, 2000:  Final follow up exam recorded. Final data examined.
University of British Columbia Thyroid Patient Assessment Study

THYROID ORBITOPATHY INDICATOR

Patient Name: ____________________________

Gender:  Female □  Male □

Date: ____________

Birth Date (DD/MM/YY): ____________

1. Were you born in Canada?

Yes □  No □

Where? ____________________________

2. Have you or any of your blood relations (e.g. siblings/parents/aunts/uncles/children) ever been diagnosed with a thyroid disease?

<table>
<thead>
<tr>
<th></th>
<th>Me</th>
<th>Family Member</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overactive thyroid gland?</td>
<td>Yes □ No □ Unsure □</td>
<td>Yes □ No □ Unsure □</td>
</tr>
<tr>
<td>Underactive thyroid gland?</td>
<td>Yes □ No □ Unsure □</td>
<td>Yes □ No □ Unsure □</td>
</tr>
<tr>
<td>Nodule or enlarged thyroid gland?</td>
<td>Yes □ No □ Unsure □</td>
<td>Yes □ No □ Unsure □</td>
</tr>
<tr>
<td>Thyroid Cancer?</td>
<td>Yes □ No □ Unsure □</td>
<td>Yes □ No □ Unsure □</td>
</tr>
<tr>
<td>Other</td>
<td>Yes □ No □ Unsure □</td>
<td>Yes □ No □ Unsure □</td>
</tr>
</tbody>
</table>

3. Do you or any of your blood relations (e.g. siblings/parents/aunts/uncles/children) suffer from any of the following diseases?

<table>
<thead>
<tr>
<th></th>
<th>Me</th>
<th>Family Member</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graves' Orbitopathy?</td>
<td>Yes □ No □ Unsure □</td>
<td>Yes □ No □ Unsure □</td>
</tr>
<tr>
<td>Myasthenia Gravis?</td>
<td>Yes □ No □ Unsure □</td>
<td>Yes □ No □ Unsure □</td>
</tr>
<tr>
<td>Diabetes Mellitus?</td>
<td>Yes □ No □ Unsure □</td>
<td>Yes □ No □ Unsure □</td>
</tr>
<tr>
<td>Other major disease?</td>
<td>Yes □ No □ Unsure □</td>
<td>Yes □ No □ Unsure □</td>
</tr>
</tbody>
</table>

4. Do you smoke?

No □  Yes □

Approximately how many cigarettes per day? ______

How long have you smoked? ______

When did you quit? ______

How long did you smoke for? ______

5. Have you ever smoked?

No □  Yes □

Approximately how many cigarettes per day? ______
University of British Columbia Thyroid Patient Assessment Study

**THYROID ORBITOPATHY INDICATOR**

Please indicate in the left column whether you are experiencing any of the following eye symptoms. If you are, please indicate in the right column how often these symptoms affect you (when applicable):

<table>
<thead>
<tr>
<th>Study #:</th>
<th></th>
</tr>
</thead>
</table>

### Exposure

6. Itching, burning or stinging sensations in your eye(s)?
   - No
   - Little
   - Moderate amount
   - A great deal
   - Constantly
   - Occasionally

7. Grittiness or sandy sensation?
   - No
   - Little
   - Moderate amount
   - A great deal
   - Constantly
   - Occasionally

8. Dryness in one or both of your eyes?
   - No
   - Little
   - Moderate amount
   - A great deal
   - Constantly
   - Occasionally

9. Eyes watering more than normal?
   - No
   - Little
   - Moderate amount
   - A great deal
   - Constantly
   - Occasionally

10. Increased sensitivity to light?
    - No
    - Little
    - Moderate amount
    - A great deal
    - Constantly
    - Occasionally

11. In general, do your eyes get tired easily?
    - No
    - Little
    - Moderate amount
    - A great deal
    - Constantly
    - Occasionally

### Inflammatory signs

12. Redness in your eyes or eyelids?
    - No
    - Little
    - Moderate amount
    - A great deal
    - Constantly
    - Occasionally

13. Pain when moving your eyes around?
    - No
    - Little
    - Moderate amount
    - A great deal
    - Constantly
    - Occasionally

14. Pain or ache behind your eyes?
    - No
    - Little
    - Moderate amount
    - A great deal
    - Constantly
    - Occasionally

15. Swelling or feeling of fullness in one or both of your upper eyelids?
    - No
    - Little
    - Moderate amount
    - A great deal
    - Constantly
    - Occasionally

16. Swelling or feeling of fullness in one or both of your lower eyelids?
    - No
    - Little
    - Moderate amount
    - A great deal
    - Constantly
    - Occasionally

17. Bags under the eyes?
    - No
    - Little
    - Moderate amount
    - A great deal
    - Constantly
    - Occasionally

18. Have you noticed a jelly-like swelling on the surface of your eye?
    - No
    - Little
    - Moderate amount
    - A great deal
    - Constantly
    - Occasionally

### Proptosis

19. Do you feel pressure behind your eyes?
    - No
    - Little
    - Moderate amount
    - A great deal
    - Constantly
    - Occasionally

20. Have your eyelashes or eyelids begun to touch your glasses/moved outward?
    - No
    - Little
    - Moderate amount
    - A great deal
    - Constantly
    - Occasionally

21. Do your eyes *feel* as if they are being pushed outwards?
    - No
    - Little
    - Moderate amount
    - A great deal
    - Constantly
    - Occasionally

22. Do your eyes *look* as if they are protruding/sticking out?
    - No
    - Little
    - Moderate amount
    - A great deal
    - Constantly
    - Occasionally
## University of British Columbia Thyroid Patient Assessment Study

### THYROID ORBITOPATHY INDICATOR

<table>
<thead>
<tr>
<th>Study #:____</th>
</tr>
</thead>
</table>

23. Is the coloured part of your eye off-centre? (too much white showing?)

<table>
<thead>
<tr>
<th>No</th>
<th>a little</th>
<th>a moderate amount</th>
<th>a great deal</th>
</tr>
</thead>
<tbody>
<tr>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
</tbody>
</table>

24. Do your eyes seem to be open too wide?

| □  | □        | □                 | □           |

## Vision effects

25. Has there been a deterioration in your vision recently?

| □  | □        | □                 | □           |

26. Is your vision blurry (even with glasses/contacts?)

| □  | □        | □                 | □           |

27. Are colours fading or looking grey?

| □  | □        | □                 | □           |

28. Do you have any transient grey-outs of vision?

| □  | □        | □                 | □           |

## Strabismus

29. Do you find yourself tilting your head instead of moving your eyes in order to see things?

| □  | □        | □                 | □           |

30. Do you feel resistance when moving your eyes around?

| □  | □        | □                 | □           |

31. Do you feel a pulling sensation when moving your eyes around?

| □  | □        | □                 | □           |

32. Do you have trouble focusing? (even with glasses/contacts?)

| □  | □        | □                 | □           |

33. Do your eyes tire quickly with near tasks like reading or working on a computer?

| □  | □        | □                 | □           |

34. Do you have trouble tracking objects? (Following objects with your eyes?)

| □  | □        | □                 | □           |

35. Do you suffer from double vision?

| No (go to 40) | Yes (go to 36) |

36. Is it constant or occasional?

| Constant (go to 39) | Occasional (go to 37) |

37. If occasional, how often does it happen?

| Less than 1/day (1-5 times/day) | More than 5/day (go to 38) |

38. If occasional, how long does it last?

| Momentary (1-10 minutes) | More than 10 mins |

39. Is it worse at certain angles?

| No | Yes |

86
University of British Columbia Thyroid Patient Assessment Study

THYROID ORBITOPATHY INDICATOR

STUDY #:____

Stress

40. Have you experienced any stressful life events in the past year?
   (eg. death of a spouse, loss of job)
   No [ ] Yes [ ] (Please describe) [ ]
   None of the time [ ] A little of the time [ ] A good bit of the time [ ] All of the time [ ]

41. Do you feel stressed? [ ] [ ] [ ] [ ]
42. Are you experiencing any emotional, financial or work-related stress? [ ] [ ] [ ] [ ]

Onset time

43. How fast did the symptoms of hyperthyroidism come on? [ ] [ ] [ ] [ ]
   Gradual (4 weeks or more) [ ] Moderate (1-4 weeks) [ ] Rapid (7 days or less) [ ] Not Applicable [ ]
44. How fast did your eye symptoms develop? [ ] [ ] [ ] [ ]
45. What date did you first notice your eye symptoms? (dd/mm/yy) ____________________ [ ]
46. Is there anything else you would like to add? [ ]
Appendix 3
HYPERTHYROIDISM AND THYROID EYE DISEASE:
AN INCEPTION COHORT ANALYSIS

A small percentage of individuals diagnosed with hyperthyroidism will go on to
develop a condition called thyroid eye disease. In this disease, the eye muscles are attacked
by the immune system, leading to inflammation and swelling around the eye area which
can disrupt vision and cause changes in facial appearance. The disease, therefore, has a
significant impact on the health and happiness of the patients who develop it. Fortunately,
if it is diagnosed early, the severity and extent of the disease may be controlled by
clinicians. You have been invited to participate in this study because you have been
diagnosed with hyperthyroidism, and are therefore at increased risk of developing thyroid
orbitopathy.

The objective of this research is to follow the development of thyroid eye disease in
patients with hyperthyroidism, particularly focussing on early signs and symptoms, in an
effort to better control the course of the disease. Patients over 16 years of age who are
willing to participate will be asked to return for a clinical assessment at 8 weeks, 24 weeks,
48 weeks and 72 weeks after recruitment. At each of these follow-up visits, the patient will
be examined by a clinician, and asked to fill out a symptomology questionnaire. Each
appointment should take no longer than 20 minutes. In cases where a patient shows signs
that s/he is developing the disease, clinicians will begin treating the patient immediately,
and may ask the patient to return for more frequent follow-up appointments.

At all times in this research, patient confidentiality will be strictly maintained
through the assignment of a study number instead of the use of names. Data concerning
the study will be analyzed and maintained solely by the research coordinator on a secure
database.

If you would like to participate, please complete and return the following to the front
desk: (Please Print)

Name: ___________________________ Date: ______________

Telephone: (Home) (____)__________ (Work) (____)__________
HYPERTHYROIDISM AND THYROID EYE DISEASE: AN INCEPTION COHORT ANALYSIS

Principal Investigators: Dr. Jack Rootman, Dr. Peter Dolman
Department of Ophthalmology
Faculty of Medicine
(604) 875-4199

A small percentage of individuals diagnosed with hyperthyroidism will go on to develop a condition called thyroid eye disease. In this disease, the eye muscles are attacked by the immune system, leading to inflammation and swelling around the eye area which can disrupt vision and cause changes in facial appearance. The disease, therefore, has a significant impact on the health and happiness of the patients who develop it. Fortunately, if it is diagnosed early, the severity and extent of the disease may be controlled by clinicians. You have been invited to participate in this study because you have been diagnosed with hyperthyroidism, and are therefore at increased risk of developing thyroid eye disease.

The objective of this research is to follow the development of thyroid eye disease in patients with hyperthyroidism, particularly focussing on early signs and symptoms, in an effort to better control the course of the disease. Patients over 16 years of age who are willing to participate will be asked to return for a clinical assessment at 8 weeks, 24 weeks, 48 weeks and 72 weeks after recruitment. At each of these follow-up visits, the patient will be examined by a clinician, and asked to fill out a symptomology questionnaire. Each appointment should take no longer than 20 minutes. In cases where a patient shows signs that s/he is developing the disease, clinicians will begin treating the patient immediately, and may ask the patient to return for more frequent follow-up appointments.

One aspect of the clinical examination is the use of Proparacaine 0.5% eye drops to numb the surface of the eye so that ocular pressure may be measured. Another part of the exam involves Computer-assisted Tomography (CT) scans to take pictures of the eye muscles to help with the diagnosis of the condition. These are both standard procedures in the diagnosis and treatment of a number of different eye conditions, and are not exclusive to this study.
At all times in this research, patient confidentiality will be strictly maintained through the assignment of a study number instead of the use of names. Data concerning the study will be analyzed and maintained solely by the research coordinator on a secure database.

If there are any questions or concerns over the course of the study, they may be addressed to the Thyroid Orbitopathy Research Coordinator, Mark Linder, at 875-4111 ext. 62880. Questions concerning patient rights or treatment as a research subject during the study may be addressed to Dr. R. D. Spratley, Director, Office of Research Services at 822-8598. Patients declining to participate in the study, or wishing to withdraw at a later date, may do so at any time without any consequences to their continuing medical care.

I have received a copy of this consent for my own records.

I agree to participate in this study.

Patient Signature ___________________________ Date ____________

Witness Signature ___________________________ Date ____________

Investigator's Signature ___________________________ Date ____________
Appendix 5
**UBC THYROID ORBITOPATHY CLINIC - 1st VISIT**

**DEMOGRAPHICS**

<table>
<thead>
<tr>
<th>Patient Name</th>
<th>Sex</th>
<th>Birthdate</th>
<th>Race</th>
<th>Occupation</th>
<th>Age</th>
</tr>
</thead>
</table>

**DISEASE ONSET**

<table>
<thead>
<tr>
<th>Symptoms</th>
<th>Orbit</th>
<th>Thyroid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Date of Onset:</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate of Onset:</td>
<td>Acute (days) Subacute (wks)</td>
<td>A/S/C</td>
</tr>
<tr>
<td>Chronic (mos)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Progress:</td>
<td>Same, Better, Worse</td>
<td>s/b/w</td>
</tr>
<tr>
<td>Tests:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment:</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**MEDICAL HISTORY**

<table>
<thead>
<tr>
<th>Allergies</th>
<th>Medical Hx</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medications</td>
<td>Surgical Hx</td>
</tr>
<tr>
<td>Smoking</td>
<td>Family Hx</td>
</tr>
</tbody>
</table>

**Subjective**

<table>
<thead>
<tr>
<th>Objective</th>
<th>OD</th>
<th>OS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vision: n/abn</td>
<td>Central Vision sc/cc 20/</td>
<td>20/</td>
</tr>
<tr>
<td>cM 20/</td>
<td>20/</td>
<td></td>
</tr>
<tr>
<td>Color vis: n/abn</td>
<td>Color Vision errors (AO) y/n</td>
<td>y/n</td>
</tr>
<tr>
<td>Pupils (aff defect)</td>
<td>normal &lt; 4</td>
<td></td>
</tr>
</tbody>
</table>

**INFLAMMATORY**

<table>
<thead>
<tr>
<th>Orbital pain</th>
<th>Chemosis (0-3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>at rest: y/n</td>
<td>Conjunctival injection (0-1)</td>
</tr>
<tr>
<td>with gaze: y/n</td>
<td>Lid injection (0-1)</td>
</tr>
<tr>
<td>Lid edema: y/n</td>
<td>Lid edema [upper] (0-3)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lid edema [lower] (0-3)</th>
<th>Progress: s/b/w</th>
</tr>
</thead>
</table>

**STRABISMUS / MOTILITY**

<table>
<thead>
<tr>
<th>Diplopia: none intermittent with gaze constant</th>
<th>Ductions (degrees)</th>
</tr>
</thead>
</table>

**Strabismus: y/n**

Prism measurement:

Inflammatory Index (worse)

- Chemosis (0-3):
- Conj (0-1):
- Lid injec (0-1):
- Lid edema (0-3):
- Rest pain (0-1):
- Mov't pain (0-1):

TOTAL (10):
<table>
<thead>
<tr>
<th>Subjective</th>
<th>Objective</th>
<th>OD</th>
<th>OS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>APPEARANCE / EXPOSURE</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lid retraction y/n</td>
<td>Fat prolapse</td>
<td>y/n</td>
<td>y/n</td>
</tr>
<tr>
<td></td>
<td>Interpalpebral fissure</td>
<td>mm</td>
<td>mm</td>
</tr>
<tr>
<td></td>
<td>Lid retractn (upper) MRD-4 (upper scleral show)</td>
<td>mm</td>
<td>mm</td>
</tr>
<tr>
<td></td>
<td>(lower scleral show)</td>
<td>mm</td>
<td>mm</td>
</tr>
<tr>
<td></td>
<td>Levator function</td>
<td>mm</td>
<td>mm</td>
</tr>
<tr>
<td></td>
<td>Lagophthalmos</td>
<td>mm</td>
<td>mm</td>
</tr>
<tr>
<td>Proptosis y/n</td>
<td>Exophthalmometry</td>
<td>mm</td>
<td>mm</td>
</tr>
<tr>
<td>Tearing y/n</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FB sensation y/n</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Progress: s/b/w</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>BIOMICROSCOPY</strong></td>
<td>Corneal erosions y/n</td>
<td>y/n</td>
<td>y/n</td>
</tr>
<tr>
<td></td>
<td>IOP - Straight mmHg</td>
<td>mmHg</td>
<td>mmHg</td>
</tr>
<tr>
<td></td>
<td>Up mmHg</td>
<td>mmHg</td>
<td>mmHg</td>
</tr>
<tr>
<td></td>
<td>Down (if ↑ in straight) mmHg</td>
<td>mmHg</td>
<td>mmHg</td>
</tr>
<tr>
<td></td>
<td>Other y/n</td>
<td>y/n</td>
<td>y/n</td>
</tr>
<tr>
<td><strong>FUNDUS</strong></td>
<td>Spont. venous pulsation y/n</td>
<td>y/n</td>
<td>y/n</td>
</tr>
<tr>
<td></td>
<td>Optic nerve Edema y/n</td>
<td>y/n</td>
<td>y/n</td>
</tr>
<tr>
<td></td>
<td>Pallor y/n</td>
<td>y/n</td>
<td>y/n</td>
</tr>
<tr>
<td></td>
<td>Choroidal folds y/n</td>
<td>y/n</td>
<td>y/n</td>
</tr>
</tbody>
</table>

**INVESTIGATIONS**
Survey Visual fields
Motility studies VEP
Photos Color vision tests
Ultrasound Serology
CT Scan Other

**DIAGNOSIS**

**GRADING / PROGRESS**

<table>
<thead>
<tr>
<th>Vision:</th>
<th>Grade</th>
<th>Progress</th>
<th>Rate of Onset:</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal / abnormal</td>
<td>s / b / w</td>
<td>A / S / C</td>
<td></td>
</tr>
<tr>
<td>Inflammatory Activity Score:</td>
<td>0 - 10: mild, mod, severe</td>
<td>s / b / w</td>
<td>A / S / C</td>
</tr>
<tr>
<td>absent, intermittent, constant</td>
<td>s / b / w</td>
<td>A / S / C</td>
<td></td>
</tr>
<tr>
<td>Appearance/Exposure:</td>
<td>normal, mild, mod, severe</td>
<td>s / b / w</td>
<td>A / S / C</td>
</tr>
</tbody>
</table>

**MANAGEMENT**
Observation
Conservative: lubricants, cool compresses
Medical: non-steroidal
oral steroid
pulse steroid
immunosuppressives
Radiotherapy
Surgery

**RESEARCH STUDY**

**FOLLOW - UP**
Appendix 6
### Table of p-values, Odds Ratios and Confidence Intervals for all symptom questions on the TOI

<table>
<thead>
<tr>
<th>Item #</th>
<th>p-value</th>
<th>Odds Ratio</th>
<th>95 % C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0.470</td>
<td>1.32</td>
<td>0.35</td>
</tr>
<tr>
<td>7</td>
<td>0.190</td>
<td>2.58</td>
<td>0.59</td>
</tr>
<tr>
<td>8</td>
<td>0.244</td>
<td>2.00</td>
<td>0.53</td>
</tr>
<tr>
<td>9</td>
<td>0.057</td>
<td>3.69</td>
<td>0.93</td>
</tr>
<tr>
<td>10</td>
<td>0.302</td>
<td>1.77</td>
<td>0.47</td>
</tr>
<tr>
<td>11</td>
<td>0.437</td>
<td>1.43</td>
<td>0.36</td>
</tr>
<tr>
<td>12</td>
<td>0.053</td>
<td>3.78</td>
<td>0.97</td>
</tr>
<tr>
<td>13</td>
<td>0.555</td>
<td>1.28</td>
<td>0.21</td>
</tr>
<tr>
<td>14</td>
<td>0.620</td>
<td>1.04</td>
<td>0.23</td>
</tr>
<tr>
<td>15</td>
<td>0.036</td>
<td>4.36</td>
<td>1.11</td>
</tr>
<tr>
<td>16</td>
<td>0.490</td>
<td>1.36</td>
<td>0.29</td>
</tr>
<tr>
<td>17</td>
<td>0.007</td>
<td>8.21</td>
<td>1.57</td>
</tr>
<tr>
<td>18</td>
<td>0.164</td>
<td>3.56</td>
<td>0.61</td>
</tr>
<tr>
<td>19</td>
<td>0.511</td>
<td>0.76</td>
<td>0.17</td>
</tr>
<tr>
<td>20</td>
<td>0.527</td>
<td>0.56</td>
<td>0.06</td>
</tr>
<tr>
<td>21</td>
<td>0.620</td>
<td>1.04</td>
<td>0.23</td>
</tr>
<tr>
<td>22</td>
<td>0.468</td>
<td>1.35</td>
<td>0.33</td>
</tr>
<tr>
<td>23</td>
<td>0.587</td>
<td>1.55</td>
<td>0.13</td>
</tr>
<tr>
<td>24</td>
<td>0.015</td>
<td>8.10</td>
<td>1.56</td>
</tr>
<tr>
<td>25</td>
<td>0.175</td>
<td>2.36</td>
<td>0.62</td>
</tr>
<tr>
<td>26</td>
<td>0.023</td>
<td>5.08</td>
<td>1.27</td>
</tr>
<tr>
<td>27</td>
<td>0.312</td>
<td>undefined</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>0.750</td>
<td>undefined</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>0.665</td>
<td>0.95</td>
<td>0.17</td>
</tr>
<tr>
<td>30</td>
<td>0.445</td>
<td>0.47</td>
<td>0.05</td>
</tr>
<tr>
<td>31</td>
<td>0.617</td>
<td>0.86</td>
<td>0.15</td>
</tr>
<tr>
<td>32</td>
<td>0.401</td>
<td>1.49</td>
<td>0.39</td>
</tr>
<tr>
<td>33</td>
<td>0.279</td>
<td>1.89</td>
<td>0.49</td>
</tr>
<tr>
<td>34</td>
<td>0.534</td>
<td>0.73</td>
<td>0.13</td>
</tr>
</tbody>
</table>