INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

Université d'Ottawa · University of Ottawa

A PILOT STUDY OF THE BEDSIDE TESTS FOR ASPIRATION IN ACUTE STROKE

by

© HENRY JAY BIEM MD

Thesis submitted to
the School of Graduate Studies and Research
in partial fulfilment of the requirements for the
MSc degree in Epidemiology

University of Ottawa

December, 1997

National Library of Canada
Acquisitions and Bibliographic Services
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada
Acquisitions et services bibliographiques
395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

0-612-28398-4

Canada

Acknowledgements

I would like to express my gratitude to my thesis supervisors, Rama Nair PhD and Andreas Laupacis MD, for their instruction, guidance and encouragement. I would like also to express my appreciation to Sally Lees MS, Margaret Harrison RN, Carl VanWalraven MD, Peter Watson MD, Henry B. Biem MD, and Sandra R. D. Biem MD for their assistance in developing and carrying out this pilot study, and to N. Birkett MD for detailed critique of and helpful suggestions on my attempt at sample size estimation for reliability studies. Finally, I would like to acknowledge the kind cooperation of nursing staff, medical staff, and patients with stroke at Etobicoke General Hospital.

Dedication

This thesis is dedicated to my wife Sandra.

Abstract

Background:

Stroke is a leading cause of death, admission to hospital and need for long-term care. Important complications of stroke are disordered swallowing and airway control, which may lead to dysphagia and aspiration. Previous investigators have found associations between aspiration on modified barium swallow videofluoroscopy (the current gold standard test) and various bedside clinical observations including: abnormal voice quality, abnormal voluntary cough quality, absent gag reflex, decreased pharyngeal sensation, and difficulty swallowing water in various amounts. Used alone, none of these clinical observations has been sufficiently accurate to predict aspiration on videofluoroscopy or clinical outcomes such as pneumonia and disruption of oral feeding. These bedside tests have not been examined for reliability. For dichotomous tests, the Kappa statistic represents agreement beyond that expected based on chance. Estimates of the proportions of observed (P_o) and chance expected (P_e) agreement can be used to determine the sample size required to estimate Kappa within a given confidence interval.

Aims and Objectives:

The aim of this thesis was to provide feasibility and statistical information that could be used to develop a grant proposal for a multidisciplinary study of the reliability and accuracy of the bedside diagnosis of aspiration and dysphagia in

acute stroke patients. The objectives of this pilot study were:

1. To determine the feasibility of doing simple bedside tests for dysphagia and aspiration and determining intra-observer and inter-observer reliability.
2. To determine the required sample sizes to measure the inter-observer reliability of each of the video-recorded bedside tests with the Kappa statistic within a desired confidence interval based on estimates of the observed and expected agreements.
3. To determine the rate of admission of acute stroke patients at one hospital center and the approximate proportion of acute stroke patients who are eligible for enrollment and give consent.
4. To determine the approximate proportion of recruited acute stroke patients with the clinical outcomes of pneumonia and disruption of oral feeding.

Design and Methods:

The pilot study used a prospective design to determine the agreement between two blinded observers for each of the various bedside tests recorded on videotape. It was conducted at Etobicoke General, a suburban community hospital.

Between 48 hours and 96 hours after the onset of stroke, alert patients were examined by the researcher daily for two days for voice quality, voluntary cough quality, pharyngeal sensation and gag reflex. The ability to swallow a teaspoonful of ice chips as well as 50 ml, 100 ml and 150 ml of water was also recorded. The researcher made a video-recording of the brief examination

with a standardized protocol. Patients were followed for one month. The patient records were reviewed for any note of symptoms and signs of pneumonia (productive cough, shortness of breath, fever, new crackles), feeding disruption (enteral tube feeding) or feeding difficulty (choking or coughing with feeding).

The video-recordings were shown to two physician observers blinded to the baseline clinical data and to each other's ratings. The recordings were viewed in random order on two occasions separated by one week. Two by two tables were created to measure intra-observer and inter-observer reliability.

A pilot sample size of 30 subjects was chosen to obtain estimates of observed (P_o) and chance expected (P_e) agreements.

Results:

It took ten months to enroll the 30 acute stroke patients. Approximately one third of patients admitted with a diagnosis of stroke were non-acute (with onset greater than 48 hours before admission). Many patients could not give consent themselves and next-of-kin were difficult to contact. Only one of the 30 patients developed probable pneumonia in the one month follow-up and none had disruption of oral feeding or feeding difficulty. The review of the videotape by the observers took approximately ten hours. For the bedside tests of voice quality, voluntary cough quality, pharyngeal sensation and ice water swallow, the required sample sizes to estimate Kappa within a 95% confidence interval of ± .10 were greater than 200 subjects. The

minimum required sample sizes were less but still large for gag reflex (171) and water swallow (171). Sample sizes were much less to estimate Kappa within a 95% confidence interval of ± .20 for voice quality (138), cough quality (96), gag reflex (64), and water swallow (64).

Conclusions:

The recruitment rate was lower than originally projected, based on prior admission data, because many patients admitted with a diagnosis of stroke were non-acute and because of difficulty in obtaining consent. The baseline rates of outcomes such as pneumonia and disruption of oral feeding appeared to be low.

Although it likely gave an optimistic estimate of observer agreement, the video-recording of the bedside tests was technically feasible for preliminary reliability testing. A simplified sample size table listing plausible values for observed and chance expected agreement showed the trade-off between the sample size required and size of the confidence interval around Kappa.

The pilot study led to important refinements of the protocol which were incorporated into a detailed study proposal. This proposal called for a sample size of 200 patients recognizing that this study would require at least three centers and a duration of at least two years. The pilot study supported the feasibility of a more definitive study of bedside diagnosis of aspiration in acute stroke.

Table of Contents

List of Tables

List of Figures

1. Introduction

Stroke is an important cause of mortality and morbidity. It is the third leading cause of death and the leading cause of disability in adults.[1,2] Important and frequent complications of stroke are disordered swallowing and airway control, which may lead to dysphagia and aspiration.[3] Dysphagia, broadly defined as difficulty swallowing ingested liquids and solids, may precipitate aspiration and disrupt oral feeding. Aspiration, defined as entry of substances (ingested solids and liquids, gastric contents, and oral secretions) into the lower respiratory tract below the true vocal cords, may lead to respiratory distress and pneumonia.[4] Spurr and Valencia[5] found 7% (31/469) of acute stroke patients admitted to a neurology ward developed pneumonia within a follow-up of 18 months. Vitanen et al.[6] studied causes of death at six months in stroke patients and found pneumonia to be the third most common.

A recent clinical practice guideline by the United States Public Health Service recommended that acute stroke patients have their swallowing assessed before oral intake is resumed. However, this was based more on strong panel consensus than on research evidence.[7] The diagnostic strategy, including the appropriate use of radiologic studies, and the efficient involvement of speech-language pathologists, depends on bedside assessment of stroke patients.[8,9]

Many bedside tests have been suggested to screen for dysphagia and aspiration in acute stroke patients. They include tests of voice quality, voluntary cough quality, gag reflex,[3,18] pharyngeal sensation and swallow [10,11] The current gold standard test is modified barium swallow videofluoroscopy in which any barium entry into the lower respiratory tract is considered abnormal.[4,12] Because of concerns about the accuracy of bedside tests, liberal use of modified barium swallow videofluoroscopy for screening in stroke patients has been advocated.[3,11] However, aspiration on modified barium swallow video-fluoroscopy is simply a surrogate for the clinical outcomes of aspiration (with either pneumonia or respiratory distress) and disruption of oral feeding (which may lead to malnutrition and dehydration). Some investigators have also questioned the value of video-fluoroscopy in screening acute stroke patients.[9] Modified barium swallow videofluoroscopy is time-consuming, expensive and not universally available for screening all stroke patients. The detection of patients at risk of dysphagia and aspiration in acute stroke currently relies on bedside clinical observations or tests. Some authors have recommended specific bedside screening tests to decide which patients are safe to feed orally.[13]

For a test to be useful, it should be both accurate (valid) and reliable (precise). For bedside tests with dichotomous results, reliability can be assessed with the Kappa statistic, which can be estimated within a confidence interval depending on the sample size. Based on guidelines for the interpretation of

Kappa suggested by Landis and Koch,[14] unreliable tests can be identified and either modified or discarded. Reliable tests merit further study for accuracy (validity) and clinical utility.

2. Background

2.1 Data Sources

A search of published studies was undertaken to find articles about dysphagia, aspiration, pneumonia, and disruption of oral feeding in stroke patients, and specifically articles about the reliability and accuracy of tests for dysphagia and aspiration in stroke patients. A *MEDLINE* search of articles published from 1966 to December 1996 was done. Additional articles were sought from the references of the retrieved articles, a bibliography of swallowing supplied by the Department of Speech-language Pathology at the Ottawa Civic Hospital, and standard textbooks of medicine and physical diagnosis. All retrieved articles were examined to determine whether they met the basic criterion of either 1) a comparison for accuracy between a clinical observation and a gold standard (either modified barium swallow videofluoroscopy or clinical outcome such as pneumonia or malnutrition), or 2) a comparison for reliability between two different observers. Other critical appraisal criteria used were: blinding of the comparison, inclusion of an appropriate spectrum of stroke patients, description of the referral pattern, description of the test in adequate detail and consideration of the clinical utility of the test. Studies that used an eventual clinical outcome were examined to see if there

was: assembly of an inception cohort, collection of baseline features and outcome criteria (preferably reproducible) and adequate follow-up (proportion lost to follow-up was not considered sufficient to compromise validity of results).[15,16]

For the retrieval of articles on stroke patients with dysphagia, aspiration, or pneumonia, the search strategy and key words used were: "cerebrovascular disorders" and ["deglutition disorders"[a] or "aspiration" or "pneumonia"][b]. This search (designated A) yielded 278 articles of which only 36 were original investigations of more than one human subject. The remaining articles were reviews, editorials, letters or case reports. There were no meta-analyses or systematic overviews on the diagnosis of aspiration and dysphagia in acute stroke patients. Fourteen articles compared a clinical observation of swallowing with modified barium swallow videofluoroscopy or an eventual clinical outcome.

For articles on stroke and disruption of oral feeding, the search strategy and key words used were: "cerebrovascular disorders" and ["nutrition disorders" or "enteral nutrition" or "gastrostomy" or "gastrointestinal intubation" or "gastrointestinal endoscopy" or "nutritional status"]. This search (designated B) yielded 91 articles of which only 25 were original investigations of more than one human subject. One

---

[a]DEGLUTITION DISORDERS was the more general MESH term for dysphagia

[b]Brackets and parentheses represent boolean logic.

additional article was found that compared a clinical observation with a gold standard (malnutrition).

In order to retrieve additional articles on the accuracy and reliability of diagnosis of aspiration and dysphagia not strictly confined to stroke patients, the search strategy and key words used were: ["aspiration" or "deglutition"] and ["diagnosis" or "bedside" or "sensitivity and specificity" or "observer variation" or "reproducibility" or "Kappa" or "reproducibility of results" or "probability" or "odds ratio"] and ["voice" or "cough" or "gagging" or "deglutition" or {"pharyngeal" and "sensation"} or {"fluoroscopy" and "endoscopy"}]. This search (designated C) yielded 53 articles. Although most of these had been found previously (in A and B), three additional articles were retrieved that studied dysphagia in general neurologic patients of whom most had stroke.

Eighteen studies permitted a comparison of a clinical observation or test with a gold standard of either modified barium swallow videofluoroscopy or eventual clinical outcome (pneumonia, disruption of oral feeding, malnutrition or mortality). These studies are summarized in Table 1. Only three of the studies stated that the comparison was done blindly. Only seven were done in consecutive acute stroke patients. The two articles by Kidd et al. reported the same group of patients. The referral pattern could only be categorized broadly by whether the study was done in either an acute or rehabilitation hospital. All articles gave more attention to a discussion of the clinical

utility of the tests, or lack thereof, than they did to the description of the tests themselves. Of the ten studies that compared clinical observations with an eventual clinical outcome, seven did this prospectively on an inception cohort. Of these, there was an adequate collection of baseline features and outcome criteria, and sufficient follow-up. None of the eighteen retrieved articles met all critical appraisal criteria and none examined reliability of the bedside clinical observations or tests.

Table 1. Summary of Studies in which a Test for Dysphagia or Aspiration was Compared with Modified Barium Swallow Videofluoroscopy or a Clinical Outcome.

Study (search)	Comparison	Blinding	Spectrum (n)	Referral
1.Horner[21] (A)	gag, voice vs. videofluoro.	unblinded	severe (47)	rehab*
2.Holas[48] (A)	videofluoro. vs. clinical (pneumonia) prospective	unblinded	severe (114)	rehab
3.DePippo[10] (A)	water swallow vs. videofluoro.	unblinded	severe (44)	rehab
4.Splaingard[34] (A)	s.l.p. vs videofluoro.	blinded	severe (107)	rehab
5.Depippo[9] (A)	Burke quest. vs. clinical (pneumonia) retrospective	unblinded*	severe (139)	rehab
6.Teasell[51] (A)	videofluoro. vs. clinical (pneumonia) retrospective	unblinded	severe (441)	rehab
7.Johnson[49] (A)	videofluoro. vs. clinical (pneumonia) retrospective	unblinded†	severe (60)	rehab
8.Gordon[18] (A)	swallow vs. clinical (pneumonia, death) prospective	unblinded	consecutive (91)	acute

* Rehab = Rehabilitation hospital
† Assumed unblinded as blinding was not mentioned in report.

Table 1. (Continued)

Study (search)	Comparison	Blinding	Spectrum (n)	Referral
9.Horner[51] (A)	gag,cough, voice vs. videofluoro.	unblinded	severe, bilateral strokes (70)	rehab
10.Kidd[3] (A)	gag, pharygeal sensation, water swallow vs. videofluoro.	unblinded	consecutive (60)	acute
11.Kidd[47] (A)	videofluoro. vs. clinical (lower respiratory infection)	unblinded	consecutive (60) (same patients as in 10.)	acute
12.Smithard[8] (A,B)	bedside assess, videofluoro. vs. clinical (pneumonia, malnutrition, mortality) prospective	unblinded (bedside vs videofluoro. blinded)	consecutive (121)	acute
13.Schmidt[52] (A)	videcfluoro. vs. clinical (pneumonia) retrospective	blinded	severe (59)	rehab
14.Barer[19] (A)	water swallow vs. clinical (mortality)	unblinded	consecutive (105)	acute
15.Finestone[15] (B)	speech lang. path. vs. clinical (malnutrition	unblinded	consecutive (49)	acute

Table 1. (Continued)

Study (search)	Comparison	Blinding	Spectrum (n)	Referral
16. Langmore[51] (C)	fiberoptic vs. videofluoro.	blinded	unspecified (21)	rehab
17. Horner[85] (C)	gag, cough vs. videofluoro.	unblinded	consecutive bilateral (38)	rehab
18. Linden[11] (C)	voice, cough, gag, swallow vs. videofluoro.	unblinded	most had stroke but other patients included (249)	rehab

2.2 Dysphagia and Aspiration in Stroke

Review of the published studies showed many definitions for dysphagia and aspiration. Logemann[4] has commented on the varied definitions. Dysphagia has been defined generally as difficulty swallowing food and liquid whereas aspiration can be one of the results of dysphagia. Dysphagia has been categorized into "oropharygneal dysphagia" (with symptoms such as difficulty in initiating a swallow and choking) and "esophageal dysphagia" (with symptoms such as heartburn and chest pressure).[17] Aspiration can occur during the swallowing of ingested substances or later with the reflux of ingested substances and secretions from the stomach into the airway. Aspiration of salivary secretions has been excluded from the definition of aspiration by some authors.[4] Many authors define aspiration simply as the subglottic penetration of food.[11] For this thesis, the term dysphagia signified oropharyngeal dysphagia rather than esophageal dysphagia, and aspiration signified aspiration of ingested substances rather than gastric contents and salivary secretions. For the published studies that follow, the working definitions used for dysphagia and aspiration have been given where possible.

Prospective studies have shown that disruption of swallowing and airway control is a frequent problem in acute stroke. Gordon et al.[18] prospectively examined 91 consecutive acute stroke patients (median age 70), most within 96 hours of admission to an

urban community hospital, and found 45% (41/91) had dysphagia, defined as the inability to drink or choking more than once while attempting to drink 50 ml of water on two occasions. Dysphagia was statistically associated with decreased level of consciousness at some point in the acute period, prolonged hospitalization (> 2 wk), and death within six weeks. Seventeen percent (7/41) of dysphagic patients developed chest infections within one week of admission compared to 8% (4/50) of patients without dysphagia. However, this difference in proportions was not statistically significant.[c] Kidd et al.[3] did modified barium swallow videofluoroscopy on 60 consecutive acute stroke patients (mean age 72) within 72 hours of admission to a university hospital. Aspiration of barium occurred in 42% (25/60) of patients. There was no control group.

The prevalence of dysphagia in acute stroke decreases over time after stroke onset. In a study investigating the use of beta-blockers in acute stroke in an university hospital setting, Barer et al.[19] estimated the frequency of dysphagia defined as choking or coughing after swallowing a mouthful of water from a cup. This swallow test was done on alert stroke patients within 24 hours of admission and within 48 hours of onset of symptoms. It was repeated one week later. Of the 105 patients (mean age 71), 29 (28%) had dysphagia at 24 hours. At one week, 13/29 (44%) had resolution of the dysphagia, 9/29 (31%) had persistent

[c] Unless otherwise specified, statistical significance is considered at the P = .05 level.

dysphagia and the remaining 7/29 (24%) were dead. Of the 105 patients, 10 had died by 1 week. Seven of these 10 had dysphagia on the initial assessment.

Disruption of swallowing may occur with strokes affecting not only the cranial nerves supplying the swallowing muscles but also higher central nervous system centres in various locations. In a study of 411 rehabilitation stroke patients by Rieve and coworkers,[20] dysphagia "confirmed by videofluoroscopy" occurred with strokes affecting a variety of neuroanatomical sites. The percentages of patients with dysphagia for each of the following anatomical types of strokes were: unilateral hemisphere 13%, bilateral hemisphere 23%, unilateral brainstem 32%, bilateral brainstem 71%, unilateral cerebellum 14%, bilateral cerebellum 100%, and multiple intracranial 18%.

Bilateral strokes are more frequently associated with dysphagia and aspiration than unilateral strokes.[21] Unilateral hemispheric strokes are associated with dysphagia especially when large.[21] Right and left unilateral strokes are associated with dysphagia and aspiration with distinct patterns. Left cortical stroke is associated with an impaired oral stage, difficulty initiating coordinated motor activity and swallowing apraxia. Right cortical stroke dysphagia is associated with more pharyngeal pooling, laryngeal penetration and pulmonary aspiration.[22]

2.3 Management of Dysphagia and Aspiration in Stroke

Swallowing assessment and therapy in stroke patients has become an important clinical activity of speech-language pathologists. A significant proportion of stroke patients are referred to the speech-language pathology department. At one centre, for example, of 160 patients admitted with stroke, 98 were referred for evaluation. Of these 98 patients, 53 had identifiable swallowing disorders.[23] In another center, stroke accounted for one quarter to one half of swallowing disorders diagnosed by speech-language pathologists on general medical wards.[24]

For patients presumed to be at high risk of aspiration and dysphagia, there are several available therapeutic options. Speech therapy, which includes teaching compensatory swallowing techniques and control of dietary consistency (e.g. thickened fluids), has become standard care for dysphagic patients although its effectiveness has not been proven in a controlled trial. An unblinded, controlled trial by DePippo et al.[25] randomized 115 patients into three graded levels of intervention of speech therapy (including compensatory swallowing techniques and control of dietary consistency). One intervention group had a single session that included advice on compensatory swallowing and dietary consistency. A group with a medium level of intervention had a session every other week. The group with highest level of intervention had daily sessions with strict supervision of

dietary consistency. A non-intervention group was considered unethical. There was no statistically significant difference in rates of pneumonia, malnutrition, or death, one year after stroke. Although the investigators concluded there was no difference in the intervention groups, the small sample size limited the power of the study.[26]

Enteral feeding is another option for the dysphagic or aspirating stroke patient. This may be done either by nasogastric tube or percutaneous gastrostomy. A recent systematic overview of published articles about tube feeding and aspiration found no randomized controlled trials comparing oral with tube feedings.[27] There has been a recent randomized trial enrolling 30 dysphagic stroke patients comparing percutaneous gastrostomy with nasogastric tube feeding. This trial found a statistically significant decrease in mortality at six weeks in patients fed by percutaneous gastrostomy. There were two deaths in 16 patients (12%) in the gastrostomy tube feeding group and 8 deaths in 14 patients (57%) in the nasogastric tube feeding group.[28]

2.4 Bedside Tests for Dysphagia and Aspiration in Stroke

2.4.1 Gag Reflex

The gag reflex is elicited by touching the posterior pharyngeal wall (pharyngeal gag reflex), soft palate (palatal gag

reflex) or the posterior tongue (lingual gag reflex) with a tongue depressor. The responses include elevation of the soft palate, "gagging" phonation and/or pharyngeal constriction. The gag reflex is highly variable even in normal individuals.[29] For example, Davies et al.[30] found a lack of pharyngeal constriction to touch of the posterior pharyngeal wall with a tongue blade in 34% (47/138) of healthy people. In acute stroke patients, Gordon et al.[18] found an abnormal gag reflex in 30% (24/81).

There have been several studies comparing the gag reflex with modified barium swallow videofluoroscopy. Horner et al.[21] reported 47 stroke patients referred for swallowing evaluation (mean age 64). Patients had both clinical and radiologic evaluation, but these were not done independently. There was an absent pharyngeal gag reflex in 60% (12/20) of patients who aspirated barium patients compared to 52% (11/21) patients who did not aspirate barium. Six patients were not tested. Unfortunately the criterion for an intact gag reflex (whether there had to be any or all of the three responses) was not given. The same investigator reported 70 patients (mean age 64) with bilateral stroke referred for swallowing evaluation.[31] Pharyngeal gag reflex was absent in 47% (16/34) of patients aspirating barium compared to 19% (7/36) of non-aspirating patients ($p < .05$). Linden et al.[11] found that an abnormal palatal gag reflex was associated with subglottic penetration of barium on modified barium swallow videofluoroscopy but an abnormal pharyngeal gag

reflex was not. These studies have led to the conclusions that an absent gag reflex does not necessarily indicate inability to swallow or control the airway[30] and that, conversely, an intact gag reflex does not assure normal swallowing and airway control.[32] Unfortunately, there has been inconsistency and inadequate detail among the studies in the descriptions of the gag reflex.

2.4.2 Pharyngeal Sensation

Kidd et al.[3] found abnormal pharyngeal sensation in 65% (39/60) of acute stroke patients. Of 25 patients with abnormal modified barium swallow videofluoroscopy, 25 patients had abnormal pharyngeal sensation giving a sensitivity of 100%. Of 35 patients with normal modified barium swallow videofluoroscopy, 21 patients had normal pharyngeal sensation giving a specificity of 60%. Pharyngeal sensation was tested by applying the tip of a tongue blade once on each side of the pharyngeal wall and asking patients to report the presence or absence of sensation. Aspiration was defined as tracheal penetration of barium that was not immediately expectorated. Unfortunately, the examinations of pharyngeal sensation and modified barium swallow videofluoroscopy were not done blindly.

2.4.3 Swallowing

Kidd et al.[3] defined dysphagia as choking, coughing, or alteration in voice quality during and after swallowing of 50 ml of water from a beaker in 5 ml aliquot. The sensitivity was 80% (20/25) and the specificity was 86% (30/35) when compared to the gold standard of tracheal penetration of barium on swallowing videofluoroscopy.

DePippo et al.[10] described the 3-oz water swallow test as a screening test for aspiration in stroke patients in a rehabilitation hospital setting. They examined 44 consecutive patients (mean age 71) admitted to a stroke rehabilitation center who had clinical features suggestive of a high risk of dysphagia as determined by their Burke screening test for dysphagia. The Burke screening test for dysphagia consists of at least one of the following clinical features: (1) bilateral hemispheric stroke, (2) brainstem stroke, (3) history of pneumonia during the acute stroke phase, (4) coughing associated with feeding, (5) failure to consume half of meals, (6) prolonged feeding time, and (7) institution of a non-oral feeding program.[11] Coughing within one minute or a post-swallow wet, hoarse voice was considered abnormal. Modified barium swallow videofluoroscopy was again the gold standard. It was recorded on videotape and later viewed by speech-language pathologists. They did not specify whether the viewing was done in a blinded manner. The sensitivity of the test was 80% (16/20) when the gold standard was defined as any

aspiration of barium. The specificity was 54% (13/24). Some authors have recommended the swallow test for screening acute stroke patients for unsafe oral feeding.[13]

2.4.4 Other Bedside Tests

Linden et al.[11] studied 249 neurological (mostly stroke) rehabilitation patients (mean age not given) referred to speech pathologists. They evaluated them with modified barium swallow videofluoroscopy and a battery of bedside clinical tests. Chi square analysis was used to determine which signs were potentially useful in predicting aspiration on videofluoroscopy. Abnormalities of voice quality (dysphonia [χ^2 = 5.97], harsh phonation [χ^2 = 5.30], breathy phonation [χ^2 = 3.98]), cough quality [χ^2 = 4.72], palatal gag reflex [χ^2 = 3.86] and laryngeal elevation [χ^2 = 6.56] were statistically associated with aspiration of barium ($p < .05$). Inability to sit upright [χ^2 = 11.5], control of secretions [χ^2 =17.8] were also statistically associated with aspiration of barium ($p < .05$). Only 2/3 of patients could be properly classified as to the presence or absence of aspiration on videofluoroscopy using a combination of the clinical predictors.

Splaingard et al.[34] evaluated bedside standardized clinical assessment (which included examination of voice quality, cough quality, oral motor function and gag reflex) by speech-language

pathologists in 107 rehabilitation patients (most of whom had stroke). The patients were categorized into six groups ranging from normal to profoundly abnormal. The investigators used videofluoroscopy as the gold standard and the comparison was blinded. In this population, the speech-language pathologist evaluation seemed to be specific at 91% (58/64) but not sensitive at 42% (18/43) for aspiration on videofluoroscopy.

DePippo et al.[9] screened 139 consecutive stroke patients (mean age 71) admitted to a rehabilitation unit with the Burke screening questionnaire (described above in 2.4.3). On follow-up for approximately three months, 12 patients developed pneumonia, recurrent upper airway obstruction or died. All 9 patients with pneumonia were identified with the questionnaire. Blinding of the investigators to the outcomes was not mentioned. The vague "non-oral feeding" criterion in the screening questionnaire can be criticized. Presumably, the choice of non-oral feeding reflects a clinical impression that oral feeding is unsafe. However, the clinical features leading to this impression are left unspecified.

Finestone et al.[35] studied 49 patients (mean age 60) admitted to a rehabilitation unit of a university hospital concentrating on the outcome of malnutrition. Malnutrition was diagnosed with the presence of any two of: decreased body weight (eg. BMI less than 20), skin fold thickness (less than 5th percentile), midarm circumference (less than 5th percentile), serum albumin (<35g/L), serum transferrin (<2.0g/L) and

lymphocyte count (<1,800/mm^3). Dysphagia was diagnosed if the patient had: choking, coughing, changes in voice quality after feeding; oral motor weakness; or swallowing difficulty. On admission to the rehabilitation unit (at a mean of 23 days after admission with acute stroke), 49% (24/49) had malnutrition, 47% (23/49) had dysphagia and 14% (7/49) had tube-feeding. Of the patients with dysphagia, 65% (15/23) had malnutrition. Of patients with tube-feeding, 85% (6/7) had malnutrition. Dysphagia and tube feeding at admission to the rehabilitation unit were statistically associated with malnutrition ($p < .05$).

2.5 Stroke Severity and Risk of Dysphagia and Aspiration

Patients with strokes that are bilateral hemispheric, severe unilateral hemispheric, brainstem, cerebellar or severe enough to cause any drop in level of consciousness early in the hospitalization, appear to be at higher risk of aspiration than patients with other types of strokes.[21] These clinical stroke types can give only a crude classification of stroke severity.

Disease-specific instruments for stroke have been reviewed recently.[16,17] Many of these neurologic stroke scales were created for specific therapeutic research trials and have undocumented reliability and validity. Some scales such as the Rankin Handicap scale, Modified Rankin Handicap Scale and the Oxford Handicap Scale have been popular for research but have been criticized.

Although they are simple and brief, they are rudimentary and lack sensitivity (i.e. responsiveness to change in disability).[38] Some of the disease-specific scales such as the NIH Stroke Scale, Canadian Neurological Scale and the European Stroke Scale have documented reliability and validity.[37] A disadvantage of the disease-specific scales is that they take longer to administer (5 minutes or more) compared to the Barthel index and Rapid Disability Rating Scale-2, each of which take only a few minutes to complete. The disease-specific scales are also generally designed to be responsive (useful in evaluation of therapy) rather than discriminative or predictive (which is key in diagnosis and prognosis).[39]

Scales of disability have been used to reflect severity of neurologic illness. The Barthel Index was developed as a tool to monitor functional disability in patients with neuromuscular disorders and has been used extensively in stroke patients.[40] It can be scored by a physician or nurse from medical records or direct observation. The index gives scores ranging from 0 to 100. Although it has limitations such as an arbitrary weighting of functions and narrow scope, it has reasonable reliability and validity in estimating prognosis. Granger et al.[41,42,43] suggested that a score of 60 or more indicates independence, a score of 40 or less, dependence and a score of 20 or less, total dependence. Kidd et al.[3] used the Barthel index as an indicator of stroke severity and found a higher risk of aspiration on videofluoroscopy in patients with a lower score. Similarly, Rieve

et al.[20] found a lower Barthel score in dysphagic versus non-dysphagic stroke rehabilitation patients.

The Linn Rapid Disability Scale is a multidimensional scale that has a wider scope than the Barthel Index and seems to have even greater reliability and validity.[40] It is best completed by a nurse or other person familiar with the patient's level of functioning rather than a physician who would be less likely to directly observe performance of the various functions.[44] The Rankin score has substantial reliability (Kappa .75) but slightly less reliability than the Barthel Index (Kappa .88) when the scales were directly compared.[45] The Katz scale, although commonly used, has considerably less evidence for validity and reliability.[40]

2.6 The Gold Standard for Dysphagia and Aspiration

The gold standard usually used for determining the accuracy of bedside tests for aspiration and dysphagia has been the modified barium swallow videofluoroscopy.[1,11,12] Follow-up studies after the gold standard test determine whether the gold standard has missed clinically significant aspiration or detected clinically unimportant aspiration.[46] A later report by Kidd et al.[47] on the same 60 acute stroke patients (from the study described on page 12) found an increased risk of lower respiratory tract infection in patients with barium aspiration on

modified barium swallow videofluoroscopy. Two weeks after the onset of stroke, 68% (17/25) of patients aspirating barium compared to 6% (2/35) of patients not aspirating barium had lower respiratory tract infection (loosely defined as productive cough and auscultatory crackles). Neither fever, leukocytosis or radiographic infiltrate were necessary for the diagnosis of lower respiratory tract infection. Unfortunately, the investigator diagnosing lower respiratory tract infection was not blinded to the results of the clinical data and modified barium swallow videofluoroscopy.

Holas et al.[49] prospectively studied 114 consecutive stroke patients (mean age 72) admitted to the Burke rehabilitation hospital who had videofluoroscopy on average one month after stroke onset. Of these 114 patients, 54% (61/114) aspirated barium on modified barium swallow videofluoroscopy. Pneumonia occurred in 13% (8/61) of aspirators compared to 2% (1/53) of non-aspirators, giving a relative risk of 7.0 (95% C.I. = 1.2 - 21, p = .03) in the 12 months of follow-up after the onset of stroke. They stated that there was no statistically significant increase in deaths, but the mortality rates were not given.

Johnson et al.[49,50] retrospectively studied patients (mean age 70) who had undergone videofluorscopy for swallowing difficulty after stroke and found that 48% (29/60) of patients had developed pneumonia in the year after the onset. They found an association between aspiration pneumonia and a variety of videofluorscopic findings including aspiration of barium, pooling of barium in the

piriform and vallecular areas of the pharynx, and a slow pharyngeal transit time.

Teasell et al.[51] retrospectively reviewed the charts of 441 consecutive patients (mean age not given) admitted to a stroke rehabilitation center at the University of Western Ontario for the four months after the onset of stroke. One hundred and six of these 441 patients underwent modified barium swallow videofluoroscopy for suspected swallowing difficulty of whom 84 had aspiration of barium. Of these 84 patients, 12% (10/84) developed pneumonia compared to 0.6% (2/357) of patients presumed to be not aspirating (i.e. asymptomatic patients who did not undergo videofluoroscopy and patients who tested negative on videofluoroscopy).

Schmidt et al.[52] did a retrospective blinded chart review of patients who had undergone modified barium swallow videofluoroscopy for suspected dysphagia at the Burke rehabilitation hospital. Twenty-six patients (mean age 73) who had aspiration of barium on videofluoroscopy were compared with thirty-three case-matched controls. Five of the 26 aspirators compared with only 1 of the 33 controls developed pneumonia after a mean follow-up of sixteen months (odds ratio 7.6, $p = .05$).

There also has been some scepticism about the predictive ability of the modified barium videofluoroscopy. Smithard et al.[6] studied 121 consecutive patients (median age 79) admitted with acute stroke. They evaluated patients with a battery of bedside tests in a "bedside assessment of swallowing" including voice

quality, cough quality, gag reflex and water swallow. They did not describe the tests in detail. Sixty patients (50%) were rated as having an abnormal bedside assessment of swallowing. Patients were followed for the outcomes of chest infections, disability, nutritional status, hospital length of stay and mortality but whether this was done in a blinded fashion was not mentioned. An abnormal bedside assessment of swallowing was statistically associated with chest infections, death, increased disability (as measured by the Barthel Index) and increased length of hospital stay. For example, 33% (20/60) of patients with an abnormal bedside assessment of swallowing had chest infections compared to 16% (9/57) of patients with a normal assessment. Patients also had modified barium swallow videofluoroscopy blinded to the results of the bedside evaluation. Videofluoroscopy did not add to the value of the bedside assessment of swallowing for the prediction of clinical outcome when used in a multivariate logistic regression analysis.

A technique recently suggested to predict risk of aspiration and dysphagia in stroke patients is fiberoptic endoscopic evaluation of swallowing (FEES). This technique uses laryngoscopy to detect aspiration of dye impregnated food and liquid. Langmore et al.[53] did a blinded comparison of fiberoptic laryngoscopy with modified barium swallow videofluoroscopy in 21 patients (9/21 had stroke, mean age 63) and reported a sensitivity of 88% and specificity of 92%. This technique has also been combined with sensory testing (FEESST). Aviv et al.[54] reported that of 18

stroke patients, 7 had normal sensation, 6 had unilateral decreased sensation and 5 had bilateral decreased sensation. Of the 5 patients with bilateral sensory loss, 2 developed aspiration pneumonia and died. None of the patients with unilateral sensory loss or normal sensation developed aspiration pneumonia. The FEESST is currently being investigated at Columbia Presbyterian Medical Center and is not widely available. It has been suggested that it may become the new gold standard for aspiration and dysphagia. However, this new technique remains to be assessed for reliability and accuracy.

In summary, the published studies support the use of the modified barium swallow as the current gold standard test. The studies show that abnormal modified barium swallow is statistically associated with pneumonia. However, many patients with an abnormal modified barium swallow remain free of pneumonia on clinical follow-up. The lack of concordance between an abnormal gold standard test and clinical outcomes such as pneumonia could be due to host resistance, protective reflexes or therapeutic interventions. On the other hand, pneumonia can occur without aspiration for a variety of reasons including poor lung defence or organism virulence. The published evidence thus supports the use of modified barium swallow as the current gold standard test for aspiration. The presence of association but not concordance between the gold standard test and clinical outcomes suggests that diagnostic studies on test validity best use both the surrogate gold standard and the eventual clinical outcomes.

2.7 Test Reliability

It is important to examine not only the accuracy but also the reliability of tests. Feinstein suggests that "hard scientific observational data" is best characterized by excellent reliability.[55] Reliability indicates the degree to which results are reproducible. It is not simply an association of two variables but a measure of concordance.[56] Under identical conditions, the unchanging subject or sample testing positive once should test positive again. There are two types of reliability: 1) *Inter-observer reliability* measures agreement between different observers testing the same subject. 2) *Intra-observer reliability* measures agreement for a single observer retesting the same subjects.

Feinstein has classified causes of unreliability into problems with the test stimulus, observer (perception and processing), environment and subjects. (See Appendix I for a discussion of the gag reflex using this framework.)

Reliability can be measured with the Kappa statistic, which expresses the agreement observed beyond that expected by chance. Guidelines have been advanced by Landis and Koch[14] to interpret Kappa. A Kappa of > .80 indicates almost perfect reliability. A Kappa of > .60 indicates substantial reliability and a Kappa of .41 -.60, moderate reliability. A Kappa .21 -.40 indicates only fair reliability and a Kappa of < .20, poor reliability. (See

Appendix II.) Although the Kappa statistic can be extended to assess reliability of polychotomous tests with the weighted Kappa, in medical decision-making a polychotomous test is frequently collapsed to a dichotomous result because the decision whether or not to treat is dichotomous.[57]

Although the reliability of the bedside tests for aspiration and dysphagia has not been examined, Ekberg et al.[46] has determined inter-observer reliability in the interpretation of modified barium swallow videofluoroscopy. Blinded radiologists with varying experience examined the videotapes of 72 patients (mean age 64) referred to the radiology suite of a university hospital. There was excellent agreement for the findings of contrast reaching trachea (Kappa = .83), substantial agreement for the findings of normal function (Kappa = .70) and moderate agreement for laryngeal penetration (Kappa = .57).

The 95% confidence limits around a Kappa value can be calculated with the standard error as derived by Fleiss and approximated by Cohen. The Cohen approximation has been used for sample size calculation[58] (Appendix II). Sample size tables can be created showing the trade-off between the sample size required and the precision desired around the Kappa estimate.

In summary, aspiration and dysphagia are frequent problem in acute stroke. While a variety of bedside tests have been suggested to predict aspiration and dysphagia after stroke, the reliability of these bedside tests remain to be determined.

3. Objectives

The aim of this thesis was to provide statistical and feasibility information that could be used to develop a grant proposal for a large multidisciplinary study of the reliability and accuracy of the diagnosis of aspiration and dysphagia in acute stroke patients. The objectives of this pilot study were as follows:

1. To determine the feasibility of video-recording simple bedside tests for dysphagia and aspiration in acute stroke patients for the measurement of intra-observer and inter-observer reliability.

2. To determine the required sample sizes to measure inter-observer reliability of each of the video-recorded bedside tests with the Kappa statistic within a desired confidence interval based on the estimates of the observed and expected agreements.

3. To determine the admission rate of acute stroke patients at one hospital and the approximate proportion of acute stroke patients who are eligible for enrollment and give consent.

4. To determine the approximate proportion of eligible acute stroke patients with the clinical outcomes of pneumonia and disruption of oral feeding.

4. Methods

4.1 Design

This pilot study prospectively enrolled acute stroke patients to determine the agreement between two blinded observers for each of the various bedside tests.

4.2 Setting

The pilot study was conducted on the medical wards at Etobicoke General Hospital, an acute care community hospital located in suburban Toronto. It is a non-teaching hospital with approximately 280 beds of which approximately 70 are designated for adult medical admissions. The protocol was accepted by the Etobicoke General Hospital ethics review board and the department of medicine with the stipulation that it not add significantly to the already heavy demands on the staff. For example, although it was initially planned to have nursing staff complete the Barthel Index and Rapid Disability Rating scale, this had to be abandoned.

4.3 Candidate Bedside Tests for Pilot Study

The study by Linden et al.[11] provided a rationale for the clinical observations of gag reflex, voice quality, and cough quality to predict aspiration. The report by DePippo et al.[10] suggested the swallow test as a predictor of aspiration. A test of swallowing of crushed ice before videoflouroscopy was recommended by Groher but has not been evaluated.[9] Because authors have used various amounts of water for the swallow test, a graduated test with increasing amounts of water (teaspoon of ice chips, 50 ml, 100 ml, 150 ml) was chosen for this pilot study. The candidate bedside tests seemed to the researcher to be clinically sensible in that they tested the swallowing and airway control mechanism. They also seemed to be clinically practical in that they were simple, easy and quick to do. Tests that are not clinically sensible or practical are unlikely to be used by clinicians.[6]

4.4 Protocol Overview

Between 48 hours and 96 hours[d] of the onset of stroke, alert

---

[d]This timing was chosen because cerebral edema caused by large infarctions is maximal between 24 to 72 hours. For patients who have been kept from food or drink by mouth (NPO) because of drowsiness caused by cerebral edema, a decision needs to be made regarding feeding and hydration by the end of the first few days beyond which time malnutrition and dehydration may develop.

patients were examined by the researcher daily for two days for the voice quality, voluntary cough quality, pharyngeal sensation and gag reflex. The ability to swallow a teaspoonful of ice chips, 50 ml, 100 ml and 150 ml of water was also recorded. A video-recording of the brief examination was made by the researcher. Video-recording the bedside tests avoided the difficulty of getting two physicians to examine the same subject at the same time. These video-recordings could be reviewed at a later date to determine observer agreement.

The baseline clinical data were documented by review of the hospital record after the bedside tests on the second day. The nursing staff caring for the patient, blinded to the results of the clinical bedside tests described above, were asked on day two for an impression about whether the patient had aspiration or dysphagia with the current oral intake. Decisions about feeding remained with the medical and nursing staff, who were blinded to the bedside test results.

Patients were followed by the researcher for one month after the onset of the stroke. The hospital records were reviewed until the date of discharge or one month after admission for any note of symptoms and signs of pneumonia (productive cough, shortness of breath, fever, new crackles), feeding disruption (eg. nasogastric tube feeding) or feeding difficulty (choking or coughing with feeding). For patients discharged before one month after stroke onset, the researcher called the patient or next-of-kin at home or the nursing staff at a convalescent or chronic

care facility to determine whether there had been symptoms of pneumonia, feeding disruption, or feeding difficulty within the first month. (See Appendix III for a flow chart of the study.)

The bedside video-recordings for each patient on each of two days were arranged in random order for review twice by two physician reviewers blinded to baseline clinical data, clinical outcomes and each other's responses. For the pilot study, two physicians, both with general medical experience, participated.[e] They were given a 15 minute training session that included an explanation of the study, definition of the clinical criteria for an abnormal test, and a review of three sample patient video-recordings (one was normal on all bedside tests, one was abnormal on all tests and one had only a few test abnormalities). The observers were instructed to view the video-tapes continuously in half-hour sessions to avoid fatigue that might lead to inattention. The review was done by the physicians at their leisure with televisions with screens ranging in size between 20 and 30 inches. The coding was done on a form that listed each bedside test corresponding to the video-recording sequence. Each test was rated as either normal, equivocal, or abnormal (allowing for borderline cases) and also rated as either normal or abnormal (forcing a decision).

---

[e]Thankfully, my father, a family doctor in active practice, and my wife, an obstetrician with general medical experience, accepted the arduous task of reviewing over ten hours of videotape each.

4.5.1 Baseline Assessment

The baseline assessment was completed by the researcher after the video-recording on the second day. An abbreviated neurologic examination was part of the bedside video-recording protocol and included observations of speech (receptive, expressive, and dysarthric abnormalities), cranial nerves (facial droop) and motor strength graded on the Medical Research Council five point scale[f] of both upper and lower extremities.[9] The findings of visual field, sensory and deep tendon reflex abnormalities were taken from the physician notes in the hospital record because they were too cumbersome to record on video. At the end of the video-recording session, the patient was asked whether there was any difficulty swallowing the water. However, this final segment was not included in the videotape shown to the blinded videotape reviewers. There was a brief examination of the lung fields to rule out dullness on percussion, and crackles or decreased breath sounds on auscultation.

The hospital chart was reviewed for age, sex, risk factors for stroke, use of current sedative medication and intravenous

---

[f] The Medical Research Council classification is as follows: 0, no movement; 1, trace of contraction; 2, active movement when gravity is eliminated; 4, active movement against gravity and resistance; 5, normal strength.

fluid therapy beyond 48 hours. The physician and nursing progress notes were reviewed for any mention of decreased level of consciousness during the 96 hours since stroke onset. The results of the initial chest radiograph and CT head scan done within three days of admission were recorded when the final radiologic report was available. (See Appendix IV for the data entry sheet.)

4.5.2 Voice Quality

The quality of the voice was recorded as either normal, abnormal (wet, harsh, hoarse) or equivocal in response to the question: *"How are you feeling today?"* The patient was also asked to "Say pa, pa, pa; la, la, la; and ga, ga, ga."

4.5.3 Cough Quality

The patient was asked to cough as described by Linden et al.[11] The cough was graded as either normal, abnormal or equivocal. The specific request was *"Please give a strong cough. Give a really strong cough."*

4.5.4 Pharyngeal Sensation

As described by Kidd et al.,[3] sensation was examined by applying the tip of a tongue depressor to each side of the pharyngeal wall at the proximal tonsillar pillar. The request was: *"Let me know if you feel it when I touch you."* The presence of sensation was recorded as normal, abnormal or equivocal.

4.5.5 Gag Reflex

The gag reflex was tested with a small tongue blade pressed gently, briefly, but firmly on the soft palate in the midline just anterior to the uvula (palatal gag reflex) and on the posterior pharynx (pharyngeal gag reflex). The elevation of the palate was recorded as normal, abnormal (absent) or equivocal. Gagging phonation was recorded as normal, abnormal (absent) or equivocal. Although some have suggested that pharyngeal constriction is the criterion for an intact gag reflex, absent pharyngeal constriction is common in healthy subjects. Davies et al.[30] found it to be absent in 43% (29/68) of elderly subjects and found observer agreement between two blinded observers for this criterion in only 10 of 15 subjects. For the purposes of this study a normal gag reflex was required for both palatal elevation and intact gagging phonation, but not for pharyngeal constriction.

4.5.6 Swallowing

Swallowing was tested with the patient sitting upright (>80 degrees if confined to bed). The patient was observed for 15 seconds between parts of the test and for 1 minute after the tests for any choking, coughing and "wet hoarseness". The patient had to be alert enough to have spontaneous eye opening, appear responsive to environmental stimuli (follow verbal command to *"open mouth"* or non-verbal prompt with imitation of examiner) and be aware enough to move the lips around the spoon. First of all, a teaspoon of crushed ice[3], heaping approximately 0.5 cm high, was presented to the patient with the request to "Swallow this bit of ice." The water was given in a cup with a spout lid commonly used for young children. The spout lid also had a small hole in it so that the water came out of the spout at a slow and steady rate. This was intended to avoid a bolus of water that could make a false positive bedside test. Initially 50 ml, then 100 ml and finally 150 ml of water were given each with the request to *"Swallow this water"*. After each swallow of ice or water, the patient was observed for 15 seconds and then requested to "Say la, la, la". If the patient was unable to hold the cup, the examiner held it to the patients lips delivering the liquid over 1 minute. All patients received the crushed ice and the 50

[3]At Etobicoke General Hospital the ice machine dispensed small ice cubes that had to be crushed with a spoon. Ice machines that dispense crushed ice are available at some hospitals.

ml swallow test. If the patient had respiratory distress with 50 ml, the 100 ml and 150 ml amounts were not administered. The water was taken from the cold tap without running it for long to obtain room temperature rather than cold water because it can affect the triggering of the swallowing reflex.[62] The videocamera field of view included the lower face and throat from sixty degrees off center.

The ice swallow was recorded as normal, abnormal or equivocal. A normal water swallow was defined as the absence of choking, coughing and wet hoarseness on all of the three volumes of water.

For all of the above bedside tests, if the patient did not speak English, a translator gave the instructions. (See Appendix V for the video-recording protocol.)

4.5.7 Heath Care Personnel's Impression

The nurse looking after the patient on the second day of the study was asked for an impression about whether the patient was aspirating or having dysphagia. The question was phrased as follows: *"Is your patient having difficulty swallowing or controlling the airway?"* Responses were recorded as yes, no or equivocal.

4.5.8 Stroke Severity

The Barthel Index and the Linn Rapid Disability Rating Scale-2[4] were completed after the video-recording and baseline information was taken on day two. The Barthel Index was completed by reviewing the chart and talking to the nursing staff directly caring for the patient. The Rapid Disability Rating Scale was also completed by the researcher recognizing that it would have been preferable to have the nursing staff do this.[44] Of the general disability scales and disease specific instruments advocated to measure the severity of acute stroke, these two were chosen because of their documented validity and reliability. These disability scales were quicker to administer than some of the more detailed disease specific scales (e.g. NIH stroke scale) and therefore potentially more acceptable as part of a screening tool. Kidd et al.[1] (in the same study mentioned on page 17) found stroke severity as measured by the Barthel Index to be associated with aspiration on modified barium swallow videofluoroscopy. Including these measures of stroke severity in the protocol was considered important for characterization of the study patients. This would help in comparing results between studies. As well, it would help readers to decide if the results could be generalized to their patients.

4.5.9 Clinical Follow-up

The outcomes of primary interest in clinical follow-up were pneumonia and disruption of oral feeding. Other outcomes of interest were symptomatic dysphagia, nutritional status and mortality.

For research purposes, pneumonia is commonly defined as a radiographic infiltrate with any two of the following criteria: productive cough, fever > 38.3C, positive sputum culture or Gram's stain or leukocytosis (WBC >13).[3] Although this definition can be inaccurate (some of the apparent pneumonias may actually be pulmonary atelectasis or infarction), more invasive techniques (such as bronchoscopic protected brush biopsy and quantitative bacterial culture) were considered inappropriate because of the associated potential morbidity, inconvenience and research costs. Some researchers have used less rigorous definitions for pneumonia. For example, Schmidt defined pneumonia as any three of: productive cough, fever [> 38.3C], positive auscultatory findings on chest exam, PaO2 below 70 or a 10 mmHg decrease from baseline, positive sputum culture or purulent sputum Gram's stain, and chest radiographic infiltrate.[12] With these definitions for this pilot study, the finding of "presumptive pneumonia" was made with the former definition (requiring the presence of a radiographic infiltrate) and "probable pneumonia", with the latter.

For this pilot study, disruption of oral feeding (such as

use of non-oral feeding methods including nasogastric, percutaneous gastrostomy and gastrojejunostomy enteral feeding, and intravenous parenteral nutrition) was documented by review of the patient record. Although monitoring body weight, arm muscle circumference (a measure of lean muscle mass) and triceps skin fold thickness (a measure of body fat) as indicators of nutritional status,[63] was planned originally, in many cases an admission body weight was not documented because a portable scale for bedridden, non-ambulatory patients was unavailable. As well, for the pilot study there were no resources for a follow-up visit to repeat the measurements of arm muscle circumference and triceps skin fold thickness.

Similar to the prospective study by Holas et al.[14], the follow-up was by chart review. If the patient was discharged before one month after the onset of stroke, follow-up was by a telephone call to the patient (or the next-of-kin) at home, or to the patient (or health care personnel) at a chronic care or rehabilitation institution. Korner-Bitensky et al.[64] have shown that telephone interview follow-up of health status in stroke patients has reasonable agreement with face-to-face interviews. In a follow-up of several months up to five years after admission to a rehabilitation hospital for stroke or an orthopedic condition, 366 patients received both telephone and face-to-face interviews three days apart. Kappa values suggested good agreement (Kappa > .60) for the occurrence of subsequent hospitalizations and disability (defined by a Barthel score of

less than 60).

4.6 Recruitment of Patients

New hospitalized stroke patients admitted through the emergency room of Etobicoke General Hospital with a diagnosis of stroke with onset within 48 hours before admission were identified by reviewing the admission book on the medical floor, and by asking the ward nurses for the names of new stroke patients. This occurred every weekday morning and on approximately half of weekend days. A log of identified patients was kept.

Stroke was defined as a neurologic deficit of presumed vascular etiology persisting for more than 24 hours. Patients who had resolution of symptoms and signs before 24 hours after onset (i.e. transient ischemic attack) were excluded. Patients who were eventually given a diagnosis other than stroke within 72 hours of symptom onset were also excluded. Clinical exclusion criteria included: pre-existent dysphagia; non-enteral feeding method or tracheostomy, co-existent probable or presumptive pneumonia at the time of identification (as defined above on page 41), decreased level of consciousness persisting beyond 96 hours since stroke onset (defined as inability to move the lips to a spoon and cup), nothing by mouth (NPO) status beyond 96 hours, and terminal palliative care.

Reasons for non-enrollment of stroke patients were categorized into: (1) non-acute stroke defined as stroke occurring two days or more before admission, (2) ineligibility based on clinical exclusion criteria (pre-existent dysphagia, coexistent respiratory symptoms and signs of pneumonia, persistent decreased level of consciousness, kept nothing by mouth by personal physician, terminal palliative care) (3) inability of the researcher to contact the patient or next-of-kin for consent within 96 hours of stroke onset, and (4) refusal to consent for participation in the study.

4.7 Blinding Strategy

4.7.1 Baseline Data Collection and Video-Recording

In the health care field, video-recording has been used extensively in medical education.[65] It has also been used in clinical studies. Video-recording was used for the neurologic examination in a therapeutic trial by LeWitt et al.[66] who did a bedside examination subsequently reviewed by blinded observers. Video-recording has been recommended for reliability testing of physical signs[67,68] and for history taking.[69] For example, video-recording of psychiatric interviews was used to standardize diagnoses and assess inter-observer reliability in a multicenter

collaborative study.[69]

For this pilot study, the researcher did both the baseline data collection and the video-recording of the bedside examination. The protocol for the videotaping sequence was simple and rigid to combat the bias from lack of blinding of these activities. The baseline clinical data were collected after the video-recording on the second day.

4.7.2 Blinding of the Observers

The individual patient video-recordings for each of the two days were edited to subsidiary tapes with the test sequence for all patients placed in random order. In the Quattro Pro spreadsheet program, each of the videosegments was numbered 1 to 60 and given an eight digit random number. The videosegments were then placed in random order by sorting the eight digit random numbers. The videotape recordings were shown to two blinded observers who viewed the bedside tests within the total test sequence and rated each bedside test as positive or negative for each patient. In order to isolate the physical examination findings, the observers were not aware of the baseline clinical historical information and did not hear the question to the patient at the end of video-recording protocol asking whether there was difficulty with swallowing.

4.7.3 Blinding of Clinical Outcome Determination

The follow-up for probable or presumptive pneumonia and disruption of oral feeding was obtained by review of the chart and, if necessary, telephone contact with the patient, the patient's next-of-kin, or staff at the chronic care or rehabilitation hospital or at home. Although the follow-up for the planned study is to be done by independent blinded personnel, for this pilot study it was done by the researcher.

4.8 Ethical, Safety and Cost Issues

Patients or next-of-kin had to give informed consent. A copy of this consent form was put in the patient's hospital record. The ethical issue of blinding of the attending nursing and medical staff to the results of the research bedside tests was included in the consent form. As the accuracy of the tests was not known and unblinding would have caused surveillance bias, this blinding was important. The patients were informed that, although the research on these bedside tests would not benefit them directly, it might improve the care of stroke patients in the future. All patients received usual hospital care as dictated by the judgement of patient's health care team. (See Appendix VI for the consent form.)

No additional costs for patient care were incurred in this

protocol. Chest radiographs, when done, were ordered by the personal physician on the basis of clinical suspicion rather than for study purposes only. Although it would have been useful to do the modified barium swallow videofluoroscopy, this was impractical for this pilot because it would have added to health care costs and was not available at Etobicoke General Hospital.

4.9 Sample Size

Thirty patients were enrolled in this ten month pilot study. This was the minimum number recommended by Koran in his review of reliability studies of clinical methods.[70] This number of patients gives crude estimates of the proportions of observed agreement (P_o) and expected agreement based on chance (P_e) as well as Kappa itself. A review of the ward admission book at the outset of the pilot had suggested that approximately 15 patients per month were admitted with a diagnosis of stroke, of whom one third were estimated to be eligible and consenting. This gave the initial projection of six months to enroll 30 patients.

A sample size equation can be derived from Cohen's approximation of the standard error of Kappa which is based on P_o and P_e (Appendix II). A simplified sample size table was created to show the trade-off between the size of the confidence interval around Kappa and the sample size for incremental and reasonable values of P_o and P_e. (See Appendix VII.)

5. Pilot Study Results

5.1 Patient Characteristics

5.1.1 Stroke Admission Log

A log of patients admitted with a diagnosis of stroke was kept during the study period. The hospital records of stroke patients were reviewed in detail for the first half of the pilot (September 1995 to January 1996). The log of admitted stroke patients during the second half of the pilot study was incomplete because of the unforseen occurrence of dramatic medical bed closures related to hospital budget cuts, which caused daily "bedspacing" of patients to non-medical wards and "overnighting" of patients in the emergency room. Reasons for non-enrollment of stroke patients during the first half of the study are listed in Table 2.

For patients within the first half of the study, the mean ages were: 71 for the non-acute stroke patients, 76 for the ineligible patients, 69 for patients who could not be contacted, 74 for all stroke patients refusing participation and 76 for the enrolled patients. The median length of hospital stay was shorter for the ineligible and non-contacted patient groups (9 and 14 days respectively) than for the refusal group (24 days) and

enrolled group (36 days). Analysis for statistically significant differences between the four groups for means ages and median length of stay was not meaningful because of the small patient numbers. There were three in-hospital deaths, two in the ineligible group and one in the non-contact group. Refusals to participate usually came not from the patients but from next-of-kin who were uncomfortable with deciding about participation in a study for their family member.

Table 2. Patients Admitted with Diagnosis of Stroke during First Half of Study (September 1995 to January 1996).

Non-acute stroke patients		24
Acute stroke patients (< 48 hours since onset)		
Ineligible:	pre-existent dysphagia	0
	coexistent pneumonia	1
	persistent decreased level of consciousness	4
	Nothing by mouth	1
Eligible:	non-contact	14
	refusal	7
	enrollment	14
Total patients admitted with diagnosis of stroke		65

5.1.2 Enrolled Patients

All enrolled patients entered the hospital through the emergency room after being referred to the internist-on-call by the emergency room physician. The age, sex, ethnicity, clinical features, Barthel Index and computed tomographic findings of all 30 enrolled patients are listed in Table 3. Approximately one third of the patients had speech difficulty and one half had signs of facial weakness. Most patients had some degree of limb weakness. Of the 28 patients with limb weakness, the worst motor deficit was MRC grade 0 - 1 in four patients, grade 2 - 3 in eight patients, and grade 4 in sixteen patients.

Based on the cut-offs of the Barthel Index proposed by Granger (see page 22), five patients had a score implying total dependence (20/100 or less) and six patients had a score suggestive of moderately severe dependence (greater than 20 and less than 40/100). Sixteen patients had a score suggestive of independence (60/100 or more). Three scored between 40 and 60.

All but one patient had a computed tomographic scan of the brain. Infarction was more common than hemorrhage (intracerebral, basal ganglia, lacunar) but several patients had only non-specific findings including atrophy and leukoencephalopathy.

None of the enrolled patients was thought by the nurses to have difficulty with swallowing or control of the airway. None of the patients complained of difficulty swallowing.

Table 3. Characteristics of the 30 Enrolled Patients.

Age mean - yrs. (range)	73(49-90)
Sex - no.(%)	
Female	13(43)
Male	17(57)
Ethnic group or race - no. (%)	
White	22(73)
Black	2(7)
Asian	6(20)
Clinical Deficit	
Speech difficulty - no.(%)	
Dysarthria	6(20)
Global	2(7)
Expressive	1(3)
Receptive	1(3)
Facial weakness - no. (%)	
Right	7(23)
Left	8(27)
Motor weakness - no. (%)	
Right	9(30)
Left	16(53)
Bilateral	3(10)
Barthel index - mean (range)	59(5-100)
CT scan findings - no. (%)	
Infarction	16(53)
Hemorrhage	5(17)
Non-specific*	8(27)
Not done	1(3)

* Includes: atrophy, subacute leukoencephalopathy

5.2 Follow-up

The outcomes on follow-up of the 30 enrolled patients are listed in Table 4. Only one of the 30 developed probable pneumonia during the one month follow-up. Six days after admission, this patient developed cough, shortness of breath, fever and crackles at the right lung base. A chest radiograph was not done. He was treated with antibiotics and improved. Although nine of the patients received intravenous fluids beyond 48 hours, none were given nasogastric feeding during the hospitalization. At one month, all patients were taking food orally and had no complaints of feeding difficulty. The median length of stay in all 30 enrolled patients was 23 days with a range of 5 - 212 days. There were no deaths in the enrolled patients within one month after onset of stroke.

Table 4. Outcomes of the 30 Enrolled Patients

CHARACTERISTIC	VALUE
Intravenous fluid administration beyond 48 h - no. (%)	9 (30)
Nasogastric feeding in first month - no. (%)	0 (0)
Pneumonia (probable) in first month - no. (%)	1 (3)
Length of hospital stay median - d (range)	23(5-212)
Deaths - no. (%)	0 (0)

5.3 Protocol Modifications and Implementation

The pilot study protocol was followed successfully with some modifications itemized as follows:

1. Identification of patients in a log of all stroke admissions was limited to the first half of the study because of the difficulty in tracking admissions in the second half (see page 48).

2. Measuring patient weight (as an index of nutritional status) had to be abandoned. Most patients were bedridden during the first week of their illness and a Hoyer lift scale was not available at Etobicoke General Hospital (see page 42).

3. Machine made crushed ice was not available on the ward at Etobicoke General. Small ice cubes had to be crushed at the bedside to get a finer consistency (see page 38).

4. The Barthel and RDSS-2 disability scales were completed by the researcher rather than by the nursing staff. The pilot protocol could not add to the already heavy workload of the nursing staff (see pages 31 and 40).

5. The videotape reviewers had difficulty seeing where the tongue blade touched the pharynx to categorize the gag reflex into

palatal and pharyngeal gag reflexes. A gag reflex was labelled intact if there was both gagging phonation and palatal elevation during this short videosegment (see page 37).

6. Although it was intended that the videotape reviewers watch the tapes in continuous half hour sessions without rewinding, both reviewers had to rewind the videotape several times per half hour session owing to either an interruption or lapse of attention. Both reviewers found the task tedious. It took the two physician observers approximately three weeks to review the ten hours[h] of video-tape.

5.4 Prevalence of Abnormalities on the Bedside Tests

Table 5 lists the average number of the 30 patients considered abnormal for each bedside test (on either of the two days for either of the two observers on either of the two viewings). The proportion of patients rated abnormal was small for pharyngeal sensation (5%) and ice swallow (5%). The proportion rated abnormal was greater for gag reflex (36%) and water swallow (33%).

---

[h] Each reviewer had to review tapes for thirty patients containing five minute video segments for each of the two days twice.

Table 5. Prevalence of Abnormalities on Bedside Tests in the 30 Enrolled Patients

Bedside Test	Overall average No. of 30 (%)	Range No. of 30 (%)
Voice quality	4.3 (14)	1-6 (3-20)
Cough quality	5.3 (18)	3-7 (3-23)
Pharyngeal sensation	1.4 (5)	1-2 (3-7)
Gag reflex	11 (36)	9-14 (30-47)
Ice swallow	1.4 (5)	0-3 (0-10)
Water swallow	10 (33)	8-12 (27-40)

Note: Overall average prevalence of abnormality is average number considered abnormal out of 30 patients for the two observers on either of the two days on either of the two viewings.

5.5 Reliability of Bedside Tests

The intra-observer agreement between first and second viewings was calculated for each of the two observers for each of the two days. The inter-observer agreement for each of the viewings was also calculated for each of the two days. The two by two tables along with calculated observed agreement (P_o), expected agreement (P_e), and Kappa (κ) with 95% confidence limits (CI) are listed in Tables 6-17.

Table 6. Intra-observer Reliability of Voice Quality

					Po	Pe	K (C.I.)§
Day 1		**Observer 1**					
		Viewing	1		.90	.79	.53(.08,.97)
		Yes*	No				
Viewing 2	Yes	25	0	25(.83)			
	No	3	2	5(.17)			
		28(.93)†	2(.07)	30			
Day 1		**Observer 2**					
		Viewing	1		.93	.77	.71(.35,1.0)
		Yes	No				
Viewing 2	Yes	25	0	25(.83)			
	No	2	3	5 (.17)			
		27(.90)	3(.10)	30			
Day 2		**Observer 1**					
		Viewing	1		.93	.77	.71(.35,1.0)
		Yes	No				
Viewing 2	Yes	25	0	25(.83)			
	No	2	3	5(.17)			
		27(.90)	3(.10)	30			
Day 2		**Observer 2**					
		Viewing	1		.97	.70	.89(.68,1.0)
		Yes	No				
Viewing 2	Yes	24	1	25(.83)			
	No	0	5	5(.17)			
		24(.80)	6(.20)	30			

* Yes = normal voice quality.
† Marginal fractions are in brackets on two by two tables.
§ Confidence limits may not be valid because of small sample size.

Table 7. Inter-observer Reliability of Voice Quality

					Po	Pe	K (C.I.)§
Day 1		**Viewing 1**					
		Observer	1		.90	.85	.35(-.22,.92)
		Yes*	No				
Observer 2	Yes	26	1	27(.90)			
	No	2	1	3(.10)			
		28(.93)†	2(.07)	30			
Day 1		**Viewing 2**					
		Observer	1		.93	.72	.76(.44,1.0)
		Yes	No				
Observer 2	Yes	24	1	25(.83)			
	No	1	4	5(.17)			
		25(.83)	5(.17)	30			
Day 2		**Viewing 1**					
		Observer	1		.90	.74	.62(.23,1.0)
		Yes	No				
Observer 2	Yes	24	0	24(.80)			
	No	3	3	6 (.20)			
		27(.90)	3(.10)	30			
Day 2		**Viewing 2**					
		Observer	1		.87	.72	.52(.11,.93)
		Yes	No				
Observer 2	Yes	23	2	25(.83)			
	No	2	3	5(.17)			
		25(.83)	5(.17)	30			
Average					.90	.76	.59‡

* Yes = normal voice quality.
† Marginal fractions are in brackets on two by two tables.
‡ Overall Kappa calculated with average values P_o and P_e.
§ Confidence limits may not be valid because of small sample size.

Table 8. Intra-observer Reliability of Cough Quality

					Po	Pe	K (C.I.)§
Day 1		**Observer 1**					
		Viewing	1		.97	.66	.90(.71,1.0)
		Yes*	No				
Viewing 2	Yes	23	0	23(.77)			
	No	1	6	7(.23)			
		24(.80)†	6(.20)	30			
Day 1		**Observer 2**					
		Viewing	1		.90	.79	.52(.04,.99)
		Yes	No				
Viewing 2	Yes	25	1	26(.87)			
	No	2	2	4(.13)			
		27(.90)	3(.10)	30			
Day 2		**Observer 1**					
		Viewing	1		.97	.70	.89(.68,1.0)
		Yes	No				
Viewing 2	Yes	24	1	25(.83)			
	No	0	5	5(.17)			
		24(.80)	6(.20)	30			
Day 2		**Observer 2**					
		Viewing	1		.97	.70	.89(.68,1.0)
		Yes	No				
Viewing 2	Yes	24	0	24(.80)			
	No	1	5	6(.20)			
		25(.83)	5(.17)	30			

* Yes = normal cough quality.
† Marginal fractions are in brackets on two by two tables.
§ Confidence limits may not be valid because of small sample size.

Table 9. Inter-observer Reliability of Cough Quality

					Po	Pe	K (C.I.)§
Day 1		**Viewing 1**					
		Observer	1		.90	.74	.62(.23,1.0)
		Yes*	No				
Observer 2	Yes	24	3	27(.90)			
	No	0	3	3(.10)			
		24(.80)†	6(.20)				
Day 1		**Viewing 2**					
		Observer	1		.90	.70	.67(.34,1.0)
		Yes	No				
Observer 2	Yes	23	3	26(.87)			
	No	0	4	4(.13)			
		23(.77)	7(.23)	30			
Day 2		**Viewing 1**					
		Observer	1		.97	.70	.89(.68,1.0)
		Yes	No				
Observer 2	Yes	24	1	25(.83)			
	No	0	5	5 (.17)			
		24(.80)	6(.20)	30			
Day 2		**Viewing 2**					
		Observer	1		.90	.70	.67(.32,1.0)
		Yes	No				
Observer 2	Yes	23	1	24(.80)			
	No	2	4	6(.20)			
		25(.83)	5(.17)	30			
Average					.92	.71	.72‡

* Yes = normal cough quality.
† Marginal fractions are in brackets on two by two tables.
‡ Overall Kappa calculated with average values P_o and P_e.
§ Confidence limits may not be valid because of small sample size.

Table 10. Intra-observer Reliability of Pharyngeal Sensation

					Po	Pe	K (C.I.)§
Day 1		**Observer 1**					
		Viewing	1		.97	.90	.65(.02,1.0)
		Yes*	No				
Viewing 2	Yes	28	0	28(.93)			
	No	1	1	2(.07)			
		29(.97)†	1(.03)	30			
Day 1		**Observer 2**					
		Viewing	1		1.0	.94	1.0‖
		Yes	No				
Viewing 2	Yes	29	0	29(.97)			
	No	0	1	1(.03)			
		29(.97)	1(.03)	30			
Day 2		**Observer 1**					
		Viewing	1		.90	.85	.35(-.22,.92)
		Yes	No				
Viewing 2	Yes	26	1	27(.90)			
	No	2	1	3(.10)			
		28(.93)	2(.07)	30			
Day 2		**Observer 2**					
		Viewing	1		1.0	.94	1.0‖
		Yes	No				
Viewing 2	Yes	29	0	29(.97)			
	No	0	1	1(.03)			
		29(.97)	1(.03)	30			

* Yes = normal pharyngeal sensation.
† Marginal fractions are in brackets on two by two tables.
§ Confidence limits may not be valid because of small sample size.
‖ Confidence limits are degenerate when Kappa = 1.

Table 11. Inter-observer Reliability of Pharyngeal Sensation

					Po	Pe	K (C.I.)§
Day 1		**Viewing 1**					
		Observer	1		1.0	.94	1.0‖
		Yes*	No				
Observer 2	Yes	29	0	29(.97)			
	No	0	1	1(.03)			
		29(.97)†	1(.03)	30			
Day 1		**Viewing 2**					
		Observer	1		.97	.90	.65(.02,1.0)
		Yes	No				
Observer 2	Yes	28	1	29(.97)			
	No	0	1	1(.03)			
		28(.93)	2(.07)	30			
Day 2		**Viewing 1**					
		Observer	1		.90	.90	-.05(-.11,.02)
		Yes	No				
Observer 2	Yes	27	2	29(.97)			
	No	1	0	1(.03)			
		28(.93)	2(.07)	30			
Day 2		**Viewing 2**					
		Observer	1		.87	.87	-.05(-.13,.03)
		Yes	No				
Observer 2	Yes	26	3	29(.97)			
	No	1	0	1(.03)			
		27(.90)	3(.10)	30			
Average					.94	.90	.33‡

* Yes = normal pharyngeal sensation.
† Marginal fractions are in brackets on two by two tables.
‡ Overall Kappa calculated with average values P_o and P_e.
§ Confidence limits may not be valid because of small sample size.
‖ Confidence limits are degenerate when Kappa = 1.

Table 12. Intra-observer Reliability of Gag Reflex

					Po	Pe	K (C.I.)§
Day 1		**Observer 1**					
		Viewing	1		.90	.57	.77(.52,1.0)
		Yes*	No				
Viewing 2	Yes	19	2	21(.70)			
	No	1	8	9(.30)			
		20(.67)†	10(.33	30			
Day 1		**Observer 2**					
		Viewing	1		.90	.57	.77(.52,1.0)
		Yes	No				
Viewing 2	Yes	19	2	21(.70)			
	No	1	8	9(.30)			
		20(.67)	10(.33	30			
Day 2		**Observer 1**					
		Viewing	1		.80	.52	.59(.29,.88)
		Yes	No				
Viewing 2	Yes	15	2	17(.57)			
	No	4	9	13(.43)			
		19(.63)	11(.37	30			
Day 2		**Observer 2**					
		Viewing	1		.87	.51	.73(.49,.97)
		Yes	No				
Viewing 2	Yes	15	1	16(.53)			
	No	3	11	14(.47)			
		18(.60)	12(.40	30			

* Yes = normal palatal elevation and gagging phonation.
† Marginal fractions are in brackets on two by two tables.
§ Confidence limits may not be valid because of small sample size.

Table 13. Inter-observer Reliability of Gag Reflex

					Po	Pe	K (C.I.)§
Day 1		**Viewing 1**					
		Observer	1		1.0	.56	1.0‖
		Yes*	No				
Observer 2	Yes	20	0	20(.67)			
	No	0	10	10(.33)			
		20(.67)†	10(.33				
Day 1		**Viewing 2**					
		Observer	1		.87	.58	.68(.40,.97)
		Yes	No				
Observer 2	Yes	19	2	21(.70)			
	No	2	7	9(.30)			
		21(.70)	9(.30)	30			
Day 2		**Viewing 1**					
		Observer	1		.83	.53	.65(.37,.93)
		Yes	No				
Observer 2	Yes	16	2	18(.60)			
	No	3	9	12(.40)			
		19(.63)	11(.37	30			
Day 2		**Viewing 2**					
		Observer	1		.83	.50	.66(.40,.93)
		Yes	No				
Observer 2	Yes	14	2	16(.53)			
	No	3	11	14(.47)			
		17(.57)	13(.43	30			
Average					.88	.54	.74‡

* Yes = normal palatal elevation and gagging phonation.
† Marginal fractions are in brackets on two by two tables.
‡ Overall Kappa calculated with average values P_o and P_e.
§ Confidence limits may not be valid because of small sample size.
‖ Confidence limits are degenerate when Kappa = 1.

Table 14. Intra-observer Reliability of Ice Swallow

					Po	Pe	K (C.I.)§
Day 1		**Observer 1**					
		Viewing	1		1.0	.88	1.0‖
		Yes*	No				
Viewing 2	Yes	28	0	28(.93)			
	No	0	2	2(.07)			
		28(.93)†	2(.07)	30			
Day 1		**Observer 2**					
		Viewing	1		.97	.85	.78(.37,1.0)
		Yes	No				
Viewing 2	Yes	27	0	27(.90)			
	No	1	2	3(.10)			
		28(.93)	2(.07)	30			
Day 2		**Observer 1**					
		Viewing	1		1.0	1.0	1.0‖
		Yes	No				
Viewing 2	Yes	30	0	0(1.0)			
	No	0	0	0(0.0)			
		30(1.0)	0(0.0)	30			
Day 2		**Observer 2**					
		Viewing	1		.93	.93	0(-.24,.28)
		Yes	No				
Viewing 2	Yes	28	2	30(1.0)			
	No	0	0	0(0.0)			
		28(.93)	2(.07)	30			

* Yes = normal ice swallow.
† Marginal fractions are in brackets on two by two tables.
§ Confidence limits may not be valid because of small sample size.
‖ Confidence limits are degenerate when Kappa = 1.

Table 15. Inter-observer Reliability of Ice Swallow

					Po	Pe	K (C.I.)§
Day 1		**Viewing 1**					
		Observer	1		1.0	.88	1.0‖
		Yes*	No				
Observer 2	Yes	28	0	28(.93)			
	No	0	2	2(.07)			
		28(.93)†	2(.07)	30			
Day 1		**Viewing 2**					
		Observer	1		.93	.82	.63(.18,1.0)
		Yes	No				
Observer 2	Yes	26	0	26(.87)			
	No	2	2	4(.13)			
		28(.93)	2(.07)	30			
Day 2		**Viewing 1**					
		Observer	1		.93	.93	.0(-.24,.28)
		Yes	No				
Observer 2	Yes	28	0	28(.93)			
	No	2	0	2 (.07)			
		30(1.0)	0(0)				
Day 2		**Viewing 2**					
		Observer	1		1.0	1.0	1.0‖
		Yes	No				
Observer 2	Yes	30	0	30(1.0)			
	No	0	0	0(0.0)			
		30(1.0)	0(0.0)				
Average					.97	.91	.63‡

* Yes = normal ice swallow.
† Marginal fractions are in brackets on two by two tables.
‡ Overall Kappa calculated with average values P_o and P_e.
§ Confidence limits may not be valid because of small sample size.
‖ Confidence limits are degenerate when Kappa = 1.

Table 16. Intra-observer Reliability of Water Swallow

					Po	Pe	K (C.I.) §
Day 1		**Observer 1**					
		Viewing	1		.90	.56	.77(.53,1.0)
		Yes*	No				
Viewing 2	Yes	19	0	19(.63)			
	No	3	8	11(.37)			
		22(.73)†	8(.27)	30			
Day 1		**Observer 2**					
		Viewing	1		.90	.57	.77(.52,1.0)
		Yes	No				
Viewing 2	Yes	19	1	20(.67)			
	No	2	8	10(.33)			
		21(.70)	9(.30)	30			
Day 2		**Observer 1**					
		Viewing	1		.90	.56	.77(.53,1.0)
		Yes	No				
Viewing 2	Yes	19	0	19(.63)			
	No	3	8	11(.37)			
		22(.73)	8(.27)	30			
Day 2		**Observer 2**					
		Viewing	1		.90	.53	.79(.56,1.0)
		Yes	No				
Viewing 2	Yes	17	2	19(.63)			
	No	1	10	11(.37)			
		18(.60)	12(.40	30			

* Yes = normal swallow free of choking and wet hoarseness on three amounts.
† Marginal fractions are in brackets cn two by two tables.
§ Confidence limits may not be valid because of small sample size.

Table 17. Inter-observer Reliability of Water Swallow

					Po	Pe	K (C.I.)§
Day 1		**Viewing 1**					
		Observer	1		.97	.59	.92(.76,1.0)
		Yes*	No				
Observer 2	Yes	21	0	21(.70)			
	No	1	8	9(.30)			
		22(.73)†	8(.27)				
Day 1		**Viewing 2**					
		Observer	1		.90	.54	.78(.55,1.0)
		Yes	No				
Observer 2	Yes	18	2	20(.67)			
	No	1	9	10(.33)			
		20(.63)	11(.37)	30			
Day 2		**Viewing 1**					
		Observer	1		.87	.55	.71(.45,.96)
		Yes	No				
Observer 2	Yes	18	0	18(.60)			
	No	4	8	12(.40)			
		22(.73)	8(.27)	30			
Day 2		**Viewing 2**					
		Observer	1		.93	.54	.86(.66,1.0)
		Yes	No				
Observer 2	Yes	18	1	19(.63)			
	No	1	10	11(.37)			
		19(.63)	11(.37				
Average					.92	.56	.81‡

* Yes = normal swallow free of choking and wet hoarseness on three amounts.
† Marginal fractions are in brackets on two by two tables.
‡ Overall Kappa calculated with average values P_o and P_e.
§ Confidence limits may not be valid because of small sample size.

Although the confidence intervals for Kappas were included, their statistical validity is threatened by the small sample size in the face of the large sample assumption necessary for the Kappa statistic to approximate the normal distribution. In any event, the confidence intervals around Kappas were usually wide, making statistical comparison of Kappa values impossible. As would be expected, the Kappa values for intra-observer reliability tended to be similar or better than those for inter-observer reliability. Taking voice quality as an example, Kappas ranged from .53 to .89 for intra-observer reliability compared to .35 to .76 for inter-observer reliability (Tables 6 and 7). For cough quality, Kappa values ranged from .52 to .90 for intra-observer reliability compared to .62 to .89 for inter-observer reliability (Tables 8 and 9).

For pharyngeal sensation, observed and expected agreements were high and Kappa values were inconsistent ranging from -.05 to 1 (Tables 10 and 11). The same occurred with the ice swallow (Tables 14 and 15). The inter-observer Kappa values for the cough quality, the gag reflex (Table 13) and the water swallow (Table 17) tended to be the highest and most consistent. All three of these had inter-observer Kappas > 0.6.

5.6 Determination of Sample Sizes with Pilot Study Estimates

Inter-observer reliability (rather than intra-observer reliability) was chosen as most relevant for the sample size calculations because the ultimate goal was to find clinical predictors that could be used by a variety of observers to detect aspiration and dysphagia in their acute stroke patients. For inter-observer reliability, the average values for P_o and P_e were used to calculate an "overall estimator of Kappa" for each of the bedside tests. With the simplified sample size table (Table 23), the pairs of values for observed (P_o) and expected (P_e) agreement as well as the average values of P_o and P_e from the inter-observer reliability tables were used to determine the sample size required for a Kappa within a 95% confidence interval of ±.10 and ±.20 . These sample sizes are tabulated in Table 18.

Table 18. Sample Sizes based on Estimates of P_o and P_e for Inter-observer Reliability.

BEDSIDE TEST	Pair	P_o	P_e	KAPPA	SAMPLE SIZE FOR κ ±.10	SAMPLE SIZE FOR κ ±.20
Voice quality	d1 o1*	.90	.85	.35	1537	384
	d1 o2	.93	.72	.76	203	m
	d2 o1	.90	.74	.62	553	138
	d2 o2	.87	.72	.52	384	96
	average	.90	.76	.59†	553	138
Cough quality	d1 o1	.90	.74	.62	553	138
	d1 o2	.90	.70	.67	384	96
	d2 o1	.97	.70	.89	124	m
	d2 o2	.90	.70	.67	384	96
	average	.92	.71	.72	384	96
Pharynx sensation	d1 o1	1.0	.94	1.0	m‡	m
	d1 o2	.97	.90	.65	1118	279
	d2 o1	.90	.90	-0.05	∞ §	∞
	d2 o2	.87	.87	0.05	∞	∞
	average	.94	.90	.33	1825	456
Gag reflex	d1 o1	1.0	.56	1.0	m‡	m
	d1 o2	.87	.58	.68	306	77
	d2 o1	.83	.53	.65	242	60
	d2 o2	.83	.50	.66	196	m
	average	.88	.54	.74	171	m
Ice swallow	d1 o1	1.0	.88	1.0	m	m
	d1 o2	.93	.82	.63	456	114
	d2 o1	.93	.93	0.0	∞	∞
	d2 o2	1.0	.99	1.0	m	m
	average	.97	.91	.63	1118	279
Water swallow	d1 o1	.97	.59	.92	70	m
	d1 o2	.90	.54	.78	171	m
	d2 o1	.87	.55	.71	242	m
	d2 o2	.93	.54	.86	90	m
	average	.92	.56	.81	171	m

* d1 o1 = day 1, viewing 1. † The "overall Kappa" is a calculated value from the average values of the estimates for P_o and P_e. ‡ When P_o (observed agreement) is equal to 1.0, the sample size equals m. § When P_o and P_e are approximately equal, the sample size becomes infinite. m = minimum sample size.

6. Discussion

6.1 Rates of Enrollment

This pilot study had an enrollment rate that was low with only 30 patients enrolled in 10 months. Although the medical admission log book in 1995 had listed the admitting diagnosis of stroke to be on average 15 patients per month, several factors seemed to contribute to the lower than projected enrollment. Firstly, a large proportion (approximately one-quarter) of patients admitted with a diagnosis of stroke actually had illness due not to acute but rather non-acute stroke; these patients were admitted with complications of an old stroke or other conditions that made them unable to cope with their stroke disability. Secondly, during the study, as the scarcity of hospital beds increased, many of the patients entered and left hospital so quickly that it was difficult to enroll them in the study. This observation is consistent with the shorter median length of stay for the non-contact group. Several of the patients who could not be contacted to participate went home before the researcher could enroll them in the study. Medical bed closures led to a queue for hospital admission at the back of the emergency room and bedspacing onto the surgical floors in the last several months of the study. Owing to this unforeseen development, it became increasingly difficult to identify stroke patients within the

first few days of admission. Thirdly, because many of the eligible acute stroke patients were elderly, had stroke-related speech difficulty and/or limited English literacy, the consent process was complicated by the need to contact the next-of-kin. This often meant setting up meetings during evening visiting hours, which was difficult. Next-of-kin can be unwilling to discuss the participation of a family member with a catastrophic illness such as stroke.[12] The consent issue has been recognized as a major obstacle in clinical research in neurologic patients.[27]

The stroke log data suggest that the enrolled patients may have differed systematically from the population of all eligible acute stroke patients. They may have been, on average, more severe because their median length of stay was longer than the patients who could not be contacted or those who refused to participate. The acute stroke patients who were ineligible (mostly because of persistent decreased level of consciousness) had the highest in-hospital mortality (2/6 patients died). In a larger study, it would be useful to better characterize the non-enrolled patients. For example, collection of baseline clinical information and stroke severity (such as with the Barthel index) on non-enrolled patients might be permissible with institutional consent.

The low enrollment rate suggests that for a larger study, intensive measures to contact all eligible patients and families would be useful. A dedicated study nurse would be invaluable in

the recruitment process. Given the problem of "bedspacing" and "overnighters", efforts to identify the patients in the emergency room, rather than on the medical ward, would be worthwhile.

6.2 Follow-up

The proportion of eligible acute stroke patients who had difficulty with swallowing, required nasogastric feeding or developed pneumonia during the first month was low. Before the pilot study, nasogastric feeding in acute stroke patients appeared to be more frequent to the researcher. However, this impression may be explained by the epidemiologic equation: *Prevalence = Incidence X Duration*. In other words, patients with stroke severe enough to develop pneumonia or require nasogastric feeding are on the wards longer than the milder stroke patients. Another reason for the lower than expected rate of outcomes may be that patients with severe strokes, such as those causing persistent decreased level of consciousness, were not eligible for this study. The finding of only one pneumonia in the 30 patients in the pilot is lower than was suggested by the study by Gordon et al.[18] who found that 12% (11/91) of their series of consecutive acute stroke patients developed chest infections. The higher proportion in the series of Gordon et al. may be because they included non-alert patients (four patients were unconscious and four were drowsy). Teasell et al.[51] recently concluded that

pneumonia was an uncommon complication of stroke. In their retrospective review of 441 stroke patients admitted to a stroke rehabilitation unit during an eight year period, only 2.7% (12/441) had pneumonia in the four months after onset of stroke. However, this study likely underestimated the incidence of pneumonia because it included only patients who were eventually admitted to a rehabilitation ward. A patient who had developed pneumonia and died would have been missed in their study.

The small proportion of enrolled patients who develop the pneumonia make the required sample size for a clinical study using this outcome large, whether it be for a prospective cohort study (determining the relative risk of developing pneumonia in patients with risk factors such as abnormal bedside tests), a randomized controlled trial (evaluating a therapeutic intervention), or a clinical prediction rule (discriminating those at low risk and high risk of pneumonia).

From sample size tables for prospective cohort studies[73] (assuming an α error of .05 and β error of .10), if the incidence of pneumonia in the group without the risk factor (bedside test negative) is 1/100 and the frequency of the risk factor (bedside test positive) is 50%, then a sample size of 3100 patients would be necessary to detect a relative risk of 2.0. A sample size of 567 would be necessary to detect a relative risk of 4.0.

For a randomized controlled trial done in patients in whom the baseline risk of pneumonia is a generous 10%, sample size tables (assuming again an α error of .05 and β error of .10) show

that to detect a 50% relative risk reduction of pneumonia (to 5%), a sample size of almost 1000 patients would be required.[74]

For the derivation of a clinical prediction rule, the sample size would also be large. With the rule of thumb of 10 outcome events per clinical predictor,[75] a clinical prediction having only three clinical predictors would require 30 outcomes. For an outcome such as pneumonia with a baseline rate of 3% (or 1 pneumonia per 30 patients) approximately 900 patients would need to be enrolled.

A longer follow-up period, such as six months for example, would require more resources but could potentially increase the number of clinical outcomes. On the other hand, if pneumonia were related to the stroke-induced dysphagia and aspiration, it would likely occur early. Extending the follow-up beyond six months would likely dilute the outcome of pneumonia related to aspiration with community acquired pneumonia.

The need for large sample sizes for studies with a clinical outcome probably explains why a surrogate outcome, such as modified barium swallow videofluoroscopy, has been chosen more often by previous investigators.

6.3 Protocol Implementation

For a larger grant-funded study, resources might be available for study personnel to collect baseline data, complete

the stroke severity scales, and follow enrolled patients. Because the video-tape review turned out to be quite time-consuming and arduous, even for this small pilot study, compensation for reviewers would be desirable if video-recording were to be used. Also helpful might be the use of a larger group of observers, as was done by Maher et al.[76] in their study of the reliability of the plantar response.

6.4 Prevalence of Abnormalities on Bedside Tests

Gordon et al.[18] found a 30%(24/81) prevalence of abnormal cough in a series of acute stroke patients, which was higher than was found in this pilot study where only 18% of patients were rated as having an abnormal cough. They also found a 30% (24/81) prevalence of an abnormal gag reflex in that series. This was similar to this pilot study in which the prevalence of an abnormal gag reflex was 36% (11/30). Unfortunately, Gordon et al. did not describe in sufficient detail how they tested the cough and gag reflex.

The study of Kidd et al.[3] done in consecutive patients, found a much higher prevalence of abnormal pharyngeal sensation (65%) than this pilot study (with prevalence of 5%). An important weakness of the study by Kidd et al. was a lack of observer blinding. Furthermore, because there was a high proportion of patients with aspiration on modified barium swallow

videofluoroscopy (25/60), it is possible that the stroke patients were at the severe end of the spectrum of all hospitalized stroke patients.

The prevalence of an abnormal swallow test in this pilot study (33%) was similar to that found by Gordon et al.[18] who noted that 45% of patients had difficulty drinking 50 ml of water and that of Barer et al.[19] who found 28% (29/105) had dysphagia with a mouthful of water at 24 hours. Kidd et al.[3] found 42% of patients were unable to swallow 50 ml of water.

The differences in the prevalence of bedside test abnormalities among the published studies are likely caused by differences in the criteria for the definition of an abnormality and by systematic and random differences in the patient selection.

Considering the prevalence of a test abnormality is important because it affects the estimates of reliability as measured by the Kappa statistic. When an abnormality is uncommon, the chance expected agreement for normality will be high. This makes the Kappa low and often unstable.[77] Grove et al.[69] have called the prevalence of abnormality the "base rate", and have suggested that the "base rate" be taken into consideration when interpreting Kappa. They recommend that the Kappa statistic not be used for tests with low "base rates" (< 5%). Spitznagel and Helzer[78] have examined the base rate problem for the Kappa statistic. They noted that in clinical studies where the base rate is relatively high (between 10% and 80%) the Kappa statistic

is relatively stable. Base rates in this range allow a better comparison among agreement studies.

6.5 Reliability of Bedside Tests

Inter-observer reliability is most relevant to the clinical use of the bedside tests. Although pooling all observations on the 30 subjects on both days and on both viewings would increase the apparent sample size, the assumption that the observations were independent, would be lost. Rather than arbitrarily selecting any one pair of P_o and P_e values, averages of the observed agreements (P_o) and expected agreements (P_e) were judged as the best estimates of the true values of P_o and P_e. The average values of P_o and P_e were used to calculate an "overall Kappa" for inter-observer reliability as the best estimate of Kappa for use in the sample size calculations. A more conservative approach would be to choose the least favourable pair of P_o and P_e (i.e. the pair that gives the lowest Kappa). This might be the "worst case Kappa" and would give higher required sample sizes.

The "overall Kappa" estimates suggest substantial reliability (>.60) for cough quality, gag reflex, ice swallow and water swallow. The "overall Kappa" estimates suggest moderate reliability (>.40) for voice quality. Pharyngeal sensation had only fair reliability (>.20).

The Kappa values for pharyngeal sensation illustrate the paradox of Kappa described by Feinstein.[77] This is the conversion of a high simple observer agreement (P_o > .90) to a relatively low Kappa (.33) because of a low prevalence of abnormality (5%) causing a high chance expected agreement (P_e >.88). As can be seen from table 11, although the two by two tables for pharyngeal sensation were quite similar in their high concordance for normality, a few discordant observations shifted the Kappa wildly from one to near zero.

The practical advantage of using a video-camera to do the reliability testing was that it was not necessary to get the same two observers to examine the same subject on the same day.[69] Video-recording has been used to compare observer agreement for physical signs between clinicians in different centers. English et al.[67] used video-recordings to examine the inter-observer reliability of physical signs of respiratory distress in children in five different countries. However, the video-recording technique comes at a cost: the reliability is probably over-estimated.

The causes of unreliability can be categorized according to test stimulus, observer perception, observer processing, environment and subjects. See Appendix I for a discussion of this framework. The use of the video-recorder would likely remove the variability that could occur in clinical practice because of the test stimulus (i.e. physical examination technique) and the environment. The variability tested by the video-recording

technique is limited to the observer perception (visual acuity and attention) and the observer processing (the observer's criterion for rating the test positive or negative). The video-recording technique with later review of the videotape could even enhance observer perception for a bedside test, such as the gag reflex, because the oropharynx was magnified when viewed on the large television screen. As well, the observers might be more attentive that in the rushed clinical situation. The idea of higher reliability in studies involving the review of videotape compared to the actual examination of patients is analogous to the finding of higher reliability for the differentiation of heart murmurs in studies involving the review of audiotape compared to bedside auscultation.[79] As the reliability of tests is probably over-estimated, the video-recording method of determining reliability is probably better at determining which tests are unreliable than determining the actual reliability of the tests. In other words, if a test has low reliability as measured under the artificial conditions of the video-recording technique, assuming good camera technique, then it is probably going to be even lower in practice.

Based on the concern about the overestimation of reliability with video-recording, a more definitive and applicable study estimating reliability should compare agreement between two bedside observers rather than two videotape reviewers. At the very least, if videotaping is to be used for an estimation of reliability in a large study, its concordance with actual bedside

examination should be proven in a representative subset of patients.

6.6 Sample Size Requirements

This pilot study gave estimates for the observed and expected agreement, allowing calculation of the sample size requirements for a larger reliability study of bedside tests determining Kappa within a reasonably small confidence interval such as ± 10% or ± 20%. From a table such as the one listed in Table 23, a decision can be made about the ultimate sample size for a study based on a trade-off between what is practically feasible and the desired precision for Kappa.[4]

It is interesting to note the rapidly escalating sample size requirements for decreases in the size of the confidence interval around the Kappa estimate. For example, N increases four-fold by decreasing the confidence interval size from κ ±.20 to κ ±.10. and again another four-fold by decreasing the confidence interval size from κ ±.10 to κ ± 05.[5]

---

[4]In addition to being useful for planning reliability studies, this simplified sample size table could also be useful in interpreting published studies. After locating the row corresponding to the simple agreement and Kappa from the study, the ± delta column is then chosen based on the approximate sample size.

$$N = \frac{1.96^2\, Po(1-Po)}{\delta^2(1-Pe)^2}$$

[5]Recall the sample size equation:

Another useful observation is the differing sample size requirements for tests with inherently different Kappa values. For a given desired precision and the same observed level of agreement, the sample size increases as the Kappa is lower. In other words, the less reliable test requires a larger sample size to get the same precision on the estimate.[k] For example, assuming

---

From the equation we see that,

$$N \propto 1/\delta^2$$

therefore as N increases fourfold when δ changes from .20 to .10 {i.e. $(1/.20)^2 = 25$ to $(1/.10)^2 = 100$ } and another fourfold when delta changes from .10 to .05. {i.e $(1/.10)^2 = 100$ to $(1/.05)^2 = 400$}.

[k]Recall that based on Cohen's approximation:

$$N \alpha \frac{Po(1-Po)}{(1-Pe)sup2}$$

Rearranging we get:

$$N \alpha \frac{\frac{Po(1-Pe)-(Po-Pe)}{(1-Pe)}}{1-Pe}$$

Rewriting in terms of P_o and Kappa gives:

$$N \alpha \frac{Po(1-K)^2}{(1-Po)}$$

Therefore for a fixed values for δ and P_o,

$$N \alpha (1-K)^2$$

the same observed agreement (P_o) and same desired precision (δ), the N required for a test with a true (inherent) Kappa of .30 is approximately twice that of a test with a true Kappa of .50 and approximately thrice that of a test with a true Kappa of .60.

With the average values of P_o and P_e as the best estimates of P_o and P_e, the required sample sizes were high (> 200) for voice quality, voluntary cough quality, pharyngeal sensation and ice swallow to obtain a Kappa within a confidence interval of ± .10.

If the standard for reliability were less strict with a Kappa of > .4 (moderate reliability as defined by Landis and Koch[14]), then the required sample sizes would be much less. If a Kappa within a larger confidence interval of ± .20 were acceptable, then the required sample sizes would approach the minimum for the gag reflex and the water swallow but would still be high (> 200) for pharyngeal sensation and the ice swallow.

These results demonstrate how the confidence limits chosen for a larger study with the Kappa statistic depend on the estimated Kappa. For example, if a pilot study showed that the Kappa was excellent (Kappa > .80), then a smaller sample size (closer to the minimum $16a^2$ required for the large sample approximation to hold) with a confidence interval of ± .20 might be sufficient because the Kappa would still be interpreted to reside in the substantial reliability range (κ>.60). On the other hand, if the Kappa estimate is in the range of .4 to .6, then a larger sample size would be needed to get a smaller Kappa

confidence interval to determine whether the actual Kappa is above a minimum standard for Kappa (such as $\kappa > .60$)

6.7 Potential Biases and Limitations of the Pilot Study

Several potential biases in this pilot study need to be addressed. Because the researcher did both the data collection and the video-recording in this pilot, there was a potential for observer bias. It could be argued that the video-recording of the physical examination could have been affected by preconceived notions from the baseline clinical data about whether the patient was at risk of aspiration pneumonia. For example, if more severe stroke patients were considered to have a higher risk of aspiration, then the researcher could have spent more time trying to detect swallowing difficulty in more severe patients. This expectation bias[80] was addressed by standardizing the video-recording protocol with identical verbal prompts and observations. Although completion of baseline data entry sheets was done after the video-recording had been done on day two, the researcher did review the chart to ascertain eligibility. If the purpose were to isolate the individual contributions to the diagnosis of the baseline characteristics and physical examination, the personnel doing the video-recording of the bedside examination should be blinded to the clinical information about the patient. In a therapeutic trial done by LeWitt et

al.[66], the physician doing the videotaped examination was aware of clinical status of the patient raising the question of whether the unblinded physician examiner could have brought subtle differences to the videotaped neurologic examination. This problem has also been recognized in studies of reliability in psychiatric diagnosis where there has been concern that the interviewer being videorecorded may, by his or her behaviour or phrasing of questions, cue the reviewers of the videotape.[69]

Observer experience and training may affect the estimates of agreement.[70] In this pilot study, both observers were physicians with general medical practice experience but different current clinical work. They were given the same amount of training beforehand. For the reliability estimates in the grant-funded study, it would be useful to assess reliability of the test methods in clinicians who would potentially use the bedside tests in actual clinical practice. Nurses in particular could be the key decision makers for feeding stroke patients and for referral to speech-language pathologists. Use of a quick and simple (rather than a long and detailed) training program would be the most relevant. A rigorous training program might increase reliability for research purposes but would not be practical if and when bedside tests were promoted for use in the clinical setting.

The setting of a study may affect the applicability of the results to other centers. This study was done at an urban community hospital over 30 km from the nearest tertiary care center. Patients usually came directly through the emergency room. Because this pilot study was based in a community hospital, it may have had a spectrum of patients different from that of a tertiary care hospital, although this is arguable. In a multi-center study, it would be useful to include a tertiary care hospital to address this concern.

Spectrum bias affects estimates of reliability as it does to estimates of accuracy.[81,82] The highest Kappa values occur when a test is applied to a group of subjects with findings evenly split between either clearly abnormal or clearly normal. P_o will be high and P_e will be approximately 50%. Lower Kappa values occur when a test is applied to patients who come from the middle of the spectrum with few patients who are clearly normal and abnormal because P_o will be lower. If all patients are at one end of the spectrum (either all patients mostly normal or all patients mostly abnormal), then P_o will be high but P_e will also be high. The most clinically relevant measurement of reliability is obtained by evaluating a realistic spectrum of patients representing the target population in whom the test is to be used in clinical practice.[83]

6.8 Potential Clinical Usefulness of Bedside Tests and Implications for Future Research

As noted in a review by Koran[70], the reliability of clinical methods is frequently moderate if determined at all. Looking at the reliability of the neurological examination in stroke patients, for instance, Shinar et al.[84] estimated inter-observer reliability to range from poor ($\kappa < .20$) to good ($\kappa > .60$) for a variety of historical and physical examination findings. Unfortunately, the study included only seventeen patients. As another example, Spiteri et al.[85] found the inter-observer reliability for respiratory physical signs to range from poor ($\kappa < .20$) to fair ($.20 < \kappa < .40$) in a study of 20 patients. In this decade, there has been renewed interest in the rational use of the clinical examination[86] and calls for more research on inter-observer reliability.[87]

After identification of the bedside tests and observations that have good reliability, the next question is which of the reliable tests have accuracy. Although 100% sensitivity would be desirable, it likely comes at the cost of low specificity. As a matter of clinical judgement, a bedside screening tool for aspiration with less than 100% sensitivity might still be acceptable as the consequences of missing a patient with silent aspiration leading to pneumonia might not be fatal. In fact, some investigators have used a history of pneumonia in the acute stroke phase as a predictor of aspiration on videofluoroscopy,

recurrent pneumonia, and need for swallowing therapy or tube feeding.[9] As well, the treatment for aspiration is proven to be highly effective. With a reasonably sensitive (perhaps 90%) screening bedside test or prediction rule, a detected patient could be referred for a more specific test such as modified barium swallow videofluoroscopy or, if validated in the future, fiberoptic endoscopy.

The clinical observations and bedside tests would perform differently in strokes of different severity. In low risk stroke patients (for example, based on the Barthel index), the negative predictive value of a bedside test would be expected to be high. In high risk patients (for example bilateral stroke patients), the negative predictive value would be lower but the positive predictive would be higher.[16]

Review of the published evidence for tests of aspiration and dysphagia in stroke showed that most of the tests were studied in rehabilitation groups. Sox has used the concept of spectrum bias to explain why accuracy described in published studies may not apply to the relevant clinical situation. Typically, reports of a clinical test begin by testing the "wellest of the well" and the "sickest of the sick". This is important at the outset to make sure the test, at the very least, is positive in obvious disease and negative in health. Subsequently, the test is studied in patients referred for the gold standard test and it is in this stage that most of the studies appear in publications. However, the patients are usually sicker than the patients in the

clinically relevant population and sensitivity is usually overestimated.[88] Therefore, given the modest sensitivity of the bedside tests found in stroke rehabilitation patients, the sensitivity for aspiration would be expected to be even worse in the clinically relevant population of acute stroke patients. Thus, a single bedside test is unlikely to be sensitive enough to be useful in acute stroke patients.

Horner et al.[89] used two bedside clinical variables (the voluntary cough and the pharyngeal gag reflex) to categorize 38 bilateral stroke patients into risk groups for aspiration. With modified barium swallow videofluoroscopy as the gold standard, they found patients with a normal gag reflex and voluntary cough had a low risk (14%) of aspiration compared to a moderate risk (approximately 50%) for patients with either of the two abnormalities and a high risk (87%) for patients with both an abnormal gag reflex and an abnormal voluntary cough. Although flawed because of its small size and lack of blinding between the bedside and videofluoroscopic examination, this study suggests that a combination of bedside tests might be a more powerful predictor of aspiration than a single test alone.

Based on the lack of published evidence for a single bedside test to accurately predict aspiration and dysphagia, it seems justifiable to propose a study designed to develop a clinical prediction rule. A clinical prediction rule has been defined as a decision-making tool that includes three or more variables obtained from history, physical examination, or simple diagnostic

tests and that provides a high probability of an outcome or suggests a diagnostic or therapeutic course of action.[60]

Many clinical prediction rules have been developed.[90] For example, Heckerling et al.[91] derived a clinical prediction rule from data on coexistent medical conditions, symptoms and signs to predict chest radiographic infiltrates (as a clinical surrogate for pneumonia). A stepwise logistic regression analysis created a rule that was then tested in a validation set. Analogously a clinical prediction rule for aspiration in stroke patients could combine clinical features and bedside tests to determine risk of clinical pneumonia. Such a prediction rule could assist in deciding which patients should be referred for more specific evaluation of swallowing and airway control followed by therapeutic maneuvers such as control of dietary consistency and non-oral feeding.

A clinical prediction rule is best composed of reliable clinical observations.[60] A field study confirming the reliability of clinical observations gives a sound basis to the derivation of a good clinical prediction rule[92] and is especially important when the observations, such as bedside clinical tests, require subjective interpretation.[55]

6.9 Clinical Utility

Even if a test or prediction rule is reliable and accurate,

if it does not affect patient management, either by determining therapy or prognosis, then it is probably not worthwhile. Diagnostic testing should ultimately affect the care of the patient. An accurate and reliable stratification of acute stroke patients for risk of aspiration and dysphagia could, at the very least, be a tool for further research on management. The highest level of evidence for a particular management strategy would come from a randomized controlled clinical trial. Such a trial could compare usual care with an intensive diagnostic and therapeutic approach, (which could, for example, include reliable and accurate diagnostic methods initiating speech pathology rehabilitative measures and/or non-oral feeding). The relevant outcomes should be medical complications (such as pneumonia, malnutrition, disruption of oral feeding), mortality, quality of life and costs. Smithard[6] recently emphasized the need for such a trial. In their commentary on a recent randomized controlled trial of percutaneous gastrostomy feeding versus nasogastric feeding for stroke patients, Koretz and Cook[93] emphasized the difficulty in currently establishing an evidenced-based policy for feeding patients with dysphagia and stroke. They recommend that clinical management trials be ranked high in priority on stroke research agendas.

Simel and Rennie [87] have recently underlined the need for more research on the clinical observations that underly clinical management trials. They make the observation that, on the one hand, great effort has been given to carrying out management

trials with reliable and valid outcomes. On the other hand, relatively little effort has been given to determining the reliability and validity of the clinical observations and diagnostic processes that identify the patients as eligible for the trial. They note that generalizability of the results of treatment trials are dependent upon the reliability and accuracy of the clinical examination.

6.10 Using the Pilot Study

The pilot study provided the groundwork for a more definitive study of the diagnosis of aspiration in acute stroke. The systematic review of the literature, the protocol with refinements based on the pilot study experience, and the sample size information were used to develop a detailed study proposal which is outlined in the next chapter.

7. Research Grant Proposal for a Study of the Reliability and Accuracy of Bedside Tests of Aspiration in Acute Stroke

7.1 Introduction

7.1.1 Statement of the Problem

Stroke is an important cause of mortality and morbidity. It is the third leading cause of death and an important cause of admission to hospital and need for long-term care.[2] A frequent complication of stroke is disordered swallowing and airway control.[1] For example, Gordon et al.[18] prospectively studied 91 consecutive acute stroke patients within 96 hours of admission and found 45% to have difficulty swallowing 50 ml of water. Disordered swallowing and airway control may lead to dysphagia and aspiration. Dysphagia, defined as difficulty swallowing ingested liquids and solids, may precipitate aspiration and disrupt oral feeding. Aspiration, defined as entry of ingested material into the lower respiratory tract below the true vocal cords, may lead to respiratory tract obstruction or infections.[4] In the same study of Gordon et al.[18], 12%(11/91) of acute stroke patients developed respiratory tract infections. Spurr and Valencia[5] found 7% (31/469) of acute stroke patients admitted to a neurology ward developed pneumonia within a follow-up of 18

months. Vitanen et al.[6] studied causes of death at six months in stroke patients and found pneumonia to be the third most common (responsible for 15% of deaths).[·]

A recent clinical practice guideline by the U.S Department of Health and Human Resources recommended that acute stroke patients have their swallowing assessed before oral intake is resumed. This particular guideline was based more on strong panel consensus than on research evidence.[7] The diagnostic strategy, including the appropriate use of radiologic studies, and the efficient involvement of speech-language pathologists, depends on bedside assessment of stroke patients. [8,9]

Many bedside tests have been suggested to screen for aspiration risk in acute stroke patients including tests of voice quality, voluntary cough quality, gag reflex, pharyngeal sensation and swallowing. [1,10,11,15] The current gold standard test is modified barium swallow videofluoroscopy in which barium entry into the lower respiratory tract during swallowing is considered abnormal.[4,14] Because of concern about the accuracy of bedside tests for aspiration risk, liberal use of modified barium swallow videofluoroscopy for screening in stroke patients has been advocated.[3] However, aspiration on modified barium swallow videofluoroscopy is a surrogate for the clinical outcomes of aspiration (with either respiratory tract infections or obstruction). Some investigators have questioned the value of

---

· Heart disease caused 35%, recurrent stroke caused 25% and pulmonary embolism caused 10%.

routine videofluoroscopy in screening acute stroke patients.[8] Modified barium swallow videofluoroscopy is time-consuming, expensive and not universally available for screening all stroke patients. The detection of patients at risk of aspiration in acute stroke currently relies on bedside clinical observations for which there is limited evidence.

7.1.2 Rationale for the Current Study

For a test to be useful, it should be both reliable (precise) and accurate (valid). For bedside tests with dichotomous results, reliability can be assessed with the Kappa statistic, which can be estimated with a confidence interval that depend on the sample size. With guidelines for the interpretation of Kappa suggested by Landis and Koch,[11] reliable tests can be identified. Reliable tests merit further study for accuracy (validity) and clinical utility.

The currently used bedside tests for aspiration have not been examined for reliability. None has been sufficiently accurate when used alone to predict aspiration on modified barium swallow videofluoroscopy or an eventual clinical outcome such as pneumonia. A validated prediction rule is a decision-making tool that includes three or more variables obtained from history, physical examination, or simple diagnostic tests that can provide the probability of an outcome or suggest a diagnostic or

therapeutic course of action.[60] A prediction rule for aspiration, based on clinical variables including the bedside tests, could be used to make an evidence-based guideline for safely feeding acute stroke patients.

7.1.3 Pilot Study

A pilot study done, at the Etobicoke General Hospital in 1995-96, was conducted to provide statistical and feasibility information for this study grant proposal. In this pilot study, the bedside tests listed above were video-recorded in 30 consenting acute stroke patients. This pilot study showed the feasibility of a more definitive study and gave valuable experience in recruitment, data collection and data analysis. It provided crude estimates of observed and chance expected agreement needed to estimate the required sample sizes to measure inter-observer reliability with the Kappa statistic with a desired confidence interval. The pilot study gave estimates of the approximate number of acute stroke patients admitted, and the proportion who are eligible and consenting during a ten month period at one community hospital center. It attempted to estimate the proportion of recruited acute stroke patients with the clinical outcomes of pneumonia and disruption of oral feeding. The pilot study also provided video-recordings that could be used in observer training for the bedside tests in the proposed study.

7.2 Objectives

1. To estimate inter-observer reliability (with the Kappa statistic) for each of the bedside tests of aspiration.

2. To estimate accuracy by determining sensitivity and specificity of each of the bedside tests of aspiration against the gold standard of modified barium swallow videofluoroscopy.

3. To derive a clinical prediction rule for aspiration on modified barium swallow videofluoroscopy based on reliable bedside tests and baseline clinical variables.

7.3 Methods

7.3.1 Setting

The pilot study suggested that more than one hospital center would be needed for sufficient recruitment within a time-frame of two years. The three Saskatoon hospitals will participate:
1) Royal University Hospital - a tertiary care teaching hospital,
2) Saskatoon City Hospital - an urban community hospital,
3) St. Paul's Hospital - an urban community hospital.

3)St. Paul's Hospital - an urban community hospital.
Because all three now come under the Saskatoon District Health Council, the principal investigator will be fortunate in having consulting privileges at all three institutions.

7.3.2 Study Group

7.3.2.1 Inclusion criteria

Patients with stroke, defined as a neurologic deficit of presumed vascular etiology persisting for more than 24 hours, will be eligible. They must have had the onset of the stroke within 96 hours[m] before admission to hospital.

7.3.2.2 Exclusion criteria

Clinical exclusion criteria will include: pre-existent dysphagia; non-enteral feeding method or tracheostomy, co-existent pneumonia (defined in 7.3.6.9 below) at the time of identification, decreased level of consciousness persisting

---

[m] Time interval increased (compared to the pilot study) to improve recruitment. Some acute stroke patients may be admitted to peripheral hospitals and then transferred a day or two later to the urban centers for neurologic consultation and CT scanning.

beyond 96 hours since stroke onset (defined as a lack of awareness such that the patient is unable to move the lips to the edge of a cup), nothing by mouth (NPO) status beyond 96 hours, terminal palliative care and pregnancy (because of the radiography).

7.3.2.3 Recruitment of patients

New hospitalized stroke patients admitted to hospital with a diagnosis of stroke will be identified by reviewing the admission lists in the emergency room and the medical wards, and by asking the charge ward nurses for the names of patients who may have had a stroke. This will occur every morning. A poster will also be placed at each nursing station giving a telephone number to a voice mail message system. A study nurse will review the charts of identified stroke patients to ascertain eligibility. Patients who have an acute neurologic deficit of presumed vascular etiology less than 24 hours old (by definition a transient ischemic attack) will be reassessed for eligibility the next day. Patients who initially have decreased level of consciousness or are being kept nothing by mouth (NPO) will be reassessed daily until 96 hours after stroke onset to determine subsequent eligibility. The consent for participation will be obtained by the study nurse from the patient or next-of-kin.

The pilot showed the importance of timely identification of

patients (necessitating daily recruitment rounds including weekends). During the pilot, the recruitment and consent process was time-consuming underlining the need to have sufficient study personnel to identify patients and obtain consent. An individual physician with other clinical responsibilities would not have sufficient time for this task. Having three study nurses, one at each hospital, will help to avoid missing eligible patients. These nurses will cover each other's hospitals on weekends and days off. They also may need to be available for evening meetings with family during visiting hours to obtain consent from the patient or next-of-kin. Reasons for non-enrollment of stroke patients will be categorized into: (1) non-acute stroke defined as stroke occurring more than 4 days before admission, (2) ineligibility based on clinical exclusion criteria (pre-existent dysphagia, coexistent respiratory symptoms and signs of pneumonia, persistent decreased level of consciousness, kept nothing by mouth (NPO) by personal physician), terminal palliative care, pregnancy) (3) inability of the research nurse to contact the patient or next-of-kin for consent within 96 hours of stroke onset and (4) refusal to consent for participation in the study. A log of identified stroke patients will be kept by the study nurse based at each site. All patients entered in this log book will entered in a data base that will include information on the baseline clinical features, Barthel score and chart review follow-up for pneumonia or disruption of oral feeding. This information will be used to compare characteristics

of the enrolled patients with the non-enrolled patients. It could also be used to estimate the frequency of pneumonia and disruption of oral feeding prospectively in all stroke patients.

7.3.3 Study Design

The study will be prospective with a) blinded determination of inter-observer agreement for the bedside test results and b) blinded comparison of each of the bedside tests with the primary outcome of barium aspiration on videofluoroscopy and secondary outcomes such as pneumonia and disruption of oral feeding.

7.3.4 Candidate Bedside Tests

A review of published studies, found by MEDLINE searching for articles on dysphagia and aspiration in stroke, suggested several simple bedside tests: gag reflex, pharyngeal sensation, swallowing, cough quality and voice quality. These studies are summarized in Table 1. The tests of pharyngeal sensation and ice swallowing were judged to have the least potential reliability based on the pilot study experience. In particular, they had a low prevalence of abnormality (below the 5% cutoff recommended by Grove et al.[69]), which made the Kappa values unstable. The test

of pharyngeal sensation was difficult or impossible in patients with stroke-induced communication problems, which is not uncommon: four of 30 (15%) patients in the pilot study had aphasia. A change in the procedure was made to allow a non-verbal response to pharygeal sensory testing. The ice swallow test was not considered feasible and abandoned. The lack of an ice machine with an automatic crusher at Etobicoke General Hospital might easily be the case for other hospitals. There was inconsistency in the size of the ice chips made at the bedside.

7.3.5 Protocol Overview

Starting between 48 hours and 96 hours after the onset of stroke, alert patients will be examined once by each of two study nurses independently for: voice quality, voluntary cough quality, palatal gag reflex, pharyngeal gag reflex and pharyngeal sensation. The ability to swallow 50 ml, 100 ml and 150 ml of water will also be recorded.

The baseline clinical data will be documented by review of the hospital record and talking with the primary care nurses before doing the bedside tests. The Barthel index will be completed immediately after the bedside testing. Then, the nursing staff caring for the patient, blinded to the results of the clinical bedside tests will be asked for their impressions about whether the patient is aspirating or having dysphagia.

Decisions about feeding will remain with the medical and nursing staff, who will be blinded to the bedside test results.

All patients will undergo an abbreviated modified barium swallow videofluoroscopy within 24 hours of enrollment in the study. Although this may entail doing the procedure on the weekend, it is important to keep this time interval short because the patient's clinical status could change over time.

Patients will be followed for one month after the onset of the stroke. The hospital records will be reviewed until the date of discharge, or until one month after admission, for any note of symptoms and signs of pneumonia (e.g. productive cough, shortness of breath, fever, new crackles), feeding disruption (e.g. nasogastric tube feeding) or feeding difficulty (e.g.choking or coughing with feeding). For patients discharged before one month after stroke onset, study personnel will telephone the patient or next-of-kin at home, or the nursing staff at a convalescent or chronic care facility to determine whether there had been symptoms of pneumonia, feeding disruption, or feeding difficulty within the first month.

7.3.6 Specific Procedures

7.3.6.1 Baseline Assessment

During daytime hours (0700 to 2100) the baseline assessment

will completed by the study nurse obtaining the consent for participation (the "recruiting nurse") before the bedside testing. An abbreviated neurologic examination will include observations of speech (any receptive, expressive, and dysarthric abnormalities or none), cranial nerves for facial strength (rated as any droop or none) and motor strength graded on the Medical Research Council five point scale[n] of both upper and lower extremities.[95] There will be an examination of the lung fields to rule out dullness on percussion, and crackles or decreased breath sounds on auscultation. The patients will be interviewed and the hospital chart reviewed for demographic data (age, sex), stroke risk factors, use of current sedative medication and intravenous fluid therapy beyond 48 hours after stroke onset. The physician and nursing progress notes will be reviewed for any mention of decreased level of consciousness during the 96 hours since stroke onset.

The "recruiting nurse" will then do bedside tests as described below in the same order followed by the scoring of the Barthel index. Then the nurse will ask the nursing staff directly caring for the patient "Do you think this patient is aspirating with oral intake?" and "Do you think this patient is having difficulty swallowing?"

A second nurse will then be called by the recruiting nurse

---

[n] The Medical Research Council classification is as follows: 0, no movement; 1, trace of contraction; 2, active movement when gravity is eliminated; 4, active movement against gravity and resistance; 5, normal strength.

to repeat the same procedure. This nurse will visit within 36 hours.

The results of the chest radiograph and CT brain scan, done within three days of admission, will be recorded approximately one week later when final radiologic reports become available. The chest radiograph will be used to confirm the patient was free of pneumonia at the time of admission. The CT brain scans will be used as a variable to estimate stroke severity. CT brain scans will be categorized by whether or not there are findings of stroke. CT brain scans with stroke findings will be subcategorized by stroke type (infarction, hemorrhage, and lacune) and stroke location (cerebrum, cerebellum, and brainstem).

7.3.6.2 Voice Quality

The quality of the voice will be recorded as either normal or abnormal (wet, harsh, hoarse) in response to the request to "Say pa, pa, pa; la, la, la; and ga, ga, ga."

7.3.6.3 Cough Quality

The patient will be asked to give a voluntary cough. The specific request will be *"Please give a strong cough. Give a*

really strong cough."

7.3.6.4 Pharyngeal Sensation

Pharyngeal sensation will be tested by applying the tip of a tongue blade once on each side of the pharyngeal wall at the proximal tonsillar pillar. The request will be: "Nod if you feel it." The blade will touch first on the left, then on the right, then on the right, and finally again on the left. The sensation will be rated normal if the patient nods for each touch of the tongue blade.[c]

7.3.6.5 Gag Reflexes

The palatal gag reflex will be tested with a tongue blade pressed gently, briefly, but firmly on the soft palate in the midline just anterior to the uvula. The pharyngeal gag reflex will be tested by touching the posterior pharynx just inferior and posterior to the soft palate. For each of the two gag reflexes, an intact reflex will be defined by both palatal elevation and gagging phonation.

---

[c] This procedure is a modification from the pilot study in which the request was a non-specific: "Let me know if you feel it when I touch you.

7.3.6.6 Water Swallowing

Swallowing will be tested with the patient sitting upright (>80 degrees if confined to bed). The patient will be observed for 15 seconds between parts of the test and for 1 minute after the tests for any choking, coughing and "wet hoarseness". The patient must be alert enough to have spontaneous eye opening, appear responsive to environmental stimuli (follow verbal command to *"open mouth"* or non-verbal prompt with imitation of examiner) and be aware enough to move the lips around the spout lid opening. The water will be given in a plastic cup with a spout lid commonly used for young children. The spout lid will have a small hole in it so that the water comes out at a slow and steady rate avoiding the bolus of water that could make a false positive bedside test. Initially 50 ml, then 100 ml and finally 150 ml of water will be given with the request to *"Swallow this water"*. After each swallow of water, the patient will be observed for 15 seconds and then requested to "Say la, la, la". If the patient is unable to hold the cup, the examiner will hold it to the patients lips delivering the liquid over about 1 minute. All patients will receive the 50 ml swallow test. If the patient has respiratory distress with 50 ml, the 100 ml and 150 ml tests will not be administered. The water will be taken from the cold tap without running it for more than a few seconds to obtain room temperature rather than cold water as cold water can affect the triggering of the swallowing reflex.

A normal water swallow will be defined as the absence of choking, coughing and wet hoarseness on all of the three volumes of water. For all of the above tests, if the patient does not speak English, a translator will give the instructions.

7.3.6.7 Stroke Severity

The Barthel Index will be completed immediately after the bedside tests by the study nurse by reviewing the hospital chart, interviewing the patient and talking to the nursing staff. Of the general disability scales and disease specific instruments advocated to measure the severity of acute stroke, the Barthel index was chosen because of its validity and reliability.[47] This disability scale is quicker to administer than the disease specific scales and therefore potentially more acceptable as part of a screening tool.

7.3.6.8 Modified Barium Swallow Videofluoroscopy

All enrolled patients will be examined by one of a team of speech-language pathologists and undergo modified barium swallow videofluoroscopy. The modified barium swallow videofluoroscopy done for the purposes of this study will be limited to a determination of whether or not there is aspiration of barium.

In the examination, patients will be given 5 ml of thin barium, then 20 ml of thin barium and finally 30 ml of thin barium. Patients will be seated upright and viewed from the lateral aspect. For patients who aspirate, the presence or absence of a cough will be noted. This examination, which can take as little as two minutes, will be videotaped as is the usual practice. The type of aspiration will be categorized as: a) any and b) large (greater than 10% of barium). For this study, the presence of any aspiration of barium below the true vocal cords will define aspiration on the gold standard test.

7.3.6.9 Clinical Follow-up

The clinical outcome of primary interest is pneumonia. Outcomes also of interest are: disruption of oral feeding, symptomatic dysphagia, and one and six month mortality.

All study patients will have a follow-up chest radiograph at two weeks after the onset of the acute stroke. A radiographic infiltrate generally remains apparent for several weeks (and may take up to six weeks to clear completely) after onset of pneumonia.[96] A timepoint of 2 weeks after stroke onset was chosen because it would capture infiltrates developing during the first and second weeks during which time feeding should have been reestablished. Choosing this timepoint would also be efficient in that the majority of patients would still be in hospital. For

example, the pilot study showed that for the enrolled patients the median length of stay was 23 days.

Independent study personnel will be responsible for the follow-up for clinical outcomes. Similar to the prospective study by Holas et al.[48], the follow-up will be by chart review. If the patient is discharged before four weeks after the onset of stroke, follow-up will be by telephone calls, at two and four weeks and six months after stroke onset, to the patient (or the next-of-kin) at home, or to the patient (or health care personnel) at a chronic care or rehabilitation institution. During the telephone calls, the patient, or next-of-kin will be asked "Has there been any cough, ... shortness of breath, ... fever, ... difficulty swallowing or ... any need to seek medical attention?" If any symptoms are present, the patient will be requested to see his or her personal physician. This physician will be contacted by study personnel and then recontacted one week later for the details of the evaluation.

For research purposes, pneumonia is commonly defined as a radiographic infiltrate with any two of the following criteria: productive cough, fever > 38.3C, positive sputum culture or Gram's stain or leukocytosis (WBC >13).[3] Although this definition can be inaccurate (some of the apparent pneumonias may actually be pulmonary atelectasis or infarction), more invasive techniques are considered inappropriate because of the associated morbidity, inconvenience and research costs. Some researchers have used less rigorous definitions for pneumonia. For example, Schmidt defined

pneumonia as any three of following: productive cough, fever [> 38.3C], positive auscultatory findings on chest exam, PaO2 below 70 or a 10 mmHg decrease from baseline, positive sputum culture or purulent sputum Gram's stain, and chest radiographic infiltrate.[52] With these definitions for this study, the finding of "presumptive pneumonia" will be made with the former definition (requiring the presence of a radiographic infiltrate) and "probable pneumonia", with the latter definition. Korner-Bitensky et al.[64] have shown that telephone interview follow-up of health status in stroke patients has agreement with face-to-face interviews.

Disruption of oral feeding (such as use of non-oral feeding methods including nasogastric, percutaneous gastrostomy [PEG] and gastrojejunostomy enteral feeding, and intravenous parenteral nutrition) will be a secondary outcome and will also be documented by review of the patient record and telephone follow-up. Although monitoring body weight, arm muscle circumference (a measure of lean muscle mass) and triceps skin fold thickness (a measure of body fat) as indicators of determining nutritional status,[97] has been done in previous studies, this will not be included. Based on the pilot study, an admission body weight is frequently impossible because a portable scale for bedridden, non-ambulatory patients is unavailable. The resources required for a follow-up visit to repeat the measurements of arm muscle circumference and triceps skin fold thickness would add to the costs of an already labour intensive protocol.

7.3.7 Blinding Strategy

7.3.7.1 Baseline Data Collection and Bedside Tests

An important design issue is whether or not to isolate the baseline clinical data collection (including medical history and neurologic exam findings), Barthel index and bedside tests from each other. On the one hand, it theoretically would be useful to examine the contribution of individual independent variables to a prediction model. This would also reduce "expectation bias" on the part of the observers. On the other hand, predictor variables are not elicited in isolation in clinical practice. Therefore, for a prediction model with applicability to clinical practice, the personnel making the observations will be aware of the baseline clinical data. Previously derived and validated prediction rules have not isolated the bedside tests from other clinical information.[98]

7.3.7.2 Blinding of Videofluoroscopy

The videofluorscopy will be done by speech pathologists blinded, at the time the procedure is done, to the baseline

clinical information and the bedside tests. The results of the videofluorscopy will not be made available to the health care team. However, the health care team will be free to order the standard modified barium swallow videofluoroscopy as indicated.

7.3.7.3 Blinding of Clinical Outcome Determination

The follow-up for "probable" or "presumptive" pneumonia and disruption of oral feeding will be obtained by review of the chart and, if necessary, contact with the patient, the patient's next-of-kin, or staff at the chronic care or rehabilitation hospital or at home. This will be done by study personnel not participating in the recruitment and bedside testing.

7.3.8 Study Organization

7.3.8.1 Personnel

Nurses rather than physicians will obtain the baseline clinical data, do the bedside tests and score the Barthel index. Notwithstanding their non-availability given recent staff cutbacks, nurses may be more suited than physicians to collect this information in clinical practice. Compared to physicians, nurses spend more time at the bedside of the acute stroke patient

and the safety of feeding is an important nursing issue[99,100,101] It is conceivable that a derived and validated clinical decision rule (perhaps in a checklist format appended to the stroke clinical pathway form) might help nurses to decide about feeding and determine patient referral to a speech-language pathologist.

A speech-pathologist at each hospital with experience in videofluoroscopy will be sought. They could cross cover each other on weekends and holidays.

As was done in the pilot study, both a nurse and a speech-language pathologist will be sought as coinvestigators.

7.3.8.2 Initial Training

A two hour orientation session for the study nurses and speech-pathologists will give an overview of the 1) study design and the importance of blinding those doing the initial data collection from those doing the outcome determination, 2) patient recruitment with the consent process and, 3) specific protocol procedures. The study nurses will be shown an edited videotape, including candidate bedside tests (voice quality, cough quality, gag reflex, pharyngeal sensation, water swallow) for one of the two days for the thirty pilot study patients. They will rate the videotapes with the data collection forms in an unblinded fashion and discuss any disagreements. They will be given an opportunity to practice the bedside tests on the researcher and a consenting

patient on the ward. Video training has been used in the field of stroke research and been considered practical and effective for improving inter-observer reliability.[102] The video training session will serve to improve the inter-observer reliability of the study nurses. The video training will not be so intensive as to be impractical, and therefore nongeneralizable, for the ultimate application of a validated decision rule. Perhaps such a training video could be used in the dissemination of a validated decision rule.

7.3.8.3 Monitoring of Performance

The investigator will monitor recruitment and data collection quality (completion of data entry sheets for the baseline clinical information, bedside tests and Barthel index). The videofluoroscopic examinations and clinical follow-up will be monitored for data collection quality and timeliness.

7.3.9 Data Collection

7.3.9.1 Baseline Assessment

A standard form, refined from the pilot study, will be completed by a study nurse doing a baseline assessment.

7.3.9.2 Bedside Tests

A standard form will be completed at the bedside by each study nurse on the day of bedside testing. For each of the bedside tests, there will be a five category rating scale defining the certainty of the finding: definitely normal, probably normal, equivocal, probably abnormal and definitely abnormal. Below this rating on the form, there will be a dichotomous rating forcing a decision between normal and abnormal. This format will allow a determination of the number of "easy" cases and the number of "difficult" cases.

7.3.9.3 Modified Barium Swallow Videofluorscopy

The videotapes of the modified barium swallow videofluoroscopies on all enrolled patients will be edited and

transferred to subsidiary videotapes in random order. These will be shown to two blinded speech pathologists skilled in videofluoroscopy to determine inter-observer reliability. The final designation of aspiration versus no aspiration status will arbitrarily be taken as the majority decision of the three different observers (the speech-language pathologist actually doing the test and the two later reviewing the video-recording).

7.3.9.4 Clinical Outcome Determination

A standard form for recording the outcomes at two weeks and four weeks will be completed at these timepoints. Two blinded observers will review the charts to ascertain the same outcomes to allow inter-observer reliability testing of the outcome. The final designation for an individual patient on whom there is disagreement between the two outcome observers will depend on the decision of a third blinded observer. A second page in the form will be completed for patients requiring telephone follow-up because of discharge before one month. These telephone calls will not be repeated for the purpose of reliability testing to avoid inconvenience to the patients.

7.3.10 Sample Size

7.3.10.1 Reliability

The sample size calculation is based on the assumption of normality for the sampling distribution for Kappa. This assumption is reasonable if the sample is reasonably large i.e. ≥64 patients (See Appendix II). The sample size calculations assume that the crude estimates are reasonable. In the pilot, the video-recording method probably led to optimistic estimates of agreement because of the removal of variability related to the test stimulus, test environment and subject status. On the other hand, the observer perception may have been limited by the video-recording technique. It seems reasonable to treat the estimates as potentially optimistic and to be conservative on the sample size estimation i.e. treat the estimate as a minimum sample size required.

For sample size calculation aimed at reliability, a sample size of 200 patients is judged to be reasonable. Based on the crude estimates of P_o and P_e for the gag reflex and water swallow in the pilot study, the sample size required to obtain a Kappa within a ±.10 confidence interval would be 171 for both tests. Based on the crude estimates of P_o and P_e for voice quality and cough in the pilot study, inclusion of 138 and 96 patients respectively would be necessary to define Kappa within a confidence interval of ±.20. This means that a sample size of 200

confidence interval of ±.20. This means that a sample size of 200 patients would allow the conclusion that the true Kappa for each test would be >.40 if the Kappa estimate were >.60. A Kappa value > .40 signifies at least moderate reliability according to the criteria of Landis and Koch.[14]

7.3.10.2 Accuracy

Estimates for sample sizes required to determine sensitivity and specificity depend on the desired 95% confidence interval for the estimate and the actual estimate. Tables have been published[103] showing confidence intervals for various sample sizes according to the observed proportion, p, and the number of positive tests. For example if the sensitivity of a particular bedside test turns out to be 90%, based on 10 patients aspirating barium on videofluoroscopy, then 95% confidence limits for that estimate are quite wide (.54, .99). If the estimate is based on 100 patients aspirating barium on videofluorscopy, then the confidences are narrower (.82, .95).

Sample sizes have also been examined for studies aiming at likelihood ratios. Simel[104] calculated the sample size required for a negative likelihood ratio less than .50 with a confidence interval that does not include 0.50. When the sensitivity and specificity are known, the prevalence of abnormality affects the sample size required. For a test with an estimated sensitivity of

90% and a specificity of 50%, 160 patients would be required if the prevalence of abnormality were 50%. Similarly, 228 patients would be required if the prevalence of abnormality were 33%, and 370 patients would be required if the prevalence of abnormality were 20%.

7.3.10.3 Prediction Rule Derivation

The sample size required for a logistic regression can be estimated with the rule of thumb of at least 5 to 10 outcomes per clinical predictor variable.[30,75] Using modified barium swallow videofluoroscopy as a surrogate gold standard will likely provide more outcomes (aspiration of barium) than the use of only clinical events (pneumonia). Therefore, if the clinical prediction rule is made up of three predictor variables, 30 patients with aspiration on modified barium swallow videofluoroscopy would be required. Some published studies have shown that up to 42% of acute stroke patients have aspiration on videofluoroscopy.[3] Assuming this estimate, as few as 75 patients would be required using three clinical predictors. More conservatively, assuming only 15% of patients have aspiration on videofluoroscopy, 200 patients would be needed for three clinical predictors. Obviously more patients would be needed if more predictors were to be included in the model.

Use of clinical outcomes (presumptive and probable

pneumonia) would require a much larger study. Based on the pilot study's crude estimate of only 1 in 30 patients with a clinical outcome of pneumonia and the rule of thumb of 10 outcomes per predictor variable, a sample size of 900 patients would be required for three predictor variables. Unfortunately, when the probability for an outcome is small (say < .10), the sample size for a logistic regression is large (>500) even with a continuous variable.[125] This supports the decision to use a surrogate gold standard as the primary outcome for the proposed study.

7.3.11 Data Analysis

7.3.11.1 Reliability

Kappa will be calculated for inter-observer agreement for each of the bedside tests as well as dichotomous baseline variables such as presence or absence of past medical history and most neurologic exam findings. The confidence intervals will be calculated based on the standard error derived by Fleiss assuming a normal distribution for the Kappa estimate. The Kappa will be interpreted according to the guidelines of Landis and Koch.[14] Assuming a 95% confidence interval of ±.10, a Kappa estimate of .50 would indicate at least fair inter-observer agreement (true Kappa > .40).

Weighted Kappa will be calculated for inter-observer agreement of the multichotomous variable of muscle strength graded from 1 to 5. Intraclass correlation coefficient will be calculated for the inter-observer agreement of the continuous variable of the Barthel index.

7.3.11.2 Accuracy

Sensitivity, specificity and likelihood ratios (with their respective confidence intervals) will be calculated for the bedside tests against the current gold standard of modified barium swallow videofluorscopy. Sensitivity is the number of true positives over the number who have aspiration. Specificity is the number of true negatives over the number who have no aspiration.[p] The likelihood ratio of a positive test result is *sensitivity/(1 - specificity)* and the likelihood ratio of a negative test is *(1-sensitivity) over specificity.*[104, 105] The formula of a confidence

---

[p] The approximate 95% confidence interval for any simple proportion, p, drawn from a binomial distribution is:

$$p \pm 1.96\sqrt{p(1-p)/n}$$

For sensitivity, the confidence interval is:

$$sens \pm 1.96\sqrt{sens(1-sens)/(number\ with\ disease)}$$

For specificity, the confidence interval is:

$$spec \pm 1.96\sqrt{spec(1-spec)/(number\ without\ disease)}$$

interval for a likelihood ratio [q] has been given by Simel et al.[105]

7.3.11.3 Prediction Rule Derivation

Assuming there are enough outcomes (> 30 patients with barium aspiration on videofluoroscopy), multivariate techniques will be used to derive a prediction rule with three or more clinical variables. The multivariate model must account for binary (aspiration on videofluoroscopy and pneumonia) and must allow use of dichotomous (e.g. abnormal water swallow test) as well as continuous predictor variables (e.g. age and Barthel score).

Logistic regression is suited for models that contain dichotomous variables as well as continuous variables.[60] A logistic regression model will be created with bedside variables that have reliability (Kappa estimate > .60 with 95% confidence interval +/.20) and have statistical associations with aspiration on videofluoroscopy (P > .05). Chi-square tests will be done on dichotomous variables and Student's t test (with a two-tailed alpha of .05) on continuous variables. Stepwise

---

[q] The confidence interval for a likelihood ratio is:

$$\ln LR = \ln p_1 \text{ over } p_2 \pm 1.96 \sqrt{\frac{1-p_1}{p_1 n_1} + \frac{1-p_2}{p_2 n_2}}$$

logistic regression will be done with the BMDPLR computer program. Tests for interaction among variables will be done as well. In order to keep the model simple, three or four variables (depending on the number of outcomes that actually occur) with the most discriminatory ability will be chosen for the final model.

A recursive partitioning model will also be examined. It may lead to a rule with higher sensitivity but potentially less overall accuracy.[60] Recursive partitioning can lead to a concise presentation of the prediction rule as a decision tree.[106] More recently available software for recursive partitioning (e.g. KnowledgeSeeker)[107] is reported to be more user-friendly and better documented compared to previous software.[108,109]

Cross validation with the study data set itself will be done using the jacknife method.[110] The prediction rule's sensitivity, specificity, positive and negative predictive values with confidence intervals will be calculated.[111]

The choice of the ultimate multivariate model used (logistic regression versus recursive partitioning) will depend upon a judgment comparing the risk of false negatives (false negative bedside testing with eventual pneumonia) compared to the cost of many false positives (false positive bedside testing leading to inappropriate institution of nasogastric or PEG feeding). The choice of the ultimate model and variable used also depends on the sensibility of the prediction rule. The rule should be judged clinically reasonable by involving tests that reflect risk of

aspiration. It should be clinically meaningful by affecting patient management. The rule should also be clinically feasible by being quick, simple to do and practical (e.g. equipment, if required, is readily available).[60]

7.3.11.4 Comparison with Clinical Outcome

Both the clinical predictors and current gold standard of modified barium swallow videofluoroscopy will be compared to the eventual clinical outcomes of pneumonia. If there are a sufficient number of clinical outcomes (i.e. pneumonias), a prediction rule will be derived that includes videofluoroscopy as an independent variable as well. The follow-up for clinical outcomes may be of some reassurance for a less than perfect prediction rule. For example, if the rule fails to predict aspiration on videofluoroscopy in some patients, but these same patients remain free of symptoms of pneumonia on follow-up, this would be useful information. This is analogous to the situation that arises in attempts to rule out pulmonary embolism with non-invasive techniques (i.e. low probability ventilation/perfusion scans and negative serial venous doppler studies). Even though a signficant proportion (as high as 30%) of patients testing negative on non-invasive tests may still have had pulmonary emboli, very few patients (untreated) die in follow-up.[112]

7.3.12 Feasibility

The pilot study showed the protocol to be generally feasible but time-consuming. Refinements suggested by the pilot have already been mentioned. Eliminating the video-recording may increase the recruitment rate. In the pilot study, several patient's families declined participation because of unease with video-recording in spite of assurances that the field of view of the cameras would include only the lower face and neck.

A prospective study published during the pilot study, supports the feasibility of this proposed study. Smithard et al.,[8] enrolled 121 consecutive patients with acute stroke. Videofluoroscopy was done on all patients blinded to the results of a standardized bedside swallowing assessment. Clinical follow-up showed 41% (29/121) to have chest infections. Unfortunately reliability of the bedside tests was not measured. Although an abnormal bedside swallowing assessment was associated with chest infection, the individual tests were not examined for accuracy.

7.3.13 Safety, Ethical and Cost Issues

The risks of this study protocol are small. Although testing patients for water swallowing and gag reflex has the potential to cause aspiration, this aspiration is unlikely to be harmful. Immediate measures to protect the airway can be taken for

respiratory distress. None of the patients in the pilot study had adverse effects associated with the bedside examination. The amount of ionizing radiation associated with abbreviated videofluoroscopy and routine chest radiography is small.

Patients or next-of-kin will be required to give informed consent. The pilot study consent form will be modified to include the risk and benefit of the early videofluoroscopy and the follow-up chest radiograph. Although it could be argued that, because of the low risk of the protocol to the patients, consent is not necessary, this situation does not meet the guidelines set out by the FDA[72] for consent waiver.[f] Studies in acute stroke patients that have not used informed consent[113] have been criticized for being disrespectful.[113]

Another ethical issue is whether the results of the abbreviated videofluoroscopy (done as part of the protocol) can be kept from the clinicians and patient. On the one hand, keeping the clinicians blinded avoids the bias that might enter into the subsequent follow-up. If the clinicians were told that there was aspiration of barium, they might feel compelled to start initial rehabilitative measures (such as nasogastric or PEG feeding) that might not have otherwise been done. A non-oral feeding method might reduce the incidence of pneumonia but could also lead to complications (e.g. stomal infections). The unblinding might lead

---

[f] These guidelines are: 1)a life-threatening situation necessitating use of an intervention that is compared to a standard treatment 2)potential patients are incapacitated 3)no time to obtain consent from legal representatives 4)no equally effective alternative therapy with similar likelihood of success is available

to surveillance bias. For example, patients who have tested positive on videofluorscopy might be observed more carefully than those who tested negative. These biases could seriously undermine the validity of the study.

On the grounds that an invalid study is itself unethical, the blinding of videofluoroscopy is considered important. However, it is important to include the issue of blinding in the consent form so that the participants are aware the videofluoroscopy is being done without affecting the patient's management. The clinicians caring for the study patients are still free to order the standard modified barium swallow videofluorscopy based on their clinical judgement. There will be some potential benefit of the protocol for the individual patient. A radiographic consolidation found on the follow-up chest radiograph might lead to early treatment.

In the consent form, there will also be consent obtained for two possible telephone follow-up calls (should the patient be discharged before one month after stroke onset). The consent will mention that should the telephone call detect any symptoms, the patient will be advised to visit his personal physician from whom study personnel will obtain a clinical update.

Approval will be sought to collect information (baseline neurologic information, Barthel index, and follow-up chart review for pneumonia and disruption of oral feeding) in acute stroke patients who are not eligible, cannot be contacted or who decline consent for the study protocol.

The budget for the study will be, for the most part, made up of salaries for three (half time) study nurses for two years and three (1/8 time) speech-language pathologists. Because all three hospitals already have videofluoroscopy equipment, positioning chairs and videorecorders, the budget should only need to account material expenses of the videotapes and chest radiographic film. Other expenses will include: telephone charges (dedicated line with voice mail message system and long distance calling may be required for patients transferred back to a rural area), stationery, computer time, and bedside equipment costs (tongue depressors and plastic cups with spout lids).

7.4 Relevance

Two general approaches to providing evidence for the utility of a diagnostic test have been described.[115] In the "one step approach", a population is subjected to a randomized controlled trial comparing usual care with a management strategy based on a diagnostic test of interest. Examples of such tests supported by "one step approach" include: endoscopy versus radiology for acute upper GI bleeding[116] and fecal occult blood testing for colorectal carcinoma.[117]

In the "stepwise approach", firstly, the proportion of patients who test positive is estimated. Next, the ability of the

test to predict an adverse outcome is determined. Finally, the effectiveness of treatment is assessed by carrying out a randomized trial of patients who test positive.[115] This approach is used most frequently and may be more efficient than the "one step approach".[115] An example of this approach is blood pressure measurement. Population studies showed that a certain proportion of the population have hypertension and prospective studies showed that hypertension predicted adverse outcomes such as stroke. Randomized controlled studies have shown the effectiveness of treating hypertension in the prevention of stroke.[118]

This research protocol will take the "stepwise approach". The study will determine the proportion of patients who test positive with the bedside tests and modified barium videofluoroscopy in a recognizable patient population. It will try to use the bedside tests to predict which patients have aspiration on the current gold standard (modified barium swallow videofluoroscopy). It will also compare the bedside tests and the current gold standard to eventual clinical outcomes such as pneumonia. If a reasonable prediction rule can be derived, the next step will be to prospectively validate the rule in a new setting.

The proposed study may provide clinical predictors that identify patients at low risk of aspiration. This information might avoid unnecessary videofluoroscopy and referral to speech-language pathologists, which would have some potential for health

care cost savings. These clinical predictors may also identify a high risk group in which interventions (such as non-oral feeding methods or compensatory swallowing techniques) can be tested for effectiveness thereby preventing pneumonia. Subsequent studies would need to assess the utility of a validated clinical prediction rule in a new population. It is hoped that this research will ultimately contribute to the development of evidence-based guidelines for feeding acute stroke patients.

8. Conclusion

In summary, this thesis centered around an empirical pilot study of an important medical problem and explored some potentially useful approaches to assessing bedside clinical tests. The pilot study also showed the challenges of research in this area. The recruitment rate was lower than projected as many patients admitted with a diagnosis of stroke were non-acute. Obtaining consent was more difficult than expected. Many patients could not give consent themselves and the next-of-kin were difficult to contact. This pilot study also showed that the baseline rates of outcomes such as pneumonia and disruption of oral feeding were likely low, suggesting that a randomized controlled trial of a management strategy aimed at these clinical outcomes would need to be large.

The pilot study used bedside video-recording to facilitate independent blinded observations. The video-recording was technically feasible but the blinded review of the videotapes was arduous. Because the variability owing to observer perception and processing was isolated from the variability owing to the test stimulus, environment and subjects, it likely gave an exaggerated estimate of the reliability of these bedside tests. However, tests found to be unreliable using the video-recording technique need to be examined and either modified or discarded. For example, the pilot study suggested that tests of pharyngeal sensation and ice swallow belong to this category.

The issue of the sample size required for reliability studies of categorical tests was approached by creating a simplified sample size table to estimate Kappa within a given confidence interval. This pilot study provided the required sample sizes to determine Kappa within a 95% confidence interval of ±.10 and ±.20 for the bedside tests. The tables demonstrate how the decision about the necessary confidence interval depends on the Kappa estimate itself and the Kappa threshold chosen for interpreting acceptable test reliability.

The review of published studies, the pilot protocol experience (which led to refinements and modifications for feasibility), the sample size determination based on pilot study estimates, and the recruitment rates were used to support a detailed study proposal. This thesis succeeded in providing statistical and feasibility information for a study of the reliability and accuracy of the bedside tests for aspiration in acute stroke patients. It is hoped that this research will ultimately contribute to the development of evidence-based guidelines for feeding acute stroke patients.

Appendix I. Reliability and the Gag Reflex

The gag reflex is a bedside clinical test that depends on the neural pathways controlling swallowing and airway control. It is elicited by touching the posterior pharynx (either on the posterior soft palate or on the posterior tongue) with a tongue depressor and observing for palatal elevation, pharyngeal constriction and/or "gagging phonation".

If the gag reflex were a completely unreliable test, such as flipping a coin, then the test results obtained by two observers might only agree to an extent predicted purely by chance. If the chance of a gag reflex on any one viewing by an observer were 50%, then the chance of agreement by two similar observers would be 50%. The chance of agreement would be the sum of the probability of both observers noting a positive reflex .50 x .50 = .25 and the probability of both observers noting a negative reflex .50 x .50 = .25, which is .50 or 50%. Similarly, if the probability of a positive gag reflex on any one viewing by an observer were 75%, then the chance of agreement between two similar observers would be 62.5%. Finally, if the probability of a positive gag reflex by one observer is 90% and the probability of a positive gag reflex by another observer is 75%, then the agreement would be .70 or 70%.

If there is 100% agreement between the two observers, the test is perfectly reliable. The causes of less than perfect agreement can be categorized according to: 1) test stimulus, 2)

observer perception, 3) observer processing, 4) environment and 5) subjects.

Test Stimulus. The firmness with which, or the location where, the tongue blade touches the mucosa of pharynx, may vary.

Observer Perception. The observers may vary in their sensory apparatus. One observer might be less attentive or perceptive visually. The examiner can be interrupted, sleep deprived or bored.

Observer Processing. Observers may use different criteria for what an intact reflex is. For example, one observer might require that all three responses including palatal elevation, "gagging" phonation and pharyngeal constriction be present to call the gag reflex present. Another observer might call the reflex present if either one of the three responses is present. Training and experience likely affect observer processing.

Environment. The environment can influence the observer or the subject. Lighting may affect the observer's ability to see the subject's response and noise may distort auditory perception. The environment can distract the observer and the subject.

Subjects. The clinical status of a subject may change. The level of consciousness could be affected by cerebral edema, sedative medication, or sleep deprivation for the gag reflex done in acute stroke patients. The degree of cooperation may be affected by the patient's emotional state or speech comprehension. Even the subject's ability to open the mouth may change because of a fluctuation in speech comprehension.

Repetition may cause some learning that can increase or decrease the response. The actual degree of abnormality of the subject may also affect whether the test is likely to be reliable. A subject with a mild abnormality (for example a mild stroke) may at his or her best be normal but at his or her worst do poorly.

The population has two characteristics that can affect reliability. It has an overall prevalence of abnormality and a spectrum of patient severity. The prevalence of abnormality affects the marginal probabilities that in turn affects the chance expected agreement on which measures of reliability such as Kappa are based.

The effect of spectrum on reliability has been discussed by Guggenmoos-Holzmann.[119] A latent class model divides the subjects into "easy to classify" and "difficult to classify". A population that has mostly "easy to classify" subjects (also called "textbook cases") will increase observed agreement. A population that has many "difficult to classify" subjects, which, because of their borderline findings, may cause the observer to guess one way or the other, and introduce randomness that decreases observed agreement. If one of the two observers is more accurate (with better observer perception by a keener eye or better observer processing by a more valid criterion), the effect of spectrum on reliability can be understood.[5] Although the use an

---

[5] By way of illustration of the importance of the spectrum of abnormality in the population, one can imagine the dipstick test to check if there is any oil in stalled cars done by two observers. One of the observers has a dipstick that is bit too short for some oil reservoirs. Imagine there is a group of cars whose oil

artificial group evenly split between "textbook cases" and normals would increase measured reliability by the Kappa statistic, for better generalizability, it is important to do reliability studies in the groups in which the test is ultimately to be used. Thompson has warned that, although it is tempting to conduct reliability studies in samples of convenience (typically comparing inpatients with medical students) with a high (50%) prevalence of the trait of interest, testing is best done in more representative groups.[120]

A general way to improve reliability is to remove some of the randomness inherent in each of the above categories by making replicate observations by the same observer on the same subject.[121] By choosing the best of three or the best of five, the test result is likely to closer to the true state.

Considering the factors that can affect reproducibility of a test result suggests what can be done to improve reliability. These also can be categorized (See Table 19).

---

reservoirs are either filled to the top or completely dry. If two observers compare their agreement for this group they may find they have quite good agreement. Now if the same observers examine another group of cars with variable levels of oil they may find that the agreement is not as good. The observer with the shorter dipstick may call some partially filled reservoirs empty.

Table 19. Improving the Reliability of the Gag Reflex

CATEGORY	MEASURE
Test:	identical equipment
	similar training of technique for: location of stimulus tongue depressor force
Observers:	
sensory apparatus	similar visual and auditory acuity
processing	standardized criterion for positive test (e.g. palatal elevation and gagging phonation), similar observer training and experience
Environment:	
	quiet
	similar lighting
Subjects:	
	free of sedatives
	same time of day

Appendix II. Reliability and Kappa

Fleiss has reviewed measures of agreement for categorical variables.[71] Kappa, which has been promoted by Feinstein,[83] expresses the agreement observed beyond that expected by chance. Kappa ranges from -1 (complete disagreement) to 1 (complete agreement) with 0 being agreement based on chance alone. Guidelines have been advanced by Landis and Koch[74] to interpret Kappa. A Kappa of > .60 indicates substantial reliability and a Kappa of .41 -.60, moderate reliability. A Kappa of .21 -.40 indicates only fair reliability and a Kappa of < .20, poor reliability. Although admittedly arbitrary, these guidelines have been widely accepted. Kappa is calculated by:

$$\kappa = \frac{Po - Pe}{1.0 - Pe}$$

P_o is simply the proportion on which there is agreement and P_e is the proportion expected based on chance. The two by two table allows one to easily calculate P_o and P_e.

	Observer one		
Observer two	Yes	No	
Yes	a	b	a+b
No	c	d	c+d
	a+c	b+d	N

In the two by two table, P_o, the simple or observed agreement is:

$$(a + d) / N$$

P_e is calculated by determining the expected values for a and d under the assumption of independence of the two observers. This is the sum of the products of the marginal proportions for positive and negative categories:

$$\frac{a+b}{N} * \frac{a+c}{N} + \frac{c+d}{N} * \frac{b+d}{N}$$

An estimation of the standard error of Kappa is required for interpretation. The standard error as derived by Fleiss et al.[122] has become accepted and its validity has been supported by Monte Carlo simulations. The standard error of Kappa used depends on the purpose.[123] To test the hypothesis that the underlying Kappa is equal to zero (implying that the observers agree to an extent based on only chance), the appropriate standard error of Kappa ("null standard error of Kappa") is given by:

$$s.e(\kappa) = \frac{1}{1 - Pe\sqrt{N}} \sqrt{Pe + Pe^2 - [(p1p2(p1 + p2) + q1q2(q1 + q2)]}$$

in which the marginal proportions are represented by:

$$\frac{a+b}{N} = p1 \quad \frac{a+c}{N} = p2 \quad \frac{c+d}{N} = q1 \quad \frac{b+d}{N} = q2$$

In order to test the hypothesis that the underlying Kappa is equal to a prespecified value other than zero, a more cumbersome

standard error is used. This "non-null standard error" is also the one used to create the confidence limits for Kappa. For example the 95% confidence limits are: $\kappa \pm 1.96 * se(\kappa)$.[121] The equation can be separated into three parts:

$$A = \frac{a}{N}[1-(p1+p2)(1-\kappa)]^2 + \frac{d}{N}[1-(q1+q2)(1-\kappa)]^2$$

$$B = (1-\kappa)^2 \left[\frac{b}{N}(p2+q1)^2 + \frac{c}{N}(p1+q2)^2 \right]$$

$$C = [\kappa - Pe(1-\kappa)]^2$$

The standard error of Kappa is calculated as follows:

$$se(\kappa) = \frac{\sqrt{A+B-C}}{(1-Pe)\sqrt{N}}$$

With this standard error of Kappa and confidence intervals, Kappa can be interpreted against the criteria of Landis and Koch.[14] This statistical procedure has been exemplified by Dunn and

Everitt.[124]

The standard error for Kappa can also be approximated for the two observer by two category case by the formula for the standard error of Kappa originally suggested by Cohen.[58,125] This formula has been reported to overestimate the standard error of Kappa and lead to conservative (wider than necessary) confidence intervals.[122]

$$se(\kappa)=\frac{\sqrt{Po-Po^2}}{(1-Pe)\sqrt{N}}$$

Construction of confidence limits with these formulae is based on the assumption that, for large samples, the sampling distribution of Kappa is approximately normal. It has been suggested that for this assumption to hold, the sample size should be at least $16a^2$, where a is the number of categories for the Kappa statistic.[126,127,128] For the dichotomous case here, $a = 2$ and thus a minimum suggested sample size is 64.

Sample size requirements for a Kappa within a known confidence interval can be calculated as follows. Let 2δ be the width of the 95% confidence interval. Thus:

$$\delta = 1.96 * se(\kappa)$$

Therefore, substituting for $se(\kappa)$ and rearranging we get:

$$N = \frac{1.96^2\ Po(1-Po)}{\delta^2(1-Pe)^2}$$

The sample size N required depends on the estimated P_o, estimated P_e and desired δ. This equation can be used to estimate the sample size required for a reliability study for dichotomous data in the two observer by two category case. The sample size N can also be formulated in terms of the estimated Kappa and the rate of test positivity although the algebra becomes quite unwieldy (Appendix XV).

Donner and Eliasziw[129] described a goodness-of-fit (GOF) approach to inference procedures for Kappa. Assuming there is no rater bias (in other words they both rate the same proportion of subjects positive), this approach can give confidence intervals with an N as small as 25. A Monte Carlo study showed that this approach was valid as long as the prevalence of positive ratings was > .10 (and < .90) and Kappa was < .90. An advantage of the GOF approach is that it not does not assume that the sampling distribution of Kappa estimates follow a normal distribution. It also does not yield degenerate upper confidence limits (i.e. > 1). However, use of this standard error is limited by the assumption that the raters are unbiased. This bias can be assessed with McNemar's test for correlated proportions. As the sample size tables in this thesis were developed with inter-

observer reliability in mind, rater bias could occur. Fortunately, the standard error of Fleiss holds in spite of rater bias.[129]

Appendix III: Flow Chart for Study

Admission to hospital (eg.Monday at noon)
|
V
Identification at t= 24-72hrs (eg.Tuesday am)
| (patient has had symptoms for at least 24 hours)
| (patient name kept in log book)
| (consent form given to patient)
V
Enrollment at 48-96 hrs. (eg.Wednesday am)
| (consent form signed)
| (videotaping of voice quality, cough quality, pharyngeal
| sensation, gag reflex and graduated swallow of ice
| chips, 50 ml, 100 ml and then 150 ml) daily x 2 days
| 48hrs,72hrs, or 72hrs,96hrs or 96hrs, 120hrs
| (eg. W,Th)
| (Hx/PE data sheet filled in by researcher)
V
Interview at 72-120 hrs. (eg. Thursday am)
| interview with nursing staff
| as to impression of aspiration and dysphagia
|
V
Outcomes:
1. vital status
2. disruption of oral feeding eg. N/G tube started
3. pneumonia probable or presumptive
4. feeding difficulty
5. body weight

- Hospital chart is reviewed until time of discharge for documentation of vital status, symptoms or signs of pneumonia, feeding disruption or feeding difficulty.

- Patient or next-of-kin or health care personnel are contacted by telephone at one month to determine if there has been any symptoms of pneumonia, feeding disruption or feeding difficulty.

Appendix IV: Study Data Entry Sheet

DEMOGRAPHIC DATA
Patient E#: _____

Date of admission: day__ month___ year___

Date of first examination: day___ month___ year___

Date of birth: day___ month___ year___

Age: ____ years

Sex: (circle) male female

How referred to EGH: (circle) ER physician family MD

Family doctor's name (responsible for subsequent care)

Family doctor's office telephone: (___)-___-_____

MISCELLANEOUS DATA

Stroke risk factors (circle)
smoking hypertension atrial fibrillation
peripheral vascular disease

Sedative medications_____

Dentition: dentures? caries

Is an IV in place (circle)?
yes no

Has assistance with feeding been required (circle)?
spoon feeding cup holding drinking with a straw

Does patient move from wheelchair to bed (circle)?
with help alone

Is there difficulty getting on and off toilet (circle)?
with help alone

Is patient able to bathe self (circle)?
with help alone

Is patient able to walk on level surface (circle)?
with help alone

Is patient able to ascend/ descend stairs (circle)?
with help alone

Is patient able to dress self (circle)?
with help alone

Is patient able to control bowels (circle)?
with help alone

Is patient able to control bladder (circle)?
with help alone

NEUROLOGICAL CLINICAL DATA

Date and approximate time of onset of symptoms:
hours (2400 format) day___ month___ year___

Any decreased level of consciousness in time since symptom onset until first examination documented in chart:
date initially noted day__ month___ year___
date resolved day__ month___ year___
worst severity (specify) _____
_____
_____

visual field defect: right homonymous left homonymous
dysconjugate gaze

facial numbness: right left

facial droop: right left

speech difficulty: receptive, motor, global, dysarthria

hemiparesis	right (location)	left (location)
upper extremity,		
best strength:	__/5__ ( )	__/5__ ( )
worst deficit:	__/5__ ( )	__/5__ ( )
lower extremity,		
best strength:	__/5__ ( )	__/5__ ( )
worst deficit:	__/5__ ( )	__/5__ ( )

hemianesthesia: right left

asymmetric reflexes: right left
(specify > or < or =)

plantar response: right left indeterminate

CT scan(s)
date of exam: _____
results:

_____
_____

date of exam:
results:

_____
_____

date of exam:
results:

_____
_____

History of previous CVA? If yes, give details

_____
_____

RESPIRATORY CLINICAL DATA:
cough in past 24 hours? yes no
shortness of breath in past 24 hours? yes no

highest temperature in past 24 hours: ____ C
highest respiratory rate in past 24 hours: ___ /min

	right	left
percussion:	dullness	dullness
auscultation:	crackles /no crackles breath sds./ decreased	crackles/ no crackles breath sds / decreased

initial CXR report
date of initial exam:

Barthel Index

(score 0 where patient unable to meet defined criteria)

Criterion	With Help	Independent
1.Feeding (if food needs to be cut up = help)	5	10
2.Moving from wheelchair to bed and return (includes sitting up in bed)	5-10	15
3. Personal toilet (wash face, comb hair, shave, clean teeth)	0	5
4. Getting on and off toilet (handling clothes, wipe, flush)	5	10
5. Bathing self	0	5
6. Walking on level surface	10	15
(or if unable to walk, propel wheelchair) * score only if unable to walk	*0	*5
7. Ascend and descend stairs	5	10
8. Dressing (includes tying shoes, fastening fasteners)	5	10
9. Controlling bowels	5	10
10. Controlling bladder	5	10

Rapid Disability Rating Scale-2

(circle response that reflects current behaviour)

Assistance with activities of daily living:

Eating	None	A little	A lot	Spoon-feed, IV
Walking (with can or walker if used)	None	A little	A lot	Does not walk
Mobility (going outside and getting about with wheelchair, etc.)	None	A little	A lot	Is housebound
Bathing (include getting supplies, supervising)	None	A little	A lot	Must be bathed
Dressing (includes help in selecting clothes)	None	A little	A lot	Must be dressed
Toileting (includes help with clothes, cleaning, or help with ostomy, catheter)	None	A little	A lot	Uses bedpan Catheter
Grooming (shaving for men, hair-dressing for women, nails, teeth)	None	A little	A lot	Must be groomed
Adaptive tasks (managing money, possessions, telephoning, buying newspaper, toilet articles, snacks)	None	A little	A lot	Cannot manage

Degree of disability:

Communication (expressing self)	None	A little	A lot	Does not Communicate
Hearing (with aid if used)	None	A little	A lot	Does not hear
Sight (with glasses, if used)	None	A little	A lot	Does not see
Diet (deviation from normal)	None	A little	A lot	Fed by IV
In bed during day (ordered or self-initiated)	None	A little (<3 hr)	A lot	Most/ all
Incontinence (urine/feces, with catheter or prosthesis if used)	None	A little	A lot (weekly)	No control
Medication	None	A little	A lot	Daily

Degree of special problems:

Confusion	None	A little	A lot	Extreme
Uncooperative (combats efforts to help with care)	None	A little	A lot	Extreme
Depression	None	A little	A lot	Extreme

Appendix V: Bedside Video-recording Protocol

1.Record consent
2.Record time, date (eg. watch)
3.Bedside mini-neuro assessment (videocamera at patient's shoulder height to include arms and legs but not face)
a.Greeting "Hello, how are you?" "How are you feeling today"(voice observation). "Say pa, pa, pa. "Say la, la, la." "Say ga, ga, ga."
b."Please sit up." "Sit up straight."
c.Chest auscultation, comprehension
"I'm going to listen to your chest. Take a deep breath in and out. Take a deep breath in and out."
d.Cough quality, comprehension
"Give a strong cough. Give a really strong cough."
e.Test of alertness, comprehension and limb strength.
"Lift up your arms and hold them there. Hold them there" (15 sec)
"Lift up this leg, keep it there (15 sec). Lift up this leg keep it there. (15 sec)"

4. Videotape closeup pharyngeal examination (patient's shoulders and face angled at 45 degrees from midline so that camera can get closeup from bedside)
a.Test of facial symmetry, facial strength and tongue protrusion
"Open your mouth. Open your mouth wide." "Stick your tongue out. Stick your tongue straight out.""Say eeeeeeh. Say eeeeeh"

b.Test of pharyngeal sensation (touch faucial arches)
"Let me know if you feel this. Let me know if you feel this"

c.Test of gag reflex touch posterior soft palate and check for palatal elevation. Touch the posterior pharynx and check for palatal elevation, pharyngeal constriction and gagging phonation.

d.Swallowing test (take view from the left with patient facing forward including neck).
"Swallow this bit of ice" Check for laryngeal elevation and listen and observe for choking for 15 sec . "Say la,la,la." Check for wet hoarseness.
"Swallow this water" (hold the spout lid at 90 degrees.)observe 15 sec "Say la, la, la." Check for wet hoarseness.
"Swallow this water" Observe 15 sec "Say la, la, la." Check for wet hoarseness.
"Swallow this water" Observe 15 sec "Say la, la, la" Check for wet hoarseness.

Patient's subjective impression (edited out later)
"How did the water go down?".

Appendix VI: Consent Form for Study of Swallowing Problems in Stroke

Principal investigator: H.J. Biem
(416-747-3511)

You have had (or your family member has had) a stroke. Strokes may affect swallowing. If swallowing is affected, food and liquids can get into the airway and cause choking, feeding difficulty and sometimes pneumonia. We are conducting a study of swallowing in stroke patients and ask you to participate. We are checking the throat, cough and voice to see how they relate to the risk of swallowing problems.

If you agree to participate in this study, a study doctor will briefly examine you daily for two days. The throat will be checked with a tongue depressor. If you are allowed to take fluids by mouth, you will be given small amounts of ice chips and water to swallow. A videotape will be made of the examination but you will not be recognizable.

Your chart will be also be reviewed while you are in hospital and on discharge. You or your family member will be contacted by an investigator twice during the next month to determine how you are doing. A visit to your personal doctor may be recommended if you have any difficulty with cough, shortness of breath, fever or any other symptoms that might point to pneumonia. We ask that you visit your personal doctor one month after this stroke for a general check-up including measurement of your weight.

If you decide to participate, all personal information and video recordings will be kept confidential. If you decide not to participate this will not affect your relationship with your personal physician.

Whether or not you participate, you will receive the same hospital care. The examination will not alter your care or your risk of problems with swallowing. The results of this study, however, may benefit stroke patients in the future.

I understand and agree to participate in this study.

_____ _____

Signature of Patient/Next-of-kin Date

_____

Signature of Witness

Appendix VII. A Sample Size Table for Kappa Reliability Studies Using Estimates of Observed and Expected Agreement

Background

A confidence limit rather than hypothesis testing approach was taken to estimate sample size because the aim was to report the Kappa values with confidence limits rather than with p values. The two approaches to sample size estimation have been compared by Bristol.[130]

Computer simulation and computer intensive methods, now widely accessible because of powerful desktop computers, can allow empiric testing of statistical models and give alternative approaches to estimation of statistical confidence intervals and power. Monte Carlo methods allow construction of confidence intervals based directly on the variation in the parameter estimates observed in the data generated from the model. Programs can also be used to examine a test statistic through systematic enumeration.[131]

A computer program was written to calculate a sample size table for reasonable values of P_o, P_e, and delta (δ). Delta was defined as half of the width of the 95% confidence interval for Kappa. This sample size table was intended to show the trade-off between the size of the confidence interval around Kappa and the sample size. For simplicity, incremental values for P_o, P_e and delta were rounded to the nearest .05. In order to explore the

behaviour of the sample size requirements near Kappa 1.0, values of P_o were included ranging from .95 to .99. It was assumed that the observed agreement (P_o) would be greater than chance expected agreement (P_e) reasoning that to include the case of negative Kappa values would only add unnecessary complexity to the table. Similarly, the maximum delta was limited to 0.30 because a larger delta would mean the Kappa estimate would be too imprecise to interpret according to the guidelines of Landis and Koch.[*] The computer program is listed in Appendix VIII.

For the verification and refinement of the above equation-derived sample size table, combinations of values for P_o and P_e were obtained from a simulation computer program generating two by two tables with various sample sizes. The two by two tables were produced by three different methods as described in Appendix IX. The computer program code for each of the methods of generation of two by two tables are listed in Appendices X, XI, XII. The relationship between P_o and P_e was examined graphically with scatterplots. The values of P_o and P_e were then rounded off to the nearest 5% and sorted on P_o, P_e, Kappa and N. For each two by two table, there was also a listing of 95% lower and upper confidence limits for Kappa, delta (1.96 * seKappa), the standard error for Kappa of Fleiss, the approximation of the standard

---

[*] If, for example, the Kappa estimate were .70 ± .35, then the true value could be in the ranges .81-1.0 (almost perfect), .61-.80 (substantial), .41 -.60 (moderate) or .21 - .40 (poor). The upper confidence limit is also "degenerate" when the Kappa is high and the sample size is small because of the violation the large sample assumption for the normality of the Kappa statistic of Fleiss's standard error of Kappa.

error of Kappa of Cohen, the absolute difference of the two standard errors, and the cell values.

The computer programs were written in C, compiled by TurboC++ (Borland), and run on an IBM 486 PC. Computer programs were written to calculate sample sizes for reasonable values of P_c and P_e (kaptable.c in Appendix VIII), to list random two by two tables (kaprand.c in Appendix X and kapsimp.c in Appendix XI), and list permutations of two by two tables (kapperm.c in Appendix XII). Scatterplots were created by the BMDP new system statistical software package. The computer generated two by two tables were sorted with the QuattroPro (Borland) spreadsheet program.

Equation-derived Sample Sizes with Cohen's Approximation

For each combination of P_c and P_e, the equation-derived sample size table (Table 20) lists the Kappa value and the various sample sizes (N) corresponding to 95% confidence limits around Kappa of plus or minus δ. The δ ranged from .05 to .30. The sample size required increased with higher values of P_e for a given P_c. As could be deduced from inspection of the sample size equation, the sample size required increased for a smaller confidence interval around Kappa. When P_c was 1.0, Kappa was 1.0 and N was 0 regardless of P_e, consistent with the breakdown of the Kappa's standard error as Kappa approaches 1.

Table 20. Equation-derived Sample Size Table

The sample size (N) depends on the observed (P_o) and the expected (P_e) agreement proportions as well as the confidence limits (± δ) desired for Kappa.

P_o	P_e	Kappa	N for	k±.05	k±.10	k±.15	k±.20	k±.25	k±.30
0.10	0.00	0.10		138	35	15	9	6	4
0.10	0.05	0.05		153	38	17	10	6	4
P_o	P_e	Kappa	N for	k±.05	k±.10	k±.15	k±.20	k±.25	k±.30
0.15	0.00	0.15		196	49	22	12	8	5
0.15	0.05	0.11		217	54	24	14	9	6
0.15	0.10	0.06		242	60	27	15	10	7
P_o	P_e	Kappa	N for	k±.05	k±.10	k±.15	k±.20	k±.25	k±.30
0.20	0.00	0.20		246	61	27	15	10	7
0.20	0.05	0.16		272	68	30	17	11	8
0.20	0.10	0.11		304	76	34	19	12	8
0.20	0.15	0.06		340	85	38	21	14	9
P_o	P_e	Kappa	N for	k±.05	k±.10	k±.15	k±.20	k±.25	k±.30
0.25	0.00	0.25		288	72	32	18	12	8
0.25	0.05	0.21		319	80	35	20	13	9
0.25	0.10	0.17		356	89	40	22	14	10
0.25	0.15	0.12		399	100	44	25	16	11
0.25	0.20	0.06		450	113	50	28	18	13
P_o	P_e	Kappa	N for	k±.05	k±.10	k±.15	k±.20	k±.25	k±.30
0.30	0.00	0.30		323	81	36	20	13	9
0.30	0.05	0.26		358	89	40	22	14	10
0.30	0.10	0.22		398	100	44	25	16	11
0.30	0.15	0.18		447	112	50	28	18	12
0.30	0.20	0.12		504	126	56	32	20	14
0.30	0.25	0.07		574	143	64	36	23	16
P_o	P_e	Kappa	N for	k±.05	k±.10	k±.15	k±.20	k±.25	k±.30
0.35	0.00	0.35		350	87	39	22	14	10
0.35	0.05	0.32		387	97	43	24	15	11
0.35	0.10	0.28		432	108	48	27	17	12
0.35	0.15	0.24		484	121	54	30	19	13
0.35	0.20	0.19		546	137	61	34	22	15
0.35	0.25	0.13		621	155	69	39	25	17
0.35	0.30	0.07		713	178	79	45	29	20
P_o	P_e	Kappa	N for	k±.05	k±.10	k±.15	k±.20	k±.25	k±.30
0.40	0.00	0.40		369	92	41	23	15	10
0.40	0.05	0.37		409	102	45	26	16	11
0.40	0.10	0.33		455	114	51	28	18	13
0.40	0.15	0.29		510	128	57	32	20	14
0.40	0.20	0.25		576	144	64	36	23	16
0.40	0.25	0.20		656	164	73	41	26	18
0.40	0.30	0.14		753	188	84	47	30	21

0.40	0.35	0.08		873	218	97	55	35	24

P_o	P_e	Kappa	N for	k±.05	k±.10	k±.15	k±.20	k±.25	k±.30
0.45	0.00	0.45		380	95	42	24	15	11
0.45	0.05	0.42		421	105	47	26	17	12
0.45	0.10	0.39		470	117	52	29	19	13
0.45	0.15	0.35		526	132	58	33	21	15
0.45	0.20	0.31		594	149	66	37	24	17
0.45	0.25	0.27		676	169	75	42	27	19
0.45	0.30	0.21		776	194	86	49	31	22
0.45	0.35	0.15		900	225	100	56	36	25
0.45	0.40	0.08		1056	264	117	66	42	29

P_o	P_e	Kappa	N for	k±.05	k±.10	k±.15	k±.20	k±.25	k±.30
0.50	0.00	0.50		384	96	43	24	15	11
0.50	0.05	0.47		426	106	47	27	17	12
0.50	0.10	0.44		474	119	53	30	19	13
0.50	0.15	0.41		532	133	59	33	21	15
0.50	0.20	0.37		600	150	67	38	24	17
0.50	0.25	0.33		683	171	76	43	27	19
0.50	0.30	0.29		784	196	87	49	31	22
0.50	0.35	0.23		909	227	101	57	36	25
0.50	0.40	0.17		1067	267	119	67	43	30
0.50	0.45	0.09		1270	317	141	79	51	35

P_o	P_e	Kappa	N for	k±.05	k±.10	k±.15	k±.20	k±.25	k±.30
0.55	0.00	0.55		380	95	42	24	15	11
0.55	0.05	0.53		421	105	47	26	17	12
0.55	0.10	0.50		470	117	52	29	19	13
0.55	0.15	0.47		526	132	58	33	21	15
0.55	0.20	0.44		594	149	66	37	24	17
0.55	0.25	0.40		676	169	75	42	27	19
0.55	0.30	0.36		776	194	86	49	31	22
0.55	0.35	0.31		900	225	100	56	36	25
0.55	0.40	0.25		1056	264	117	66	42	29
0.55	0.45	0.18		1257	314	140	79	50	35
0.55	0.50	0.10		1521	380	169	95	61	42

P_o	P_e	Kappa	N for	k±.05	k±.10	k±.15	k±.20	k±.25	k±.30
0.60	0.00	0.60		369	92	41	23	15	10
0.60	0.05	0.58		409	102	45	26	16	11
0.60	0.10	0.56		455	114	51	28	18	13
0.60	0.15	0.53		510	128	57	32	20	14
0.60	0.20	0.50		576	144	64	36	23	16
0.60	0.25	0.47		656	164	73	41	26	18
0.60	0.30	0.43		753	188	84	47	30	21
0.60	0.35	0.38		873	218	97	55	35	24
0.60	0.40	0.33		1024	256	114	64	41	28
0.60	0.45	0.27		1219	305	135	76	49	34
0.60	0.50	0.20		1475	369	164	92	59	41
0.60	0.55	0.11		1821	455	202	114	73	51

P_o	P_e	Kappa	N for	k±.05	k±.10	k±.15	k±.20	k±.25	k±.30
0.65	0.00	0.65		350	87	39	22	14	10
0.65	0.05	0.63		387	97	43	24	15	11
0.65	0.10	0.61		432	108	48	27	17	12
0.65	0.15	0.59		484	121	54	30	19	13

0.65	0.20	0.56	546	137	61	34	22	15
0.65	0.25	0.53	621	155	69	39	25	17
0.65	0.30	0.50	713	178	79	45	29	20
0.65	0.35	0.46	827	207	92	52	33	23
0.65	0.40	0.42	971	243	108	61	39	27
0.65	0.45	0.36	1156	289	128	72	46	32
0.65	0.50	0.30	1398	350	155	87	56	39
0.65	0.55	0.22	1726	432	192	108	69	48
0.65	0.60	0.13	2185	546	243	137	87	61

P_o	P_e	Kappa	N for	k±.05	k±.10	k±.15	k±.20	k±.25	k±.30
0.70	0.00	0.70		323	81	36	20	13	9
0.70	0.05	0.68		358	89	40	22	14	10
0.70	0.10	0.67		398	100	44	25	16	11
0.70	0.15	0.65		447	112	50	28	18	12
0.70	0.20	0.63		504	126	56	32	20	14
0.70	0.25	0.60		574	143	64	36	23	16
0.70	0.30	0.57		659	165	73	41	26	18
0.70	0.35	0.54		764	191	85	48	31	21
0.70	0.40	0.50		896	224	100	56	36	25
0.70	0.45	0.45		1067	267	119	67	43	30
0.70	0.50	0.40		1291	323	143	81	52	36
0.70	0.55	0.33		1594	398	177	100	64	44
0.70	0.60	0.25		2017	504	224	126	81	56
0.70	0.65	0.14		2634	659	293	165	105	73

P_o	P_e	Kappa	N for	k±.05	k±.10	k±.15	k±.20	k±.25	k±.30
0.75	0.00	0.75		288	72	32	18	12	8
0.75	0.05	0.74		319	80	35	20	13	9
0.75	0.10	0.72		356	89	40	22	14	10
0.75	0.15	0.71		399	100	44	25	16	11
0.75	0.20	0.69		450	113	50	28	18	13
0.75	0.25	0.67		512	128	57	32	20	14
0.75	0.30	0.64		588	147	65	37	24	16
0.75	0.35	0.62		682	170	76	43	27	19
0.75	0.40	0.58		800	200	89	50	32	22
0.75	0.45	0.55		952	238	106	60	38	26
0.75	0.50	0.50		1152	288	128	72	46	32
0.75	0.55	0.44		1423	356	158	89	57	40
0.75	0.60	0.38		1801	450	200	113	72	50
0.75	0.65	0.29		2352	588	261	147	94	65
0.75	0.70	0.17		3201	800	356	200	128	89

P_o	P_e	Kappa	N for	k±.05	k±.10	k±.15	k±.20	k±.25	k±.30
0.80	0.00	0.80		246	61	27	15	10	7
0.80	0.05	0.79		272	68	30	17	11	8
0.80	0.10	0.78		304	76	34	19	12	8
0.80	0.15	0.76		340	85	38	21	14	9
0.80	0.20	0.75		384	96	43	24	15	11
0.80	0.25	0.73		437	109	49	27	17	12
0.80	0.30	0.71		502	125	56	31	20	14
0.80	0.35	0.69		582	145	65	36	23	16
0.80	0.40	0.67		683	171	76	43	27	19
0.80	0.45	0.64		813	203	90	51	33	23
0.80	0.50	0.60		983	246	109	61	39	27
0.80	0.55	0.56		1214	304	135	76	49	34
0.80	0.60	0.50		1537	384	171	96	61	43
0.80	0.65	0.43		2007	502	223	125	80	56
0.80	0.70	0.33		2732	683	304	171	109	76

0.80	0.75	0.20		3934	983	437	246	157	109

P_o	P_e	Kappa	N for	k±.05	k±.10	k±.15	k±.20	k±.25	k±.30
0.85	0.00	0.85		196	49	22	12	8	5
0.85	0.05	0.84		217	54	24	14	9	6
0.85	0.10	0.83		242	60	27	15	10	7
0.85	0.15	0.82		271	68	30	17	11	8
0.85	0.20	0.81		306	77	34	19	12	9
0.85	0.25	0.80		348	87	39	22	14	10
0.85	0.30	0.79		400	100	44	25	16	11
0.85	0.35	0.77		464	116	52	29	19	13
0.85	0.40	0.75		544	136	60	34	22	15
0.85	0.45	0.73		648	162	72	40	26	18
0.85	0.50	0.70		784	196	87	49	31	22
0.85	0.55	0.67		968	242	108	60	39	27
0.85	0.60	0.63		1225	306	136	77	49	34
0.85	0.65	0.57		1599	400	178	100	64	44
0.85	0.70	0.50		2177	544	242	136	87	60
0.85	0.75	0.40		3135	784	348	196	125	87
0.85	0.80	0.25		4898	1225	544	306	196	136

P_o	P_e	Kappa	N for	k±.05	k±.10	k±.15	k±.20	k±.25	k±.30
0.90	0.00	0.90		138	35	15	9	6	4
0.90	0.05	0.89		153	38	17	10	6	4
0.90	0.10	0.89		171	43	19	11	7	5
0.90	0.15	0.88		191	48	21	12	8	5
0.90	0.20	0.88		216	54	24	14	9	6
0.90	0.25	0.87		246	61	27	15	10	7
0.90	0.30	0.86		282	71	31	18	11	8
0.90	0.35	0.85		327	82	36	20	13	9
0.90	0.40	0.83		384	96	43	24	15	11
0.90	0.45	0.82		457	114	51	29	18	13
0.90	0.50	0.80		553	138	61	35	22	15
0.90	0.55	0.78		683	171	76	43	27	19
0.90	0.60	0.75		864	216	96	54	35	24
0.90	0.65	0.71		1129	282	125	71	45	31
0.90	0.70	0.67		1537	384	171	96	61	43
0.90	0.75	0.60		2213	553	246	138	89	61
0.90	0.80	0.50		3457	864	384	216	138	96
0.90	0.85	0.33		6147	1537	683	384	246	171

P_o	P_e	Kappa	N for	k±.05	k±.10	k±.15	k±.20	k±.25	k±.30
0.95	0.00	0.95		73	18	8	5	3	2
0.95	0.05	0.95		81	20	9	5	3	2
0.95	0.10	0.94		90	23	10	6	4	3
0.95	0.15	0.94		101	25	11	6	4	3
0.95	0.20	0.94		114	29	13	7	5	3
0.95	0.25	0.93		130	32	14	8	5	4
0.95	0.30	0.93		149	37	17	9	6	4
0.95	0.35	0.92		173	43	19	11	7	5
0.95	0.40	0.92		203	51	23	13	8	6
0.95	0.45	0.91		241	60	27	15	10	7
0.95	0.50	0.90		292	73	32	18	12	8
0.95	0.55	0.89		360	90	40	23	14	10
0.95	0.60	0.88		456	114	51	29	18	13
0.95	0.65	0.86		596	149	66	37	24	17
0.95	0.70	0.83		811	203	90	51	32	23
0.95	0.75	0.80		1168	292	130	73	47	32
0.95	0.80	0.75		1825	456	203	114	73	51

0.95	0.85	0.67		3244	811	360	203	130	90
0.95	0.90	0.50		7299	1825	811	456	292	203

P_c	P_e	Kappa	N for	k±.05	k±.10	k±.15	k±.20	k±.25	k±.30
0.96	0.00	0.96		59	15	7	4	2	2
0.96	0.05	0.96		65	16	7	4	3	2
0.96	0.10	0.96		73	18	8	5	3	2
0.96	0.15	0.95		82	20	9	5	3	2
0.96	0.20	0.95		92	23	10	6	4	3
0.96	0.25	0.95		105	26	12	7	4	3
0.96	0.30	0.94		120	30	13	8	5	3
0.96	0.35	0.94		140	35	16	9	6	4
0.96	0.40	0.93		164	41	18	10	7	5
0.96	0.45	0.93		195	49	22	12	8	5
0.96	0.50	0.92		236	59	26	15	9	7
0.96	0.55	0.91		291	73	32	18	12	8
0.96	0.60	0.90		369	92	41	23	15	10
0.96	0.65	0.89		482	120	54	30	19	13
0.96	0.70	0.87		656	164	73	41	26	18
0.96	0.75	0.84		944	236	105	59	38	26
0.96	0.80	0.80		1475	369	164	92	59	41
0.96	0.85	0.73		2623	656	291	164	105	73
0.96	0.90	0.60		5901	1475	656	369	236	164
0.96	0.95	0.20		23603	5901	2623	1475	944	656

P_c	P_e	Kappa	N for	k±.05	k±.10	k±.15	k±.20	k±.25	k±.30
0.97	0.00	0.97		45	11	5	3	2	1
0.97	0.05	0.97		50	12	6	3	2	1
0.97	0.10	0.97		55	14	6	3	2	2
0.97	0.15	0.96		62	15	7	4	2	2
0.97	0.20	0.96		70	17	8	4	3	2
0.97	0.25	0.96		79	20	9	5	3	2
0.97	0.30	0.96		91	23	10	6	4	3
0.97	0.35	0.95		106	26	12	7	4	3
0.97	0.40	0.95		124	31	14	8	5	3
0.97	0.45	0.95		148	37	16	9	6	4
0.97	0.50	0.94		179	45	20	11	7	5
0.97	0.55	0.93		221	55	25	14	9	6
0.97	0.60	0.92		279	70	31	17	11	8
0.97	0.65	0.91		365	91	41	23	15	10
0.97	0.70	0.90		497	124	55	31	20	14
0.97	0.75	0.88		715	179	79	45	29	20
0.97	0.80	0.85		1118	279	124	70	45	31
0.97	0.85	0.80		1987	497	221	124	79	55
0.97	0.90	0.70		4472	1118	497	279	179	124
0.97	0.95	0.40		17886	4472	1987	1118	715	497

P_c	P_e	Kappa	N for	k±.05	k±.10	k±.15	k±.20	k±.25	k±.30
0.98	0.00	0.98		30	8	3	2	1	1
0.98	0.05	0.98		33	8	4	2	1	1
0.98	0.10	0.98		37	9	4	2	1	1
0.98	0.15	0.98		42	10	5	3	2	1
0.98	0.20	0.97		47	12	5	3	2	1
0.98	0.25	0.97		54	13	6	3	2	1
0.98	0.30	0.97		61	15	7	4	2	2
0.98	0.35	0.97		71	18	8	4	3	2
0.98	0.40	0.97		84	21	9	5	3	2
0.98	0.45	0.96		100	25	11	6	4	3
0.98	0.50	0.96		120	30	13	8	5	3
0.98	0.55	0.96		149	37	17	9	6	4

0.98	0.60	0.95	188	47	21	12	8	5
0.98	0.65	0.94	246	61	27	15	10	7
0.98	0.70	0.93	335	84	37	21	13	9
0.98	0.75	0.92	482	120	54	30	19	13
0.98	0.80	0.90	753	188	84	47	30	21
0.98	0.85	0.87	1339	335	149	84	54	37
0.98	0.90	0.80	3012	753	335	188	120	84
0.98	0.95	0.60	12047	3012	1339	753	482	335

P_c	P_e	Kappa	N for	k±.05	k±.10	k±.15	k±.20	k±.25	k±.30
0.99	0.00	0.99		15	4	2	1	1	0
0.99	0.05	0.99		17	4	2	1	1	0
0.99	0.10	0.99		19	5	2	1	1	1
0.99	0.15	0.99		21	5	2	1	1	1
0.99	0.20	0.99		24	6	3	1	1	1
0.99	0.25	0.99		27	7	3	2	1	1
0.99	0.30	0.99		31	8	3	2	1	1
0.99	0.35	0.98		36	9	4	2	1	1
0.99	0.40	0.98		42	11	5	3	2	1
0.99	0.45	0.98		50	13	6	3	2	1
0.99	0.50	0.98		61	15	7	4	2	2
0.99	0.55	0.98		75	19	8	5	3	2
0.99	0.60	0.97		95	24	11	6	4	3
0.99	0.65	0.97		124	31	14	8	5	3
0.99	0.70	0.97		169	42	19	11	7	5
0.99	0.75	0.96		243	61	27	15	10	7
0.99	0.80	0.95		380	95	42	24	15	11
0.99	0.85	0.93		676	169	75	42	27	19
0.99	0.90	0.90		1521	380	169	95	61	42
0.99	0.95	0.80		6085	1521	676	380	243	169

P_c	P_e	Kappa	N for	k±.05	k±.10	k±.15	k±.20	k±.25	k±.30
1.00	0.00	1.00		0	0	0	0	0	0
1.00	0.05	1.00		0	0	0	0	0	0
1.00	0.10	1.00		0	0	0	0	0	0
1.00	0.15	1.00		0	0	0	0	0	0
1.00	0.20	1.00		0	0	0	0	0	0
1.00	0.25	1.00		0	0	0	0	0	0
1.00	0.30	1.00		0	0	0	0	0	0
1.00	0.35	1.00		0	0	0	0	0	0
1.00	0.40	1.00		0	0	0	0	0	0
1.00	0.45	1.00		0	0	0	0	0	0
1.00	0.50	1.00		0	0	0	0	0	0
1.00	0.55	1.00		0	0	0	0	0	0
1.00	0.60	1.00		0	0	0	0	0	0
1.00	0.65	1.00		0	0	0	0	0	0
1.00	0.70	1.00		0	0	0	0	0	0
1.00	0.75	1.00		0	0	0	0	0	0
1.00	0.80	1.00		0	0	0	0	0	0
1.00	0.85	1.00		0	0	0	0	0	0
1.00	0.90	1.00		0	0	0	0	0	0
1.00	0.95	1.00		0	0	0	0	0	0

Verification of Sample Size Table with Computer Simulation

For each computer generated two by two table (of size N and cell values a, b, c, and d), there was a listing of P_o, P_e, and Kappa. Inspection of tables showed that for a given level of observed agreement (P_o), P_e was limited by a minimum.[1] Inspection of the two by two tables in which P_e was at minimum for a given P_o showed that these cells had "balanced marginals" as termed by Feinstein.[132] This means that the row marginal proportions (p1, q1) and column marginal proportions (p2, q2) were approximately 0.5. (See Appendix II for calculations of marginal proportions.) This implies that the numbers in the agreement (a and d) cells approximate each other. A minimum P_e also occurred when there was "asymmetric unbalance" as termed by Feinstein. This means that one of the non-agreement cells (b or c) had a low number. A scatterplot of P_o and P_e (neither rounded off) showed the constraints on P_e for a given value of P_o (Figure 1). With the characteristics of a two by two table with a minimum P_e, a more exact minimum value for P_e for a given value of P_o could be derived algebraically (Appendix XIII).

Inspection of the random two by two tables showed that Kappa values for a given P_o was constrained by a maximum that occurred when P_e was at a minimum, as could have beem simply deduced from

---

[1] The minimum for P_e was approximately .50 if P_o was ≥ .80, .45 if P_o was .75-.60, .40 if P_o was .55-.50 and .35 if P_o was 0.45. The minimum P_e was .30 if P_o was .35 and .25 if P_o was .30. The minimum P_e was .20 if P_o was .25.

the Kappa equation. Although the maximum Kappa values for a given P_o could be estimated from a scatterplot of Kappa versus P_o (Figure 2).[v] the more exact values for a Kappa maximum for a given P_o could be derived algebraically (Appendix XIV). The algebraically derived minimum values for P_e and maximum values for Kappa for a given value of P_o are listed in Table 21.

[v] For P_o = 1, $Kappa_{max}$ was 1. For P_o = .90, $Kappa_{max}$ was .80. For P_o = .80, $Kappa_{max}$ was .60. For P_o = .70, $Kappa_{max}$ was .45. For P_o = .60, $Kappa_{max}$ was .30.

Figure 1. Scatterplot of P_o and P_e from computer generated two by two tables (patterned random number method).

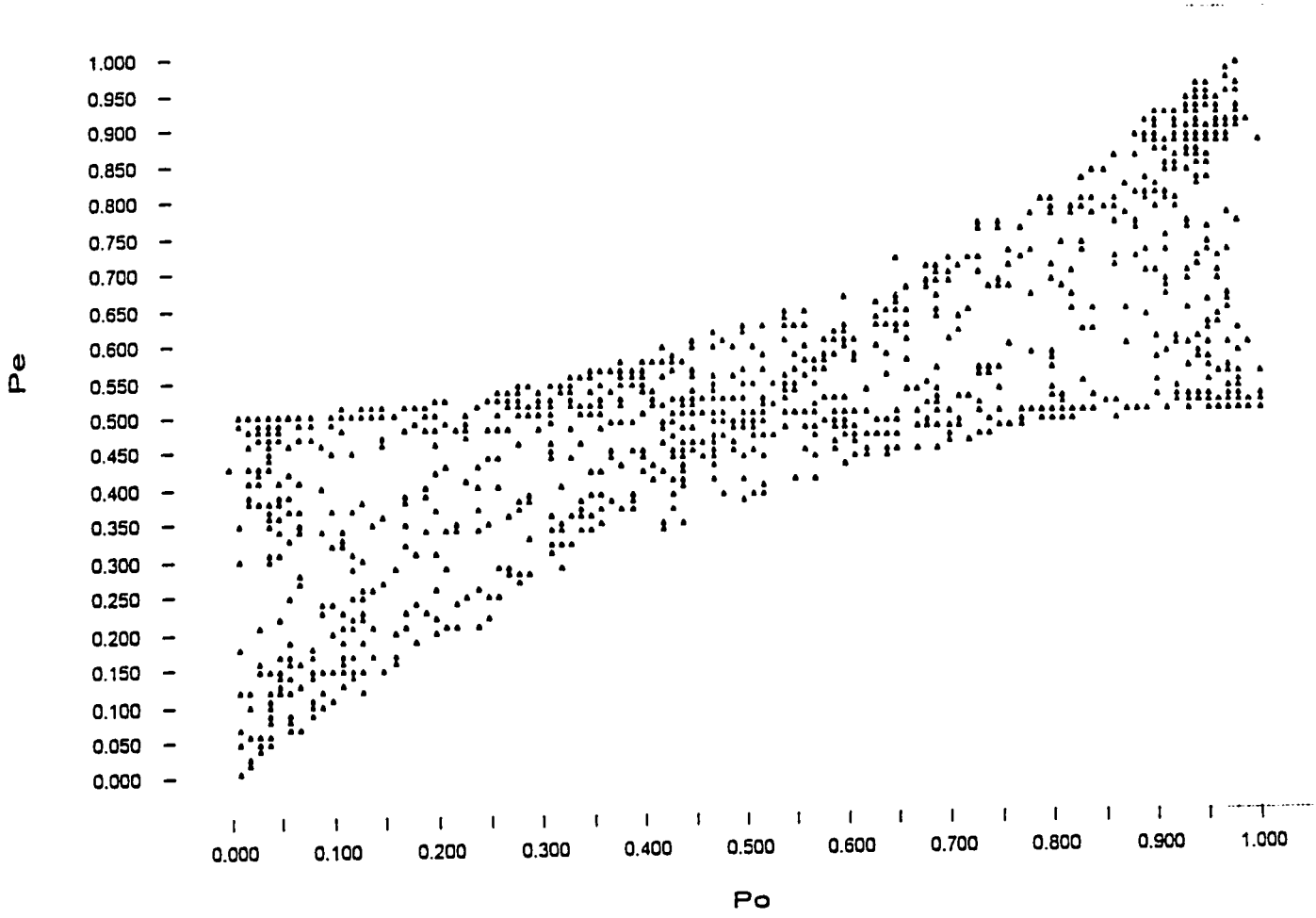


Figure 2. Scatterplot of P_o and Kappa from computer generated two by two tables (patterned random number method).

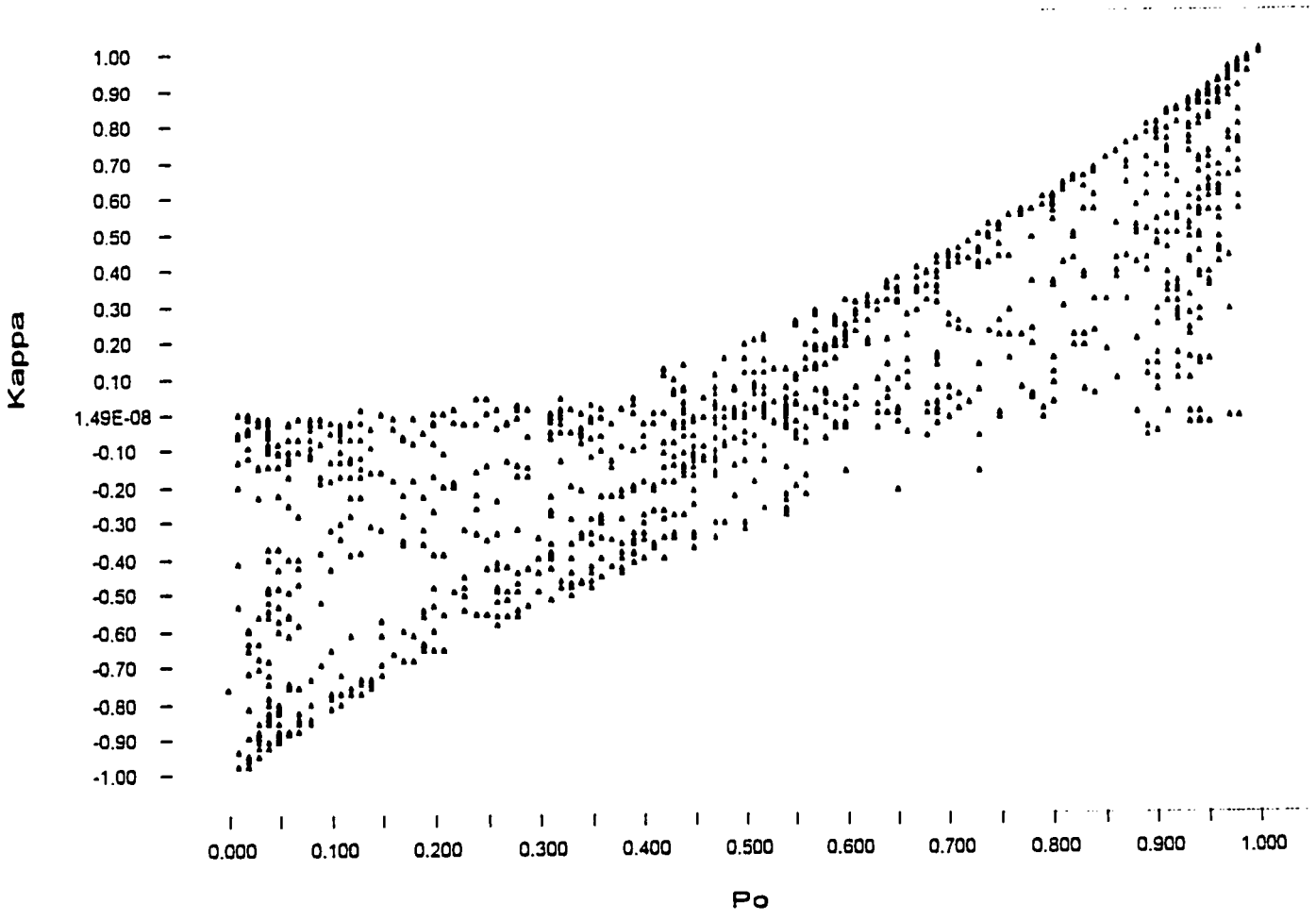


Table 21. Minimum P_e and Maximum Kappa for a Given P_o.
Minimum P_e and Maximum Kappa values were derived algebraically (see Appendices XIII and XIV) and rounded to nearest .001.

P_o *	Pe_{min} †	$Kappa_{max}$ ‡
1	.5	1
.95	.499	.900
.9	.495	.802
.85	.489	.707
.80	.480	.615
.75	.469	.529
.70	.455	.450
.65	.439	.376
.60	.420	.310
.55	.399	.252
.50	.375	.200
.45	.349	.155
.40	.320	.118
.35	.289	.086
.30	.255	.060
.25	.219	.040
.20	.180	.024
.15	.139	.013
.10	.095	.006
.05	.049	.001
0	0	0

* P_o is observed agreement.
† Pe_{min} is the minimum possible Pe value for the given P_o.
‡ $Kappa_{max}$ is maximum possible Kappa value for the given P_o.

The verification programs listed the standard error for Kappa (of Fleiss), the approximate standard error for Kappa (of Cohen) and the absolute value of their difference (SeDiff). Inspection of the listing of random two by two tables, showed that the approximate standard error by Cohen was usually close (absolute value of the difference < .02) to the standard error of Fleiss. The approximation of Cohen invariably overestimated the standard error of Fleiss. The approximation of Cohen frequently (but not always) deviated from the standard error of Fleiss when the calculated Kappa was small (< .40), N was small (< 20), or when the two by two table contained a cell with a value of < 3. A scatterplot showed how the absolute value of the difference of the two standard errors (SeDiff) could become large when Kappa is < .40 (Figure 3). Figures 4 and 5 each show scatterplots of SeDiff versus N for two by two tables with Kappa $\leq$.40 and two by two tables with Kappa >.40 respectively. For two by two tables with Kappa > .40, the SeDiff was less than .05 except at small sample sizes.

Figure 3. Scatterplot of Kappa and |SeFleiss - SeCohen| for random two by two tables.

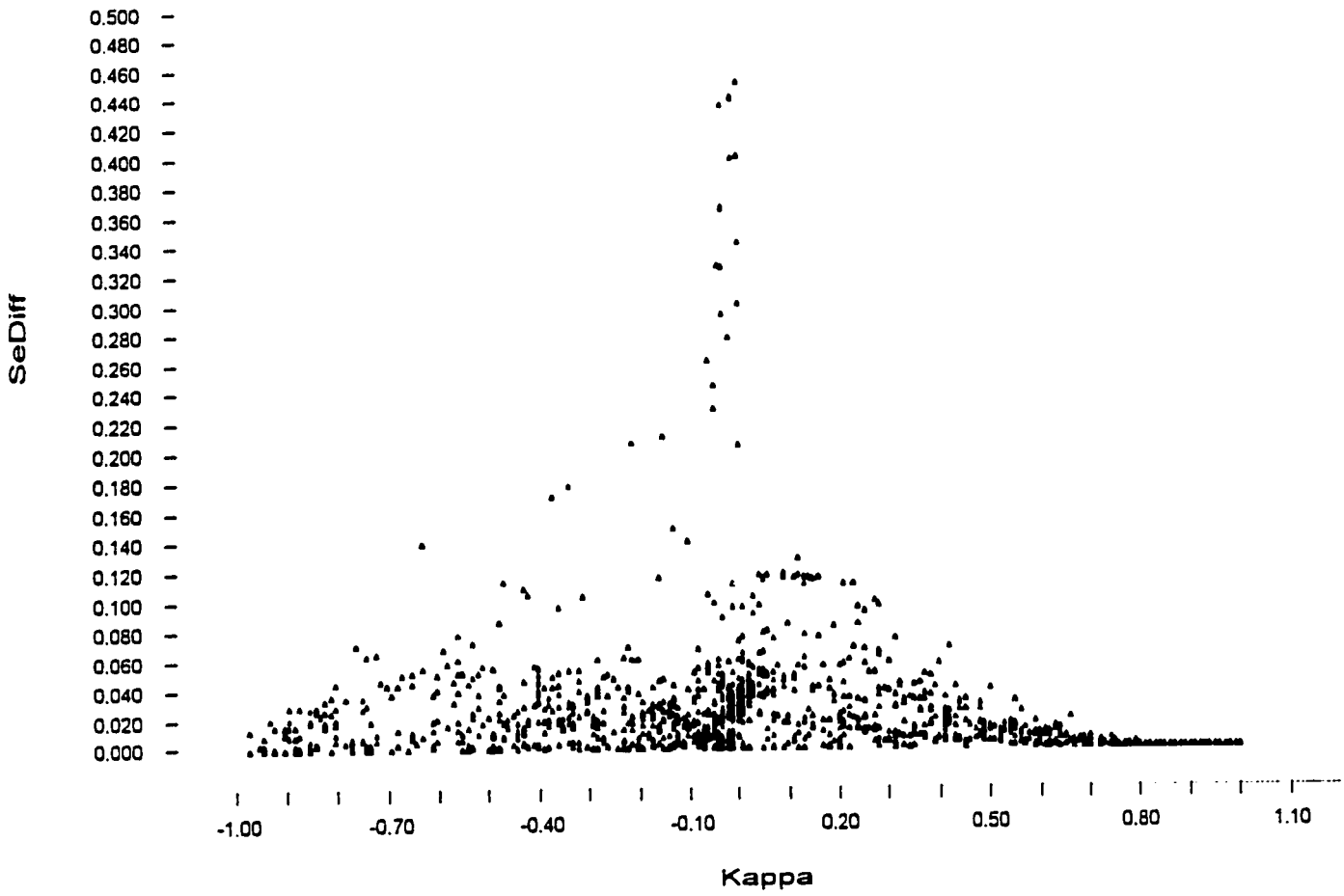


Figure 4. Scatterplot of N and |SeFleiss - SeCohen| for random two by two tables with Kappa <.40.

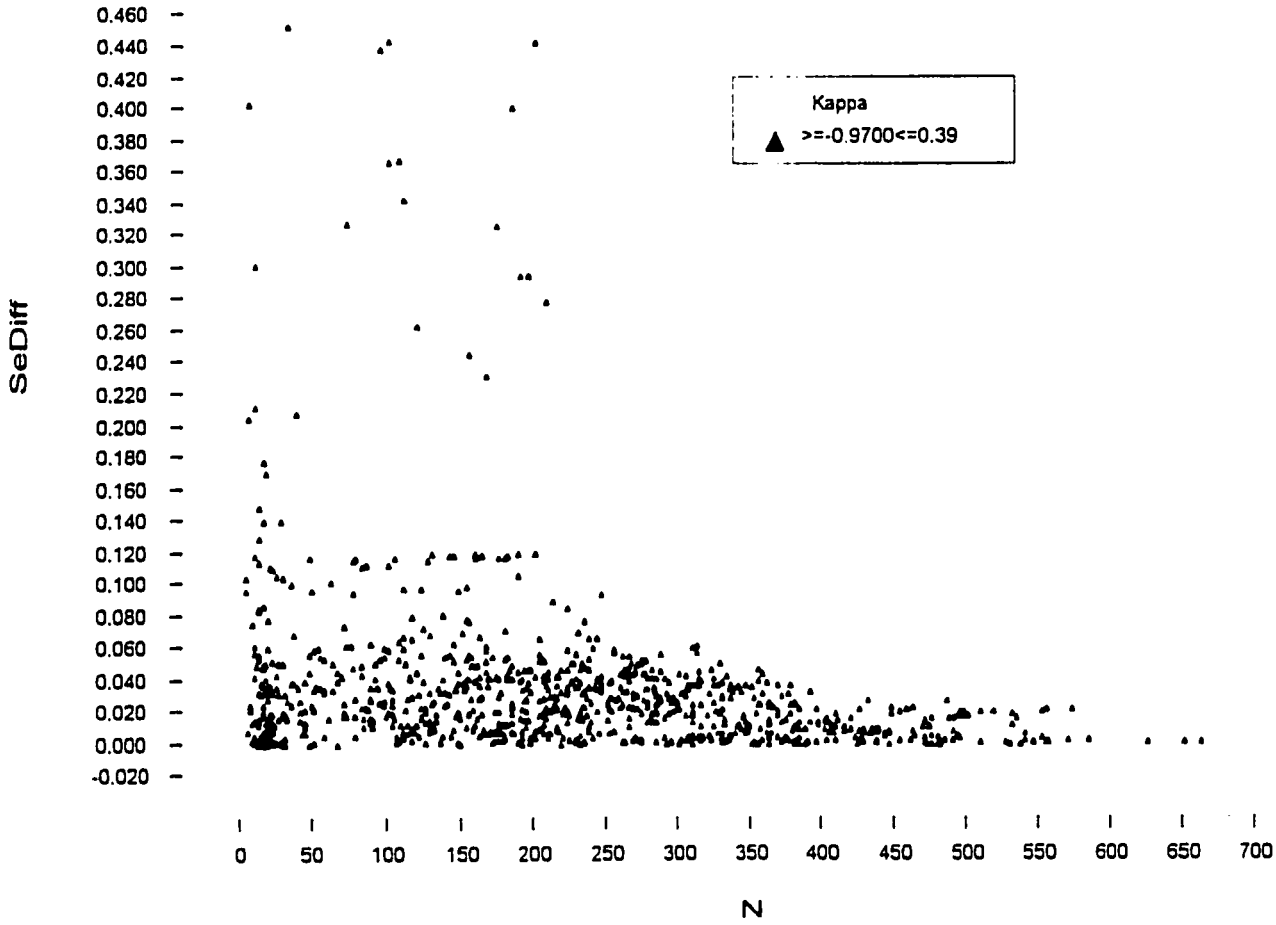
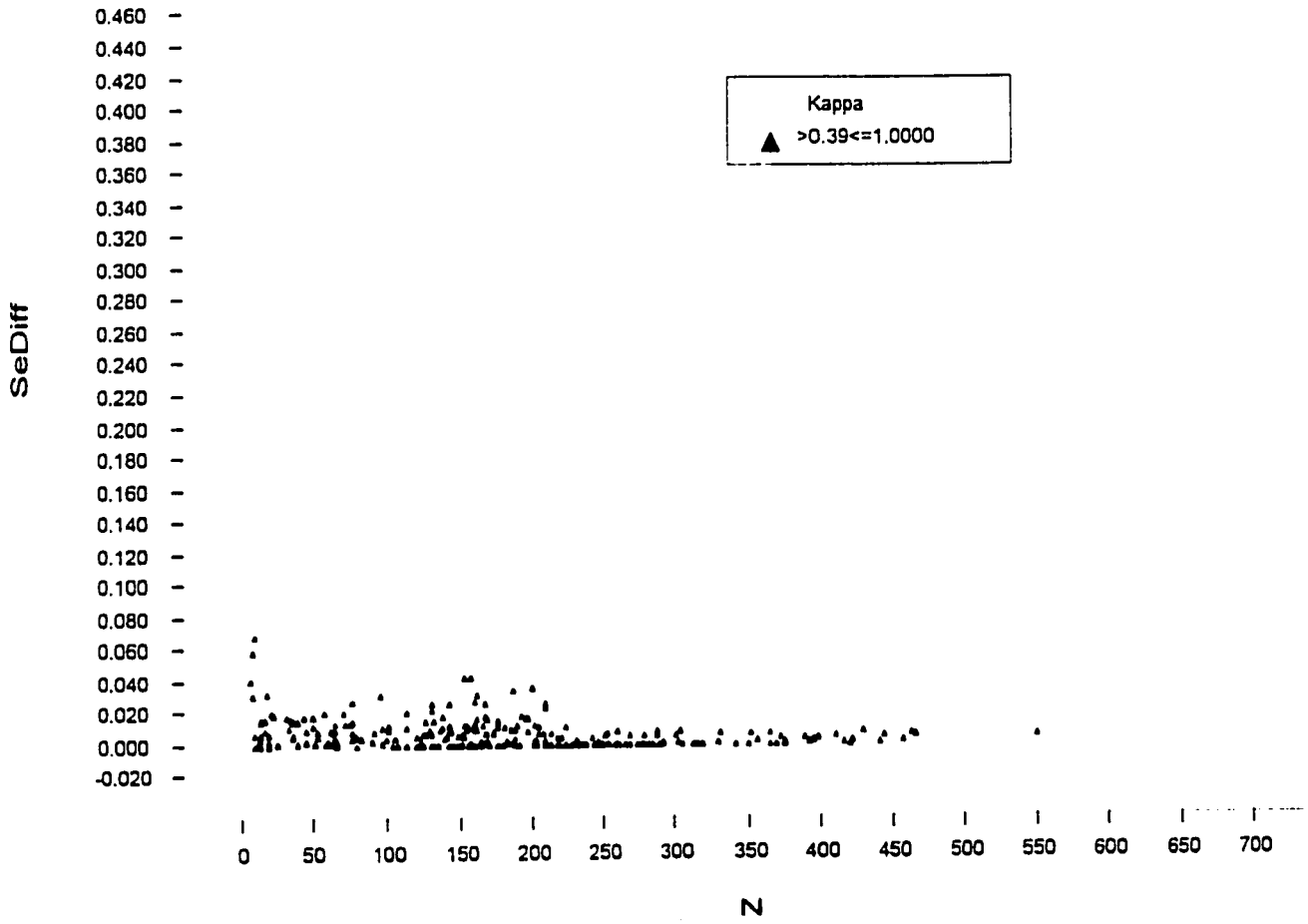


Figure 5. Scatterplot of N and |SeFleiss - SeCohen| for random two by two tables with Kappa >.40.



The conditions under which the standard error of Fleiss and the approximation of Cohen failed to hold were explored by plotting them against each other. See Figure 6 for a scatterplot of SeFleiss and SeCohen grouped by Kappa. This showed that the relationship broke down when Kappa approached 0. A scatterplot of the standard error of Fleiss versus the standard error of Cohen created from two by two tables with Kappa > .05 is shown in Figure 7. The relationhip was roughly linear with the standard error of Cohen overestimating the standard error of Fleiss especially at low Kappa values. When Kappa is greater than .40, the relationship was approximately linear. See Figure 8. The standard error of Cohen and the standard error of Fleiss were almost identical when Kappa was greater than .80. It can be easily shown algebraically that both standard errors equal 0 when observed agreement is 100%.

Figure 6. Scatterplot of SeFleiss and SeCohen for random two by two tables with Kappa > 0.

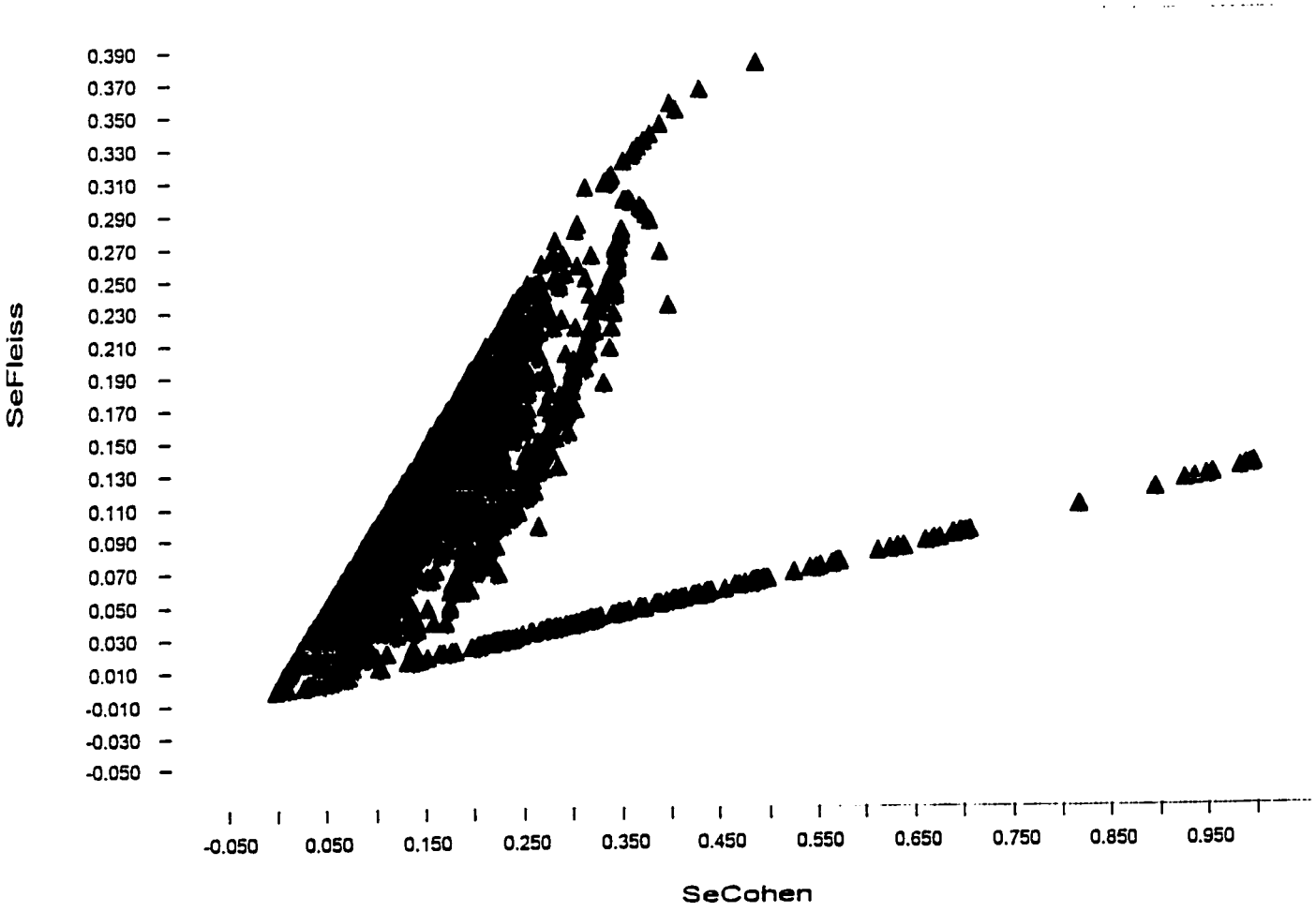


Figure 7. Scatterplot of SeFleiss and SeCohen for random two by two tables with Kappa > .05.

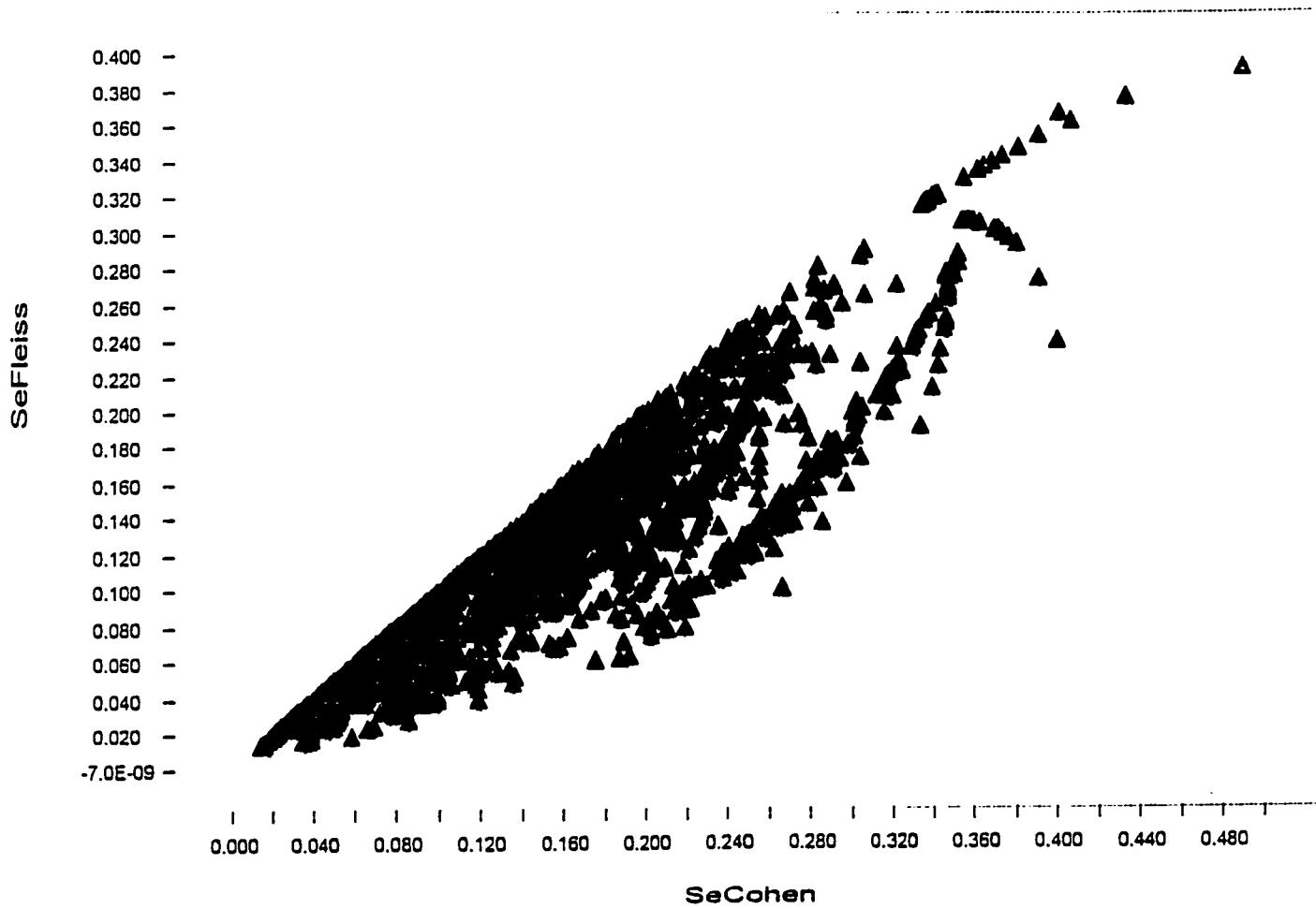
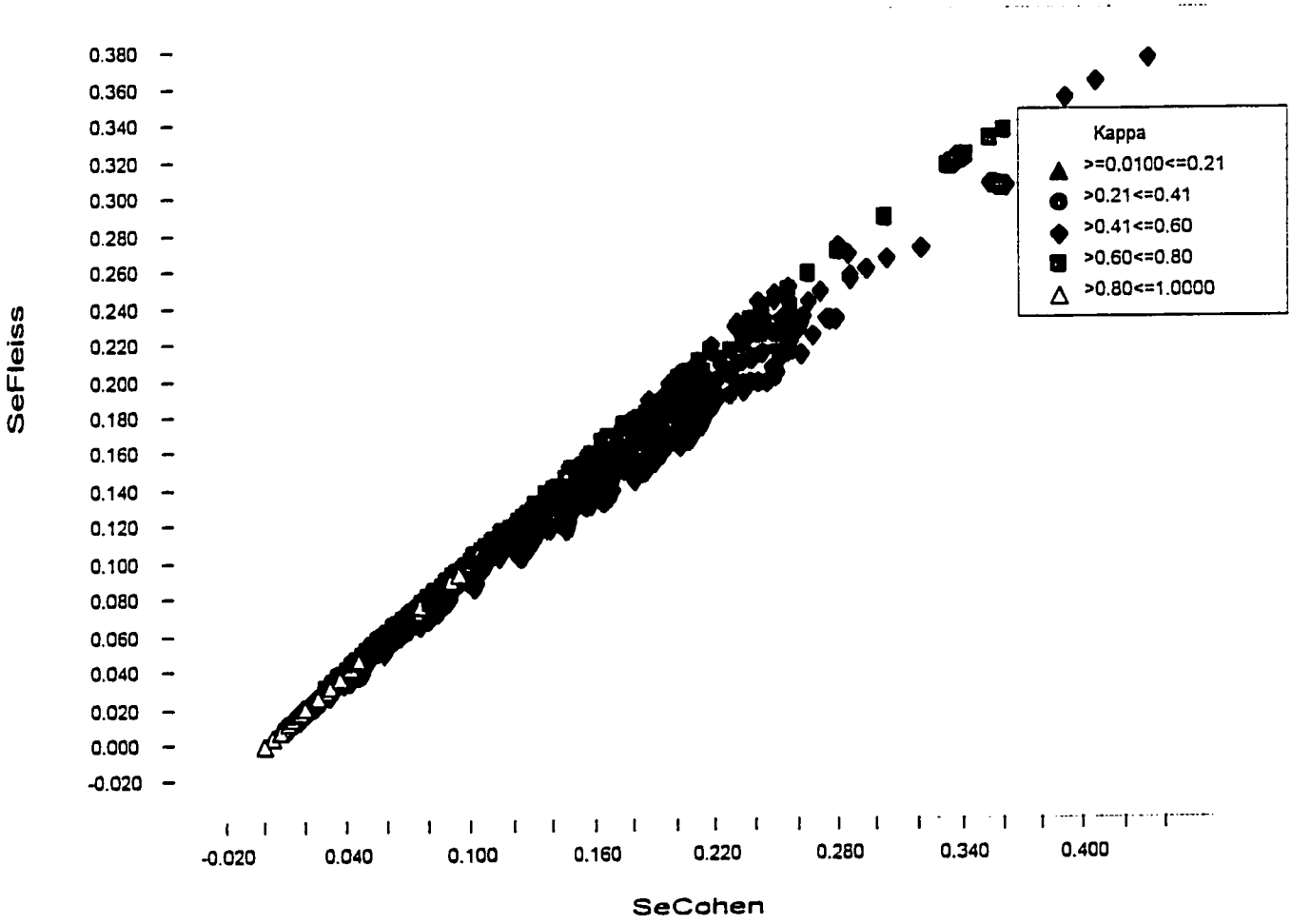


Figure 8. Scatterplot of SeFleiss and SeCohen for random two by two tables with Kappa >.40.



As a check of consistency*, the sample sizes (N) and precision (δ) obtained from the random two by two tables were compared with those N and δ obtained from the equation-derived table. Table 22 lists the random two by two tables for P_o = .80 and P_e = .50 and the random two by two tables for P_o = .90 and P_e = .50.

From the two by two tables with P_o = .80 and P_e = .50, a random number generated two by two table with N = 114 had a Kappa of .63 and a δ of .14 (giving a confidence interval of .50 - .77). The equation-derived sample size table (table 22) gave a required sample size (N) of 109.

From the two by two tables with P_o = .90, P_e = .50, a random number generated two by two table with N = 119 had a Kappa of .81 and a δ of .10 (giving a confidence interval of .71 - .92). The equation-derived sample size table (Table 20) gave a required sample size (N) of 138.

*The results are not exact because the random number generated tables have values of P_o and P_e that are rounded off.

Table 22. Listing of Random Two by Two Tables Generated with 1) P_o=.80 and P_e=.50 and 2) P_o=.90 and P_e=.50.

Po	Pe	N	Kappa	LowLim	UppLim	Delta	SeFleiss	SeCohen	SeDiff	a	b	c	d
0.8	0.5	9	0.57	0.1	1	0.47	0.2415	0.2673	0.0258	4	0	2	3
0.8	0.5	19	0.57	0.19	0.94	0.37	0.1913	0.1918	0.0006	6	2	2	9
0.8	0.5	20	0.62	0.3	0.93	0.31	0.1588	0.172	0.0132	8	4	0	8
0.8	0.5	35	0.58	0.33	0.83	0.25	0.1282	0.1411	0.0129	19	7	0	9
0.8	0.5	51	0.57	0.35	0.79	0.22	0.1122	0.1143	0.0021	21	8	3	19
0.8	0.5	61	0.55	0.34	0.76	0.21	0.1077	0.1099	0.0022	17	9	4	31
0.8	0.5	61	0.64	0.45	0.83	0.19	0.0962	0.0988	0.0027	21	2	9	29
0.8	0.5	67	0.57	0.38	0.77	0.2	0.101	0.1012	0.0002	22	7	7	31
0.8	0.5	89	0.64	0.48	0.8	0.16	0.0809	0.0813	0.0003	36	10	6	37
0.8	**0.5**	**114**	**0.63**	**0.5**	**0.77**	**0.14**	**0.0707**	**0.072**	**0.0013**	**45**	**16**	**5**	**48**
0.8	0.5	145	0.6	0.47	0.73	0.13	0.0655	0.0663	0.0007	62	20	9	54
0.8	0.5	153	0.57	0.44	0.69	0.12	0.0626	0.0656	0.0031	58	29	5	61
0.8	0.5	176	0.62	0.51	0.73	0.11	0.0561	0.0591	0.003	59	31	3	83
0.8	0.5	198	0.65	0.55	0.75	0.1	0.0506	0.0531	0.0025	88	33	2	75
0.8	0.5	229	0.58	0.47	0.69	0.11	0.0539	0.054	0	97	23	25	84
0.8	0.5	257	0.57	0.48	0.66	0.09	0.0475	0.0522	0.0047	135	53	1	68
0.8	0.5	277	0.63	0.55	0.71	0.08	0.0424	0.0456	0.0032	109	0	53	115
0.8	0.5	353	0.64	0.57	0.72	0.08	0.0393	0.0404	0.0011	146	53	11	143
0.8	0.5	373	0.6	0.52	0.67	0.07	0.0375	0.041	0.0035	175	0	77	121
0.8	0.5	448	0.57	0.5	0.64	0.07	0.0349	0.038	0.0031	155	96	4	193

Po	Pe	N	Kappa	LowLim	UppLim	Delta	SeFleiss	SeCohen	SeDiff	a	b	c	d
0.9	0.5	23	0.82	0.59	1	0.23	0.1195	0.1195	0	9	1	1	12
0.9	0.5	48	0.83	0.68	0.99	0.15	0.0787	0.0798	0.0011	20	0	4	24
0.9	0.5	70	0.83	0.7	0.96	0.13	0.0663	0.0667	0.0004	32	1	5	32
0.9	0.5	82	0.8	0.67	0.93	0.13	0.0676	0.068	0.0004	29	6	2	45
0.9	0.5	84	0.83	0.71	0.95	0.12	0.0609	0.0619	0.0009	31	0	7	46
0.9	0.5	86	0.78	0.65	0.91	0.13	0.0677	0.0694	0.0017	49	0	9	28
0.9	0.5	89	0.84	0.73	0.95	0.11	0.0568	0.0571	0.0004	38	6	1	44
0.9	0.5	90	0.77	0.64	0.9	0.13	0.0684	0.0684	0.0001	48	5	5	32
0.9	**0.5**	**119**	**0.81**	**0.71**	**0.92**	**0.1**	**0.0533**	**0.0533**	**0**	**58**	**5**	**6**	**50**
0.9	0.5	120	0.83	0.73	0.93	0.1	0.0512	0.0515	0.0003	64	8	2	46
0.9	0.5	146	0.84	0.76	0.93	0.09	0.0452	0.0452	0.0001	54	4	7	81
0.9	0.5	148	0.8	0.7	0.89	0.1	0.0497	0.0497	0	63	6	9	70
0.9	0.5	159	0.79	0.69	0.88	0.1	0.0491	0.0491	0	68	9	8	74
0.9	0.5	164	0.81	0.72	0.9	0.09	0.0471	0.0472	0	57	7	8	92
0.9	0.5	178	0.8	0.71	0.89	0.09	0.0455	0.0456	0.0001	71	12	6	89
0.9	0.5	259	0.79	0.72	0.87	0.07	0.0369	0.0376	0.0007	118	26	1	114
0.9	0.5	296	0.83	0.77	0.89	0.06	0.0324	0.0325	0	145	11	14	126
0.9	0.5	299	0.77	0.7	0.84	0.07	0.0363	0.037	0.0007	123	32	3	141

Note: Listing in bold is compared to equation-derived sample size table.

Simplification of the Sample Size Table

A simplified sample size table resulted after deleting the rows in the equation-derived sample size table that contained P_a below the algebraic minimum value for a given P_c (Table 23). Because of the large sample assumption for the normal approximation of the Kappa statistic, cells with sample sizes (N) of less than 64 were replaced with the letter m for minimum (see Appendix II). Furthermore, the asymptotic normal distribution theory will only approximate the Kappa distribution if the sample size is large when Kappa is close to 1.[133] Cells associated with degenerate confidence intervals were replaced with the letter d. For example, a cell with a Kappa of .91 and precision of ±.10 would give an impossible upper confidence limit of > 1.0.

The cells with P_c values of .97 were included to show the behaviour of the sample size requirements at very high levels of agreement. The upper confidence limit becomes degenerate at high levels of Kappa. As mentioned, the large sample assumption breaksdown as Kappa approaches 1.0 (say > .95). The Kappa values could be greater than .95 for $P_c \geq .97$.

Using the Simplified Sample Size Table

The table is used in the following manner. From a pilot or

published study giving a two by two table of agreement, the user calculates observed (P_o) and chance expected agreement (P_e). (See Appendix II for the calculation of P_o and P_e.) Along the left margin, the P_o and P_e (rounded to the nearest .05) identify a row. Along the row, the Kappa value is listed followed by the sample size N for various 95% confidence limits around the Kappa value.

Table 23. Simplified Sample Size Table for Two by Two Reliability Studies Based on Estimates of Observed (P_o) and Expected (P_e) Agreement.

P_o	P_e	Kappa	N for κ±.05	κ±.10	κ±.15	κ±.20	κ±.25	κ±.30
0.35	0.30	0.07	713	178	79	m	m	m
0.40	0.35	0.08	873	218	97	m	m	m
0.45	0.35	0.15	900	225	100	m	m	m
0.45	0.40	0.08	1056	264	117	66	m	m
0.50	0.40	0.17	1067	267	119	67	m	m
0.50	0.45	0.09	1270	317	141	79	m	m
0.55	0.40	0.25	1056	264	117	66	m	m
0.55	0.45	0.18	1257	314	140	79	m	m
0.55	0.50	0.10	1521	380	169	95	m	m
0.60	0.45	0.27	1219	305	135	76	m	m
0.60	0.50	0.20	1475	369	164	92	m	m
0.60	0.55	0.11	1821	455	202	114	73	m
0.65	0.45	0.36	1156	289	128	72	m	m
0.65	0.50	0.30	1398	350	155	87	m	m
0.65	0.55	0.22	1726	432	192	108	69	m
0.65	0.60	0.13	2185	546	243	137	87	m
0.70	0.50	0.40	1291	323	143	81	m	m
0.70	0.55	0.33	1594	398	177	100	64	m
0.70	0.60	0.25	2017	504	224	126	81	m
0.70	0.65	0.14	2634	659	293	165	105	73
0.75	0.50	0.50	1152	288	128	m	m	m
0.75	0.55	0.44	1423	356	158	89	m	m
0.75	0.60	0.38	1801	450	200	113	72	m
0.75	0.65	0.29	2352	588	261	147	94	65
0.75	0.70	0.17	3201	800	356	200	128	89
0.80	0.50	0.60	983	246	109	m	m	m
0.80	0.55	0.56	1214	304	135	76	m	m
0.80	0.60	0.50	1537	384	171	96	m	m
0.80	0.65	0.43	2007	502	223	125	80	m
0.80	0.70	0.33	2732	683	304	171	109	76
0.80	0.75	0.20	3934	983	437	246	157	109
0.85	0.50	0.70	784	196	87	m	m	m
0.85	0.55	0.67	968	242	108	m	m	m
0.85	0.60	0.63	1225	306	136	77	m	m
0.85	0.65	0.57	1599	400	178	100	m	m
0.85	0.70	0.50	2177	544	242	136	87	m
0.85	0.75	0.40	3135	784	348	196	125	87
0.85	0.80	0.25	4898	1225	544	306	196	136

0.90	0.50	0.80	553	138	m	m	m,d†	m,d
0.90	0.55	0.78	683	171	76	m	m,d	m,d
0.90	0.60	0.75	864	216	96	m	m	m,d
0.90	0.65	0.71	1129	282	125	71	m	m,d
0.90	0.70	0.67	1537	384	171	96	m	m
0.90	0.75	0.60	2213	553	246	138	89	m
0.90	0.80	0.50	3457	864	384	216	138	96
0.90	0.85	0.33	6147	1537	683	384	246	171
0.95	0.50	0.90	292	73	m,d	m,d	m,d	m,d
0.95	0.55	0.89	360	90	m,d	m,d	m,d	m,d
0.95	0.60	0.88	456	114	m,d	m,d	m,d	m,d
0.95	0.65	0.86	596	149	m,d	m,d	m,d	m,d
0.95	0.70	0.83	811	203	90	m,d	m,d	m,d
0.95	0.75	0.80	1168	292	130	73	m,d	m,d
0.95	0.80	0.75	1825	456	203	114	73	m,d
0.95	0.85	0.67	3244	811	360	203	130	90
0.95	0.90	0.50	7299	1825	811	456	292	203

P_o	P_e	Kappa	N for	k±.05	k±.10	k±.15	k±.20	k±.25	k±.30
0.97	0.50	0.94		179	m,d	m,d	m,d	m,d	m,d
0.97	0.55	0.93		221	m,d	m,d	m,d	m,d	m,d
0.97	0.60	0.92		279	m,d	m,d	m,d	m,d	m,d
0.97	0.65	0.91		365	m,d	m,d	m,d	m,d	m,d
0.97	0.70	0.90		497	124	m,d	m,d	m,d	m,d
0.97	0.75	0.88		715	179	m,d	m,d	m,d	m,d
0.97	0.80	0.85		1118	279	124	m,d	m,d	m,d
0.97	0.85	0.80		1987	497	221	124	m,d	m,d
0.97	0.90	0.70		4472	1118	497	279	179	124
0.97	0.95	0.40		17886	4472	1987	1118	715	497

*m = minimum sample size of 64 subjects.
†d = degenerate upper confidence interval (upper limit of interval > 1.0).

Appendix VIII. Computer Program for Equation-derived Sample Size Table

```
//   kaptable.c
//  A program that calculates N for various Po and Pe
//  using Cohens approximation for the standard error of Kappa.
#include <stdio.h>
#include <math.h>
main()
{
 double Pe;
 double Po;
 double kappa;
 double samplesize5;
 double samplesize10;
 double samplesize15;
 double samplesize20;
 double samplesize25;
 double samplesize30;
 double delta;
Po =.10;
while (Po < 1.05)
 {
 printf( "\n Po   Pe  kappa  N for  k+/-.05   k+/-.10  k+/-.15  k+/-.20 k+/-.25 k+/-.30 \n");
 Pe = 0;
 while (Pe < Po)
        {
        delta = .05;
        samplesize5 = pow(1.96,2)*Po*(1-Po)/(pow(delta,2)*pow(1-Pe,2));
        delta = .10;
        samplesize10 = pow(1.96,2)*Po*(1-Po)/(pow(delta,2)*pow(1-Pe,2));
        delta = .15;
        samplesize15 = pow(1.96,2)*Po*(1-Po)/(pow(delta,2)*pow(1-Pe,2));
        delta = .20;
        samplesize20 = pow(1.96,2)*Po*(1-Po)/(pow(delta,2)*pow(1-Pe,2));
        delta = .25;
        samplesize25 = pow(1.96,2)*Po*(1-Po)/(pow(delta,2)*pow(1-Pe,2));
        delta = .30;
        samplesize30 = pow(1.96,2)*Po*(1-Po)/(pow(delta,2)*pow(1-Pe,2));
        kappa =  (Po - Pe) / (1 - Pe);
        printf(" %.2f %.2f %.2f        %.0f   %.0f   %.0f   %.0f      %.0f        %.0f \n", Po,
Pe, kappa, samplesize5, samplesize10, samplesize15, samplesize20, samplesize25, samplesize30 );
        Pe = Pe + .05;
        }
 printf("\n");
 Po = Po + .05;
 }
// Added to expand table at high levels of simple agreement.
Po =.96;
while (Po < 1.01)
 {
 printf( "\n Po   Pe  kappa  N for  k+/-.05   k+/-.10  k+/-.15  k+/-.20  k+/-.25  k+/-.30 \n");
 Pe = 0;
 while (Pe < Po)
        {
        delta = .05;
        samplesize5 = pow(1.96,2)*Po*(1-Po)/(pow(delta,2)*pow(1-Pe,2));
        delta = .10;
        samplesize10 = pow(1.96,2)*Po*(1-Po)/(pow(delta,2)*pow(1-Pe,2));
        delta = .15;
        samplesize15 = pow(1.96,2)*Po*(1-Po)/(pow(delta,2)*pow(1-Pe,2));
        delta = .20;
        samplesize20 = pow(1.96,2)*Po*(1-Po)/(pow(delta,2)*pow(1-Pe,2));
        delta = .25;
        samplesize25 = pow(1.96,2)*Po*(1-Po)/(pow(delta,2)*pow(1-Pe,2));
        delta = .30;
        samplesize30 = pow(1.96,2)*Po*(1-Po)/(pow(delta,2)*pow(1-Pe,2));
        kappa =  (Po - Pe) / (1 - Pe);
        printf(" %.2f %.2f %.2f        %.0f       %.0f      %.0f        %.0f      %.0f
%.0f \n", Po, Pe, kappa, samplesize5, samplesize10, samplesize15, samplesize20, samplesize25,
samplesize30 );
        Pe = Pe + .05;
        }
 printf("\n");
 Po = Po + .01;
 }
}
```

Appendix IX. Computer Generation of Two by Two Tables for Kappa

Computer programs were written to generate a wide variety of two by two tables with sample sizes (potentially ranging from 0 to over 500). For each two by two table created, there was a listing of following: proportion of observed agreement (P_o), chance expected agreement (P_e), sample size (N), Kappa , 95% confidence lower (LowLim) and upper limits (UpperLim), standard error for Kappa of Fleiss (SeFleiss), approximation of standard error for Kappa of Cohen (SeCohen), absolute value of the difference of the two standard errors (SeDiff), and the actual two by two cell values (a, b, c, and d of the two by two table in Appendix II).

Patterned Random Number Method

This method attempted to increase the number of asymmetrical and unbalanced two by two tables. This program, kaprand.c (Appendix X), used a random number generator to give random numbers for each cell from either of two ranges. The range was specified as either large (between 0 and 250) or small (between 0 and 10). The various patterns of the two by two tables have been described previously.[112] The twelve patterns were as follows[x]:

---

[x]There are actually sixteen possible permutations of the two by two table with either a large or a small number in each cell. The P_o, P_e and Kappa for the two by two are the same when the values

1. All cells had large ranges. In this pattern, the observer agreement was random in a large sample size.

2. All cells had small ranges. In this pattern, the observer agreement was random in a small sample size.

3. Agreement cells (a and d) both had large ranges.[y] Observers generally agreed on normality and abnormality.

4. Disagreement cells (b and c) had large ranges. The observers generally agreed to disagree.

5. Cell "a" had a large range. Observers agreed on a "high prevalence" of normality.

6. Cell "d" had a large range. Observers agreed on a low prevalence of normality (or a high prevalence of abnormality).

7. One of the disagreement cells had a large range. The observers differed in their frequency of rating of normality. There has been termed "bias" in that the marginal proportions are different between the observers.[77]

8. The agreement on normality cell and either one of the disagreement cells had large ranges. This was a combination of the "high prevalence" and "bias".[77]

9. The agreement on abnormality cell and either one of the disagreement cells had large ranges.

10. The agreement on normality cell and both disagreement cells had large ranges.

11. The agreement on abnormality cell and both disagreement

---

in "b cell" and "c cell" are interchanged.

[y]Others have small ranges.

cells had large ranges.

12. Both the normality and abnormality cells and either one of the disagreement cells had large ranges.

(The scatterplot in Figure 1 plots P_o versus P_e using the patterned random number generated two by two tables.)

Simple random number method

This simple method used a random number generator to give random numbers within a range (0 to 250) for each of the four cells in the two by two table. The program kapsimp.c is listed in listed Appendix XI. This method was less satisfactory than the patterned method because it gave relatively few two by two tables at the extremes of P_o. See the scatterplot in Figure 9.

Permutation method

This method took all the permutations of the two by two tables with sample size N equalling 0 to 10. These permutations of the two by two tables (there were over a 1000 of them) were listed in a computer text file that was used by the kapperm.c program (Appendix XII) to calculate the values for P_o, P_e and Kappa. This method was less satisfactory than the patterned method because it was limited to small sample sizes but it did

include extreme values such as P_{σ} = 1.0 and -1.0. See scatterplot in Figure 10.

Figure 9. Scatterplot of P_o and P_e from computer-generated two by two tables (simple random number method).

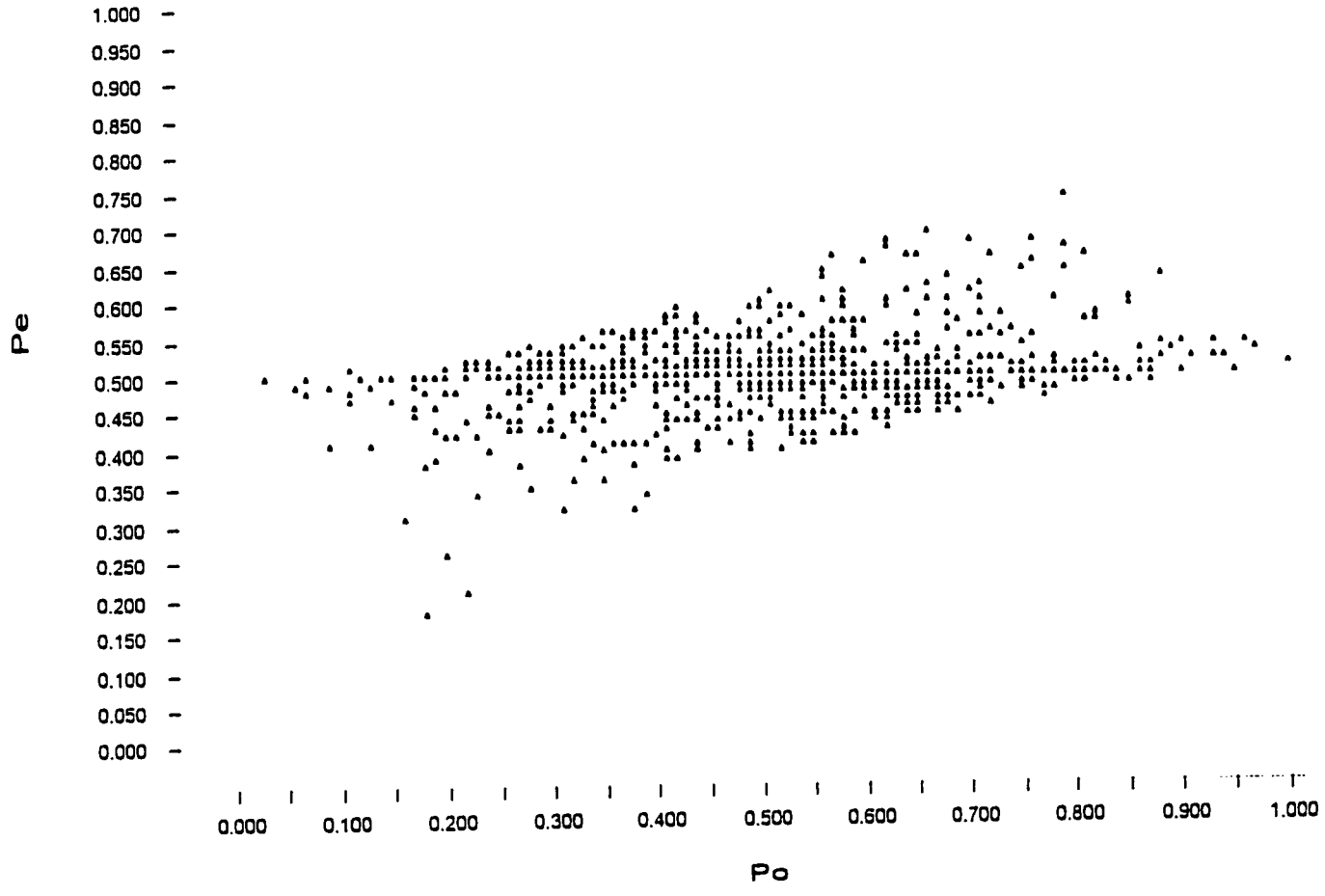
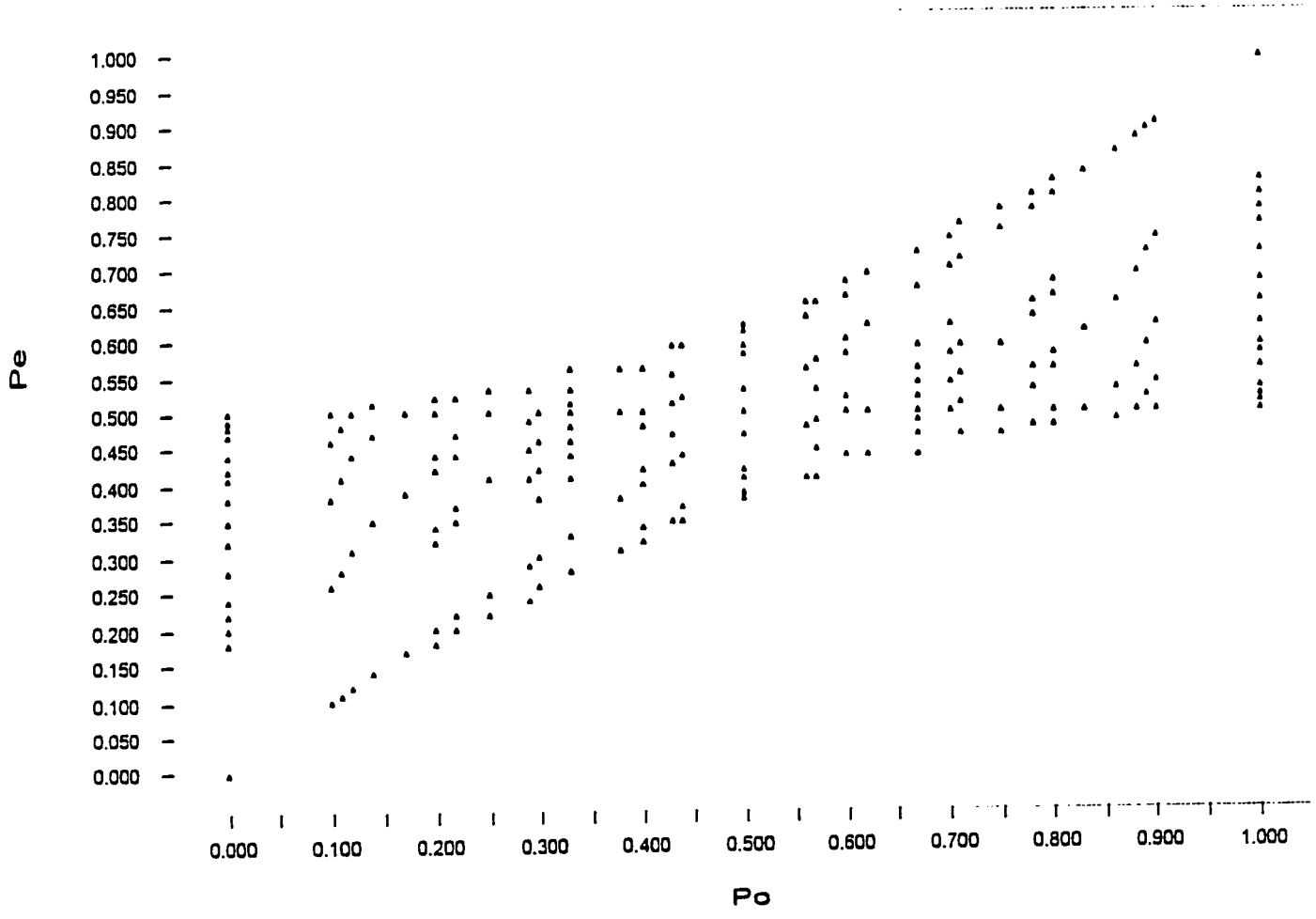


Figure 10. Scatterplot of P_o and P_e from computer-generated two by two tables (permutation method).



Appendix X. Computer Program for Generation of Random Two by Two Tables by Random Patterned Numbers.

```
//   kaprand.c
//   March 1997
// This program randomly generates two by two tables
// and lists Po, Pe, Kappa, N, confidence limits, the standard error
// of Fleiss, the approximation of the standard error of Cohen and the
// absolute value of their difference.

#include <stdio.h>
#include <math.h>
#include <stdlib.h>
#include <time.h>


double roundoff(double Px)
//This function rounds off the proportions to the nearest .05.
{
double x, xdiff;
Px = Px * 10;
 x = floor(Px);
 xdiff = Px - x;
 if (xdiff < .25)
   Px = x;
 if (xdiff >= .25 && xdiff <= .75)
   Px = x + .5;
 if (xdiff >= .75)
   Px = x + 1;
   Px = Px/10;
 return Px;
}

void generate2x2table (double a,double b, double c, double d)
// This function constructs the two by two table
// and computes the statistics.
{
 double N ;
 double p1;
 double p2;
 double q1;
 double q2;
 double Pe, rPe ;
 double Po, rPo;
 double kappa;
 double samplesize;
 double delta;
 double A;
 double B;
 double C;
 double sekappafleiss;
 double sekappacohen;
 double sediff;
 double upperlim;
 double lowerlim;

 N = a + b + c + d;
 p1 = (a + b ) / N;
 p2 = (a + c ) / N;
 q1 = (c + d ) / N;
 q2 = (b + d ) / N;
 // These lines can be used to draw the two by two table
 // on the computer print out.
 //printf("_____| \n");
 //printf("|     %.0f   %.0f  | \n",a,  b );
 //printf("|                  |\n");
 //printf("|     %.0f   %.0f  | \n",c,  d );
 //printf("|_____|  %.0f\n", N);

 // Calculation of Po, Pe, Kappa
 Po = (a + d) / N;
 Pe = (p1 * p2) +  (q1 * q2);
 kappa =  (Po - Pe) / (1 - Pe);

 // Calculation of se for C.I. as per Fleiss
    A = (a/N) *   pow( 1 -  (p1+p2)*(1-kappa) , 2 ) +
        (d/N) *   pow( 1  - (q1+q2)*(1-kappa)   , 2 );
```

```
    B = pow(  (1 -kappa),2 ) *
        (
        b/N *    pow(  ( p2 + q1) ,  2)  +
        c/N *    pow(  ( p1 + q2) ,  2)

        )       ;

     C = pow (  ( kappa-Pe*(1-kappa))    , 2 ) ;

 sekappafleiss = sqrt ( A + B - C) *  (1/(1-Pe)) *  1/(sqrt(N));

 // Cohen approximation of the standard error of Kappa.
 sekappacohen = sqrt((Po - (pow( Po, 2))) / ( N * (pow (1-Pe, 2))));

 // Calculation of the lower and upper confidence limits.
 lowerlim = kappa - sekappafleiss*1.96;
 upperlim = kappa + sekappafleiss*1.96;


// Round off the Po and Pe before a sort
 rPe = roundoff(Pe);
 rPo = roundoff(Po);

 // Calculation of delta and the difference between the apporximation
 // and the standard error given by Fleiss.
 delta = sekappafleiss*1.96;
 sediff = fabs(sekappafleiss - sekappacohen);

 // List all of the values in a line.
 printf("%.2f,%.2f,%.0f,%.2f,%.2f,%.2f,%.2f,%.4f,%.4f,%.4f,%.0f,%.0f,%.0f,%.0f\n",
         Po,  Pe, N, kappa, lowerlim, upperlim, delta, sekappafleiss, sekappacohen, sediff,
a,b,c,d);

}

// Main program generating two by two tables in
// the various configurations.

main()
{
 int randnum;
 int i;
 const int largerange = 250;
 const int smallrange = 10;
 const int iterations = 100;
 double randoma;
 double randomb;
 double randomc;
 double randomd;
 randomize();

// 1. All cells have large ranges.

 for (i=1;i< iterations;i++)
 {
 randoma = random (largerange);
 randomb = random (largerange);
 randomc = random (largerange);
 randomd = random (largerange);
 generate2x2table (randoma, randomb, randomc, randomd);
 }

// 2. All cells have small ranges.

 for (i=1;i< iterations;i++)
 {
 randoma = random (smallrange);
 randomb = random (smallrange);
 randomc = random (smallrange);
 randomd = random (smallrange);
 generate2x2table (randoma, randomb, randomc, randomd);
 }


// 3. Agreement cells (a,d) have large ranges.

 for (i=1;i< iterations;i++)
 {
```

```
 randoma = random (largerange);
 randomb = random (smallrange);
 randomc = random (smallrange);
 randomd = random (largerange);
 generate2x2table (randoma, randomb, randomc, randomd);
 }

// 4. Disagreement cells (b,c) have large ranges

 for (i=1;i< iterations;i++)
 {
 randoma = random (smallrange);
 randomb = random (largerange);
 randomc = random (largerange);
 randomd = random (smallrange);
 generate2x2table (randoma, randomb, randomc, randomd);
 }

// 5. Agreement with a high prevalence of normality.
// Agreement on high prevalence of normality.
 for (i=1;i< iterations;i++)
 {
 randoma = random (largerange);
 randomb = random (smallrange);
 randomc = random (smallrange);
 randomd = random (smallrange);
 generate2x2table (randoma, randomb, randomc, randomd);
 }

// 6. Agreement with a low prevalence of normality.
// Agreement on high prevalence of abnormality.

 for (i=1;i< iterations;i++)
 {
 randoma = random (smallrange);
 randomb = random (smallrange);
 randomc = random (smallrange);
 randomd = random (largerange);
 generate2x2table (randoma, randomb, randomc, randomd);
 }

// 7. One of the disagreement cells has a large range.

for (i=1;i< iterations;i++)
 {
 randoma = random (smallrange);
 randomb = random (largerange);
 randomc = random (smallrange);
 randomd = random (smallrange);
 generate2x2table (randoma, randomb, randomc, randomd);
 }

// 8. The agreement on normality cell and
// one of the disagreement cells have large ranges.

for (i=1;i< iterations;i++)
 {
 randoma = random (largerange);
 randomb = random (largerange);
 randomc = random (smallrange);
 randomd = random (smallrange);
 generate2x2table (randoma, randomb, randomc, randomd);
 }

// 9. The agreement on abnormality cell and
// one of the disagreement cells have large ranges.

for (i=1;i< iterations;i++)
 {
 randoma = random (smallrange);
 randomb = random (largerange);
 randomc = random (smallrange);
 randomd = random (largerange);
 generate2x2table (randoma, randomb, randomc, randomd);
 }

// 10. The agreement on normality and both of the
// disagreement cells have large ranges.
```

```
for (i=1;i< iterations;i++)
 {
 randoma = random (largerange);
 randomb = random (largerange);
 randomc = random (largerange);
 randomd = random (smallrange);
 generate2x2table (randoma, randomb, randomc, randomd);
 }

// 11. The agreement on abnormality and both the
// disagreement cells have large ranges.

for (i=1;i< iterations;i++)
 {
 randoma = random (smallrange);
 randomb = random (largerange);
 randomc = random (largerange);
 randomd = random (largerange);
 generate2x2table (randoma, randomb, randomc, randomd);
 }

// 12. The agreement on normality, agreement on abnormality, and
// one of the disagreement cells have large ranges.

for (i=1;i< iterations;i++)
 {
 randoma = random (largerange);
 randomb = random (largerange);
 randomc = random (smallrange);
 randomd = random (largerange);
 generate2x2table (randoma, randomb, randomc, randomd);
 }

}
```

Appendix XI. Computer Program for Generation of Random Two by Two Tables by Simple Random Numbers.

```
//   kapsimp.c
//   March 1997
// This program randomly generates two by two tables
// and lists Po, Pe, Kappa, N, confidence limits, the standard error
// of Fleiss, the approximation of the standard error of Cohen and the
// absolute value of their difference.

#include <stdio.h>
#include <math.h>
#include <stdlib.h>
#include <time.h>


double roundoff(double Px)
//This function rounds off the proportions to the nearest .05.
{
double x, xdiff;
Px = Px * 10;
 x = floor(Px);
 xdiff = Px - x;
 if (xdiff < .25)
   Px = x;
 if (xdiff >= .25 && xdiff <= .75)
   Px = x + .5;
 if (xdiff >= .75)
   Px = x + 1;
   Px = Px/10;
 return Px;
}

void generate2x2table (double a,double b, double c, double d)
// This function constructs the two by two table
// and computes the statistics.
{
 double N ;
 double p1;
 double p2;
 double q1;
 double q2;
 double Pe, rPe ;
 double Po, rPo;
 double kappa;
 double samplesize;
 double delta;
 double A;
 double B;
 double C;
 double sekappafleiss;
 double sekappacohen;
 double sediff;
 double upperlim;
 double lowerlim;

 N = a + b + c + d;
 p1 = (a + b ) / N;
 p2 = (a + c ) / N;
 q1 = (c + d ) / N;
 q2 = (b + d ) / N;
 // These lines can be used to draw the two by two table
 // on the computer print out.
 //printf("_____| \n");
 //printf("|     %.0f   %.0f  | \n",a,  b );
 //printf("|                  |\n");
 //printf("|     %.0f   %.0f  | \n",c,  d );
 //printf("|_____|  %.0f\n", N);

 // Calculation of Po, Pe, Kappa
 Po = (a + d) / N;
 Pe = (p1 * p2) +  (q1 * q2);
 kappa =  (Po - Pe) / (1 - Pe);

 // Calculation of se for C.I. as per Fleiss
    A = (a/N) *   pow( 1 -  (p1+p2)*(1-kappa) , 2 ) +
        (d/N) *   pow( 1  - (q1+q2)*(1-kappa)   , 2 );
```

```
    B = pow(  (1 -kappa),2 ) *
        (
        b/N *    pow(  ( p2 + q1) ,  2)  +
        c/N *    pow(  ( p1 + q2) ,  2)

        )       ;

     C = pow (  ( kappa-Pe*(1-kappa))    , 2 ) ;

 sekappafleiss = sqrt ( A + B - C) *  (1/(1-Pe)) *  1/(sqrt(N));

 // Cohen approximation of the standard error of Kappa.
 sekappacohen = sqrt((Po - (pow( Po, 2))) / ( N * (pow (1-Pe, 2))));

 // Calculation of the lower and upper confidence limits.
 lowerlim = kappa - sekappafleiss*1.96;
 upperlim = kappa + sekappafleiss*1.96;


// Round off the Po and Pe before a sort
 rPe = roundoff(Pe);
 rPo = roundoff(Po);

 // Calculation of delta and the difference between the apporximation
 // and the standard error given by Fleiss.
 delta = sekappafleiss*1.96;
 sediff = fabs(sekappafleiss - sekappacohen);

 // List all of the values in a line.
 printf("%.2f,%.2f,%.0f,%.2f,%.2f,%.2f,%.2f,%.4f,%.4f,%.4f,%.0f,%.0f,%.0f,%.0f\n",
        Po,  Pe, N, kappa, lowerlim, upperlim, delta, sekappafleiss, sekappacohen, sediff,
a,b,c,d);

}

// Main program generating two by two tables in
// the various configurations.

main()
{
 int randnum;
 int i;
 const int range = 200;
 const int iterations = 1000;
 double randoma;
 double randomb;
 double randomc;
 double randomd;
 randomize();

// 1. All cells have large ranges.

 for (i=1;i< iterations;i++)
 {
 randoma = random (range);
 randomb = random (range);
 randomc = random (range);
 randomd = random (range);
 generate2x2table (randoma, randomb, randomc, randomd);
 }
}
```

Appendix XII. Computer Program for Generation of Random Two by Two Tables by Permutation.

```
//      kapperm.c
//      Jan 1997
//
#include <stdio.h>
#include <string.h>
#include <math.h>

// function to round off values of Po and Pe.
double roundoff(double Px)
{
double x, xdiff;
Px = Px * 10;
 x = floor(Px);
 xdiff = Px - x;
 if (xdiff < .25)
   Px = x;
 if (xdiff >= .25 && xdiff <= .75)
   Px = x + .5;
 if (xdiff >= .75)
   Px = x + 1;
   Px = Px/10;
 return Px;
}

// function to calculate Po, Pe, Kappa, 95% confidence limits,
// standard error of Fleiss, approximate standard error of Cohen
// difference of the two standard errors

kappacalc(double yy, double yn, double ny, double nn)
{
 double N ;
 double a1;
 double a2;
 double b1;
 double b2;
 double Pe, rPe ;
 double Po, rPo;
 double kappa;
 double sumterm;
 double Pesquared;
 double ComplementPesquared;
 double zerovarkappa;
 double zerosekappa;
 double sekappacohen;
 double sediff;
 double A;
 double B;
 double C;
 double sekappafleiss;
 double upperlim;
 double lowerlim;
 double samplesize;
 double delta;
 N = yy + yn + ny + nn;
 if (N < 1)
   printf("N too small\n");
 else
  {
 a1 = (yy + yn) / N;
 a2 = (ny + nn) / N;
 b1 = (yy + ny) / N;
 b2 = (yn + nn) / N;


// printf("               \n");
// printf("                          \n");
// printf("              %.0f      %.0f        a1= %.2f \n",yy, yn, a1);
// printf("              %.0f      %.0f        a2= %.2f \n", ny, nn, a2);
// printf("                                       \n");
// printf("        b1=%.2f b2=%.2f   N= %.0f  \n", b1, b2, N);

 // Pe and Po calculation
 Po = (yy + nn) / N;
 Pe = (a1 * b1) +  (a2 * b2);
```

```
 // need to handle special case of Pe = 1;
 if (Pe == 1)  Pe = .99;


// Kappa calculation
 kappa = (Po - Pe) / (1 - Pe);

// need to handle special case of Kappa = 0
 if (kappa == 0) kappa = .01;

 sumterm =(a1*b1*(a1+b1)) + (a2*b2*(a2+b2));
 Pesquared = pow (Pe, 2);
 ComplementPesquared = pow( (1 - Pe), 2);

 A = (yy/N) *   pow( 1 -  (a1+b1)*(1-kappa) , 2 ) +
     (nn/N) *   pow( 1 -  (a2+b2)*(1-kappa)   , 2 );

 B = pow(  (1 -kappa),2 ) *
       (
       yn/N *    pow(  ( b1 + a2) ,  2)  +
       ny/N *    pow(  ( b2 + a1) ,  2)
       );

 C = pow (  ( kappa-Pe*(1-kappa))   , 2 ) ;
  if ((A + B - C) > 0)
      {
      sekappafleiss = sqrt ( A + B - C) *  (1/(1-Pe)) *  1/(sqrt(N));
      }
  else
     sekappafleiss = 0;

 // Cohen approximation
 sekappacohen = sqrt((Po - (pow( Po, 2))) / ( N * (pow (1-Pe, 2))));

 // round off the Po and Pe
 rPe = roundoff(Pe);
 rPo = roundoff(Po);
 delta = sekappafleiss*1.96;
 sediff = fabs(sekappafleiss - sekappacohen);


 //95% confidence limits around kappa.
 upperlim = kappa + sekappafleiss*1.96;
 lowerlim = kappa - sekappafleiss*1.96;
 if( upperlim > 1)
   upperlim = 1;
 if (lowerlim < -1)
    lowerlim = -1;

printf("%.2f,%.2f,%.0f,%.2f,%.2f,%.2f,%.2f,%.4f,%.4f,%.4f,%.0f,%.0f,%.0f,%.0f\n",
        Po, Pe, N, kappa, lowerlim,upperlim,delta, sekappafleiss, sekappacohen,
sediff,yy,yn,ny,nn );


 } // end of else
}


main()
{
 FILE *infile;
 char string[81];
 char a[3];
 char b[3];
 char c[3];
 char d[3];
 float Acell;
 float Bcell;
 float Ccell;
 float Dcell;
 // Open input file with all the permutations of the two by two
 // with sample sizes ranging from 1..10.
 if ((infile = fopen("c:\\tc\\permut.txt", "r" ))
   == NULL)
      {
      printf("fopen failed.\n");
      }
```

```
while (fgets(string, 80, infile) !=NULL)
  {
   // printf("\n\nFile input\n");
  //  printf("0123456789\n");
  //  printf("%s\n", string);
    a[0] = string[0];
    a[1] = string[1];
    a[2] = '/0';
    b[0] = string[3];
    b[1] = string[4];
    b[2] = '/0';
    c[0] = string[6];
    c[1] = string[7];
    c[2] = '/0';
    d[0] = string[9];
    d[1] = string[10];
    d[2] = '/0';
    Acell = atof (a);
    Bcell = atof (b);
    Ccell = atof (c);
    Dcell = atof (d);
  //  printf("A= %.0f, B=%.0f, C=%.0f, D=%.0f\n\n", Acell,Bcell,Ccell,Dcell);
  //  can scale cells up with scalar eg. 5, 10, 50
    kappacalc (Acell,Bcell,Ccell,Dcell);
    kappacalc (Acell*5,Bcell*5,Ccell*5,Dcell*5);
    kappacalc (Acell*10,Bcell*10,Ccell*10,Dcell*10);
    kappacalc (Acell*50,Bcell*50,Ccell*50,Dcell*50);
  }
}
```

Appendix XIII. Minimum P_e for Given P_o.

Inspection of the random two by two tables showed that for a given value of P_o there was a minimum value for P_e. Further inspection of the two by two tables showed that the minimum P_e occurred when the "a" cell value approximated the "d" cell and when either one of the non-agreement cells was large and the other was small. Therefore, assume that a = d, b > c and c = 0. Recall for any given two by two table:

N = a + b + c + d.

P_o = (a + d)/N.

and P_e is given by:

$$\frac{a+b}{N}*\frac{a+c}{N} + \frac{c+d}{N}*\frac{b+d}{N}$$

Expressing the equations in terms of a, b and N, we get:

N = 2a + b

P_o = 2a/N

$$Pe=\frac{a+b}{N}*\frac{a}{N} + \frac{a}{N}*\frac{b+a}{N}$$

Summing we get:

$$Pe=\frac{2a^2+2ab}{N^2}$$

Arranging, we get:

$$Pe = \frac{\frac{2a}{N}(a+b)}{N}$$

Replacing 2a/N with P_o gives:

$$Pe = \frac{Po(a+b)}{N}$$

Substituting for b gives:

$$Pe = \frac{Po(N-a)}{N}$$

Replacing a/N with $P_o/2$ we get:

$$Pe_{min} = Po(1 - \frac{Po}{2})$$

Appendix XIV. Maximum Kappa for Given P_o

Given that there is a minimum P_e for a given P_o, it follows that there is a maximum Kappa because:

$$\kappa = \frac{Po - Pe_{min}}{1.0 - Pe_{min}}$$

Replacing P_e min with $P_o(1 - P_o/2)$ and rearranging we get:

$$K_{max} = \frac{Po^2}{2 - 2Po + Po^2}$$

Appendix XV. Sample Size Table for Kappa Reliability Studies Using Estimates of Kappa and Test Positivity

Sample size estimation can be based on the parameters of estimated test reliability (κ) and test positivity (π). As mentioned on page 146, a Goodness-of-Fit statistic using these parameters was developed by Donner and Eliasziw[129] with the assumption the raters were unbiased (i.e. had the same rates of test positivity).

A method of obtaining sample sizes based on the parameters of Kappa and π would be to represent the standard error of Kappa as given by Cohen in terms of Kappa and π. The algebra quickly becomes unwieldy as is shown below. For P_e to be written in terms of Kappa and π we can use the fact, in the same formulation as used Donner and Eliasziw, that:

$$P(disagreement) = 2\pi(1-\pi)(1-\kappa)$$

Then P_o is simply the complement of probability of disagreement:

$$Po = 1-2\pi(1-\pi)(1-\kappa)$$

In order to get P_e in terms of Kappa and π, we must assume that the observers are unbiased, i.e. b = c in the two by two table.

Recalling that,

$$P_e = \frac{a+b}{N} * \frac{a+c}{N} + \frac{b+d}{N} * \frac{c+d}{N}$$

and since,

$$\pi = \frac{a+b}{N} = \frac{a+c}{N}$$
$$(1-\pi) = \frac{b+d}{N} = \frac{c+d}{N}$$

therefore:

$$P_e = \pi^2 + (1-\pi)^2$$

Using the equation of Cohen

$$se(\kappa) = \frac{\sqrt{Po - Po^2}}{(1-Pe)\sqrt{N}}$$

and substituting for P_o and P_e we get:

$$se(\kappa) = \frac{\sqrt{(1-2\pi(1-\pi)(1-\kappa)) - (1-2\pi(1-\pi)(1-\kappa))^2}}{(1-(\pi^2+(1-\pi)^2))\sqrt{N}}$$

This equation can be solved for N based on κ, π and δ.

$$N = \frac{1.96^2((1-2\pi(1-\pi)(1-\kappa))-(1-2\pi(1-\pi)(1-\kappa)))^2}{\delta^2(1-(\pi^2+(1-\pi)^2))^2}$$

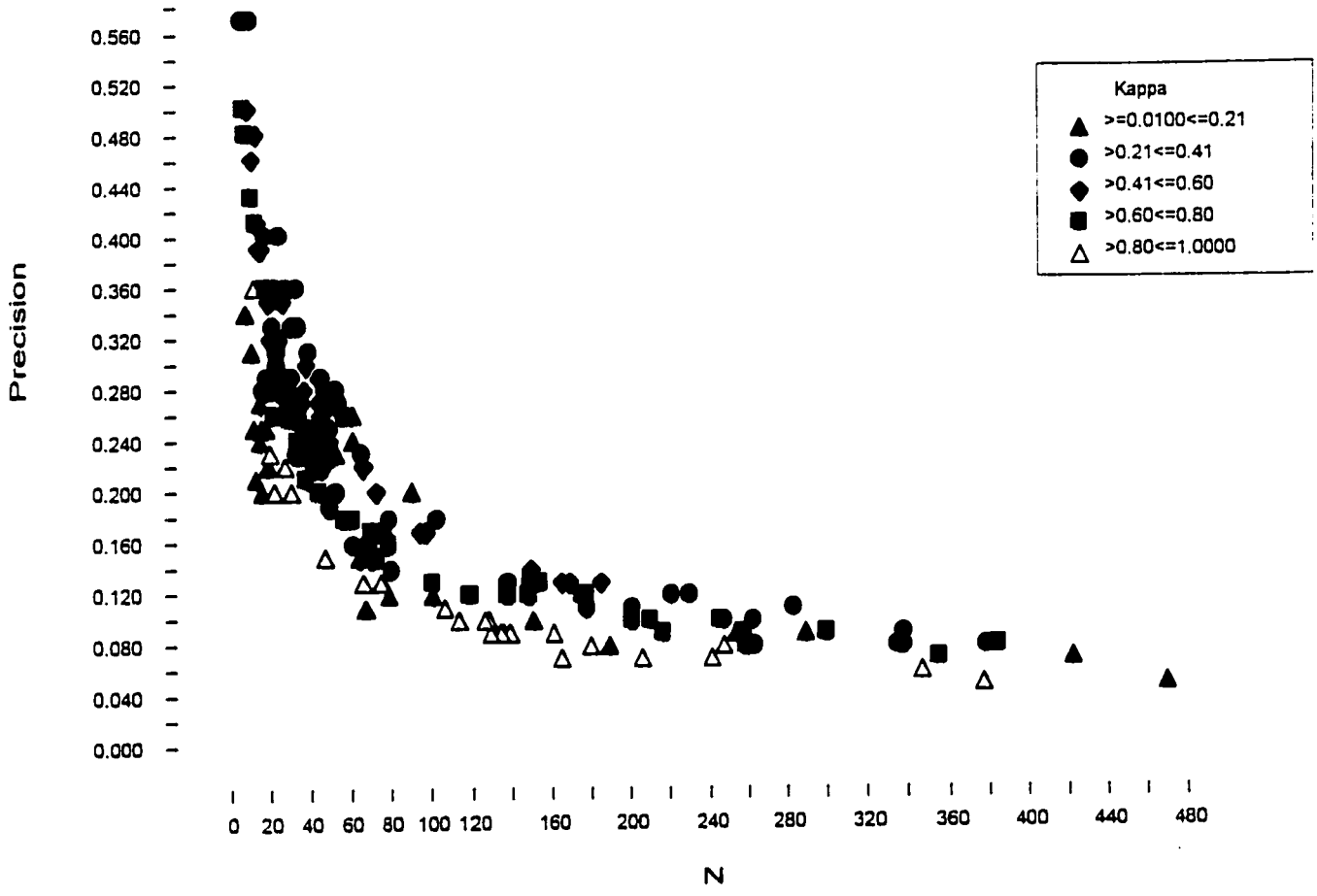
Using this equation and incremental values of Kappa and π one could create an equation-derived sample size table as was done in Appendix VII.

Using a computer simulation approach, sample size can be estimated without the assumption about rater bias. From the random two by two tables, a listing was created that included test positivity (π) defined as the average positive test prevalence among the two observers. From the two by two table:

$$\pi = \frac{(a+b)/N+(a+c)/N}{2}$$

The sample size (N) was plotted against the precision around the Kappa estimate (δ). As long as the Kappa was > .10 and < .90, the relationship was parabolic for a given level of test postivity(π). See Figure 11 for an example of a scatterplot of N versus δ for a π of .30.

Figure 11. Scatterplot of sample size (N) and precision (δ) grouped by Kappa for a test positivity (π) of approximately 30%.



Scatterplots were created with test positivity (π) set at approximately 50%, 40%, 30%, 20% and 10% and test reliability (κ) set at approximately .10, .30, .50, .70 and .90. From these scatterplots, the sample sizes corresponding to the various precisions were interpolated and tabulated in a sample size table. (See Table 24.)

As with the sample size table described in Appendix VII, there is the need to assume a minimum sample size for the large sample assumption for the Normal approximation to hold. As well, the upper confidence interval will be degenerate for higher Kappa values with wide confidence intervals. These cases are again identified on the table with the letters *m* for minimum sample size. An asterisk indicates when the upper confidence interval is degenerate i.e. extends above 1 (which is impossible).

Table 24. Sample Size Table for Two by Two Reliability Studies Based on the Expected Test Positivity (π) and Kappa (κ).

π	Kappa	N for κ±.05	κ±.10	κ±.15	κ±.20	κ±.25	κ±.30
0.1	0.1	>500	200	150	80	m†	m
0.1	0.3	>500	>500	280	150	130	100
0.1	0.5	>500	>500	270	250	200	80
0.1	0.7	>500	>500	220	150	130	90
0.1	0.9	>500	200*	190*	80*	m*	m*
0.2	0.1	>500	300	80	m	m	m
0.2	0.3	>500	360	140	100	m	m
0.2	0.5	>500	360	180	120	80	m
0.2	0.7	>500	360	140	100	m	m
0.2	0.9	>500	300*	m*	m*	m*	m*
0.3	0.1	480	100	m	m	m	m
0.3	0.3	>500	200	120	80	m	m
0.3	0.5	>500	220	140	100	m	m
0.3	0.7	>500	200	120	80	m	m
0.3	0.9	380	100	m*	m*	m*	m*
0.4	0.1	360	120	m	m	m	m
0.4	0.3	560	180	80	m	m	m
0.4	0.5	560	200	120	m	m	m
0.4	0.7	560	180	80	m	m	m
0.4	0.9	320	120	m*	m*	m*	m*
0.5	0.1	440	100	m	m	m	m
0.5	0.3	640	180	80	m	m	m
0.5	0.5	640	200	120	m	m	m
0.5	0.7	640	180	80	m	m	m
0.5	0.9	440	100	m*	m*	m*	m*

* Upper confidence limit is degenerate.
† m = minimum sample size.

Bibliography

1. Collins JG. Prevalence of selected chronic conditions, United States, 1983-85. Hyattsville, Maryland: Public Health Service, 1988; DHHS publication no. (PHS) 88-1250, (Vital and health statistics, no. 155)

2. Dobkin B. The economic impact of stroke. Neurol 1995;45(Suppl 1):S6-9.

3. Kidd D, Lawson J, MacMahon J. Aspiration in acute stroke: a clinical study with videofluoroscopy. Q J Med 1993;86:825-9.

4. Logemann JA. Treatment for aspiration related to dysphagia: an overview. Dysphagia 1986;1:34-8.

5. Spurr SA, Valencia AM. Aspiration pneumonia in stroke: a prospective study. Neurol 1995;45(Suppl 4):A223.

6. Vitanen M, Winblad B, Asplund K. Autopsy-verified causes of death in hemiplegia. Acta med Scand 1987;222:401-5.

7. Gresham GE, Duncan PW, Stason WB et al. Post-stroke rehabilitation: clinical practice guideline. Gaithersburg, Maryland: Aspen, 1996:59 (AHCPR publication no: 95-0662)

8. Smithard DG, O'Neill PA, Park C, Morris J, Wyatt R, England R, Martin DF. Complications and outcome after acute stroke. Stroke 1996;27:1200-4.

9. DePippo KL, Holas MA, Reding MJ. The Burke screening test for dysphagia: validation of its use in patients with stroke. Arch Phys Med Rehabil 1994;75:1284.

10. DePippo KL, Holas MA, Reding MJ. Validation of the 3-oz water swallow test for aspiration following stroke. Arch Neurol 1992;49:1259-61.

11. Linden P, Kuhlemeier KV, Patterson C. The probability of correctly predicting subglottic penetration from clinical observations. Dysphagia 1993;8:170-9.

12. Palmer JB. Kuhlemeier KV Tippett DC Lynch C. A protocol for the videofluoroscopic swallowing study. Dysphagia 1993; 8:209-14.

13. Ullman T, Reding M. Gastrointestinal dysfunction in stroke. Semin Neurol 1996;16:269-75.

14. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.

15. Sackett DL, Haynes B, Guyatt G, Tugwell P. Clinical epidemiology: a basic science for clinical medicine 2 ed. Boston: Little & Brown, 1991.

16. Laupacis A, Wells G, Richardson S, Tugwell P. User's guides to the medical literature: how to use an article on prognosis. JAMA 1994:272:234-7.

17. Logemann JA. Screening, diagnosis, and management of neurogenic dysphagia. Semin Neurol 1996; 16:319-27.

18. Gordon C, Hewer RL, Wade DT. Dysphagia in acute stroke. Br Med J 1987; 295:411-4.

19. Barer DH. Lower cranial nerve motor function in unilateral vascular lesions of the cerebral hemisphere. Br Med J 1984; 289:1622.

20. Rieve JE, Kirby NA, Siebens AA. Clinical correlates of Dysphagia in stroke patients, [abstract] Arch Phys Med Rehabil 1986; 70:55-6.

21. Horner J, Massey MD, Riski JE, Lathrop MA, Chase MS. Aspiration following stroke: clinical correlates and outcome. Neurol 1988;38:1359-62.

22. Robbins J, Levine RL. Swallowing after unilateral stroke of the cerebral cortex: preliminary experience. Dysphagia 1988;3:11-7.

23. Gresham, SL. Clinical assessment and management of swallowing difficulties after stroke. Med J Aust. 1990;153:399.

24. Groher MD, Bukatman R. The prevalence of swallowing disorders in two teaching hospitals. Dysphagia 1986;1:3-6.

25. DePippo KL. Holas MA, Reding MJ, Mandel FS, Lesser ML. Dysphagia therapy following stroke: a controlled trial. Neurol 1994;44:1655-60.

26. Detsky AS, Sackett DL. When was a 'negative' clinical trial big enough? how many patient you needed depends on what you found. Arch Intern Med 1985;145:709-12.

27. Finucane TE, Bynum JP. Use of tube feeding to prevent aspiration pneumonia. Lancet 1996;348:1421-4.

28. Norton B, Homer-Ward M, Donnelly MT, Long RG, Holmes GK. A randomized prospective comparison of percutaneous endoscopic gastrostomy and nasogastric tube feeding after acute dysphagic stroke. Br Med J 1996;6:312:13-6,

29. Buchholz D. Neurologic evaluation of dysphagia. Dysphagia 1987; 1:187-192.

30. Davies AE, Kidd D. Stone P MacMahon J. Pharyngeal sensation and gag reflex in healthy subjects. Lancet 1995;345:487-8.

31. Horner J, Massey EW, Brazer SR Aspiration in bilateral stroke patients. Neurol 1990;40:1686-88.

32. Logemann J. Evaluation and treatment of swallowing disorders. San Diego: College Hill Press, 1983:110

33. DePippo KL, Holas MA, Reding MJ. The Burke screening test for dysphagia: validation of its use in patients with stroke [Abstract]. Stroke 1993;24:173.

34. Splaingard M, Hutchins B, Sulton LD, Chaudhuri G. Aspiration in rehabilitation patients: videofluoroscopy versus bedside clinical assessment. Arch Phys Med Rehabil 1988; 69:637-40.

35. Finestone HM Greene-Finstone LS, Wilson ES, Teasell RW. Malnutrition in stroke patients on the rehabilitation service and at follow-up: prevalence and predictors. Arch Phys Med Rehabil 76:76:310-16.

36. Hantson L, DeKeyser J. Neurological scales in the assessment of cerebral infarction. Cerebrovasc Dis 1994:4(suppl2):7-14.

37. D'Olhaberriague L, Litvan I, Mitsias P, Mansbach HH. A reappraisal of reliability and validity studies of stroke. Stroke 1996;27:2331-6.

38. Bowling A. Measuring disease: a review of disease-specific quality of life measurement scales. Philadelphia: Open University Press, 1995:175-89.

39. Guyatt GH, Jaeschke R. Measurement in clinical trials; choosing the appropriate approach. In: Spilker B, ed. Quality of life assessments in clinical trials. New York: Raven Press, 1990:37-46.

40. McDowell I, Newell C Measuring Health A Guide to Rating Scales and Questionnaires. New York: Oxford University Press, 1996:56-63.

41. Granger CV, Albrecht GL, Hamilton BB. Outcome of comprehensive medical rehabilitation: measurement by PULSES Profile and the Barthel Index. Arch Phys Med Rehabil 1979;60:145-54.

42. Granger CV, Sherwood CC, Greer DS. Functional status measures in a comprehensive stroke care program. Arch Phys Med Rehabil 1977;58:555-61.

43. Granger CV, Dewis LS, Peters NC, Sherwood CC, Barrett JE. Stroke rehabilitation:analysis of repeated Barthel index measures. Arch Phys Med Rehabil 1979;60:14-7.

44. Linn MW, Linn BS. The rapid disability rating scale-2. J Am Geriatr Soc 1982;30:378-82.

45. Wolfe CDA, Taub NA, Woodrow EJ, Burney PGJ, Asessment scales of disability and handicap for stroke patients. Stroke 1991;22:1242-4.

46. Ekberg O, Nylander G, Fork F, Sjoberg S, Birch-Iensen M, Hillary B. Interobserver variability in cineradiographic assessment of pharyngeal function during swallow. Dysphagia 1988;3:46-8.

47. Kidd D, Lawson J, Nesbitt R, MacMahon J. The natural history and clinical consequences of aspiration in acute stroke. Q J Med 1995;88:409-13.

48. Holas MA, DePippo KL, Reding MJ. Aspiration and relative risk of medical complications following stroke. Arch Neurol 1994;51:1051-3.

49. Johnson ER, McKenzie SW, Sievers A. Aspiration pneumonia in stroke. Arch Phys Med Rehabil 1993;74:973-6.

50. Johnson ER, McKenzie SW, Sievers A. Aspiration pneumonia in stroke [erratum]. Arch Phys Med Rehabil 1994;75:665.

51. Teasell RW, McRae M, Marchuk Y, Finestone HM. Pneumonia associated with aspiration following stroke. Arch Phys Med Rehabil 1996;77:707-9.

52. Schmidt J, Holas M, Halvorson K, Reding M. Videofluoroscopic evidence of aspiration predicts pneumonia and death but not dehydration following stroke. Dysphagia 1994; 9:7-11.

53. Langmore SE, Schatz K, Olson N. Endoscopic and videofluoroscopic evaluation of swallowing and aspiration. Ann Oto Rhinol Laryngol 1991;199:678-81.

54. Rothman S. Test may cut post stroke aspiration pneumonia risk, otolaryngology update. Med Post May 1996.

55. Feinstein AR. Clinimetrics. New Haven: Yale University Press, 1987:167.

56. Kramer MS, Feinstein AR Clinical biostatistics: the biostatistics of concordance. Clin Pharmacol Ther 1981;29:111-23.

57. Fletcher RH, Fletcher SW, Wagner EH. Clinical epidemiology: the essentials. 3rd ed. Baltimore:Williams & Wilkens 1996:44.

58. Norman GR, Streiner DL. Biostatistics: the bare essentials. St Louis: Mosby, 1994.

59. Groher ME. Dyphagia: diagnosis and management. Boston: Butterworth and Heineman, 1992:158.

60. Laupacis A, Sekar M, Stiell IG. Clinical prediction rules: a review and suggested modifications of the methodologic standards. JAMA 1997;277:488-94.

61. Stern LZ, Bernick C. Motor system and gait. In: Walker HK, Hall WD, Hurst JW eds. Clinical Methods, 3rd ed. Boston: Butterworths, 1990:353.

62. Lazzara GL, Lazzarus C, Logemann J. Impact of thermal stimulation on the triggering of the swallowing reflex. Dysphagia 1986;1:73-7.

63. Grant JP, Custer PB, Thurlow J. Current techniques of nutritional assessment. Surg Clin North Am 1981;61:437.

64. Korner-Bitensky N, Wood-Dauphinee S, Siemiatycki J, Shapiro S, Becker R. Health-related information postdischarge: telephone versus face-to-face interviewing. Arch Phys Med Rehabil 1994:75:1287-96.

65. Westberg J, Jason H. Teaching creatively with video. New York: Springer, 1994.

66. LeWitt PA, Miller LP, Levine RA, Lovenberg W, Newman RP, Papavasiliou A, Rayes A, Eldridge R, Burns RS. Tetrahydrobiopterin in dystonia: Identification of abnormal metabolism and therapeutic trials. Neurol 1986;36:760-5.

67. English M, New L, Peshu N, Marsh K. Video-assessment of simple respiratory signs. Br Med J 1996;313:1527-8.

68. English M, Murphy S, Mwangi I, Crawley J, Peshu N, Marsh K. Interobserver variation in respiratory signs of severe malaria. Arch Dis Child 72:334-6.

69. Grove WM, Andreasen NC, McDonald-Scott P, Keller MB, Shapiro RW. Reliability studies of psychiatric diagnosis: theory and practice. Arch Gen Psychiatry 1981;38:408-13.

70. Koran LM. The Reliability of Clinical Methods, Data and Judgements (Second of two Parts) N Enol J Med 1975;293:695-702.

71. Fleiss JL. Measuring agreement between two judges on the presence or absence of a trait. Biometrics 1975;31:651-9.

72. Alves WA. Macciocchi SN. Ethical considerations in clinical neurosciences: current concepts in neuroclinical trials. Stroke 1996;27:1903-8.

73. Kahn HA, Sempos CT. Statistical methods in epidemiology. New York: Oxford University Press, 1989:66.

74. Friedman LM, Furberg CD, DeMets DL. Fundamentals of clinical trials. 2nd ed. Littleton, Mass: PSG Publishing Company, 1985:90.

75. Concato J, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. Ann Intern Med 1993;118:201-10.

76. Maher J, Reilly M, Daly L, Hutchinson M. Plantar power: reproducibility of the plantar response. Br Med J 1992;304:482.

77. Byrt T, Bishop J, Carlin JB. Bias, prevalence and Kappa. J Clin Epidemiol 1993;46:423-9.

78. Spitznagel EL, Helzer JE. A proposed solution to the base rate problem in the Kappa statistic. Arch Gen Psychiatry 1985;42:725-728.

79. Etchells E, Bell C, Robb K. Does this patient have an abnormal systolic murmur? JAMA 1997; 277:564-71.

80. Sackett DL. Bias in analytic research. J Chronic Dis 1979;32:51-63.

81. Lachs MS, Nachamkin I, Edelstein PH, et al. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. Ann Intern Med 1992;117:135-40.

82. O'Connor PW, Tansey CM, Detsky AS, Mushlin AI, Kucharczyk W. The effect of spectrum bias on the utility of magnetic resonance imaging and evoked potentials inthe diagnosis of suspected multiple sclerosis. Neurol 1996;47:140-4.

83. Feinstein AR. Clinical Epidemiology: the architecture of clinical research. W. B. Saunders Company, 1985:640.

84. Shinar D, Gross CR, Mohr JP et al. Interobserver variability in assessment of neurologic history and examination in the stroke data bank. Arch Neurol 1985;557-65.

85. Spiteri MA, Cook DG, Clarke SW. Reliability of eliciting physical signs in examination of the chest. Lancet 1988;April: 873-5.

86. Sackett, DL. A primer on the precision and accuracy of the clinical examination. JAMA 1992;267:2638-44.

87. Simel DL, Rennie D. The clinical examination: an agenda to make it more rational. JAMA 1997;277:572-4.

88. Sox HC, Blatt MA, Higgens MC, Marton KI. Medical Decision Making Boston: Butterworths, 1988:116-125.

89. Horner J Brazer SR, Massey EW Aspiration in bilateral stroke patients: a validation study Neurol 1993;43:430-3.

90. Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules: applications and methodologic standards. N Enol J Med 1985;313:793-9.

91. Heckerling PS, Tape TG, Wigton RS et al. Clinical prediction rule for pulmonary infiltrates. Ann Int Med 1990;113:664-70.

92. Stiell IG, Greenberg GH, McKnight RD, Nair RD, McDowell I, Worthington JR. A study to develop clinical decision rules for the use of radiography in acute ankle injuries. Ann Emerg Med 1992;384-90.

93. Koretz R, Cook DJ. Commentary on "Enteral feeding by endoscopic gastrotomy reduced 6-week mortality after dysphagic stroke." ACP J Club.1996 July-Aug;125:5. Comment on: Norton B, Homer-Ward M, Donnelly MT, Long RG, Holmes GK. A randomized prospective comparison of percutaneous endoscopic gastrotomy and nasogastric tube feeding after acute dysphagic stroke. Br Med J 1996;6:312:13-6.

94. Palmer JB. Kuhlemeier KV Tippett DC Lynch C. A protocol for the videofluoroscopic swallowing study. Dysphagia 1993; 8:209-14.

95. Stern LZ, Bernick C. Motor system and gait in Clinical Methods, third edition Walker HK, Hall WD, Hurst JW (eds.). Boston: Butterworths, 1990:353.

96. Levison ME. Pneumonia, including necrotizing pulmonary infections (lung abscess). In: Isselbacher KJ. Braunwald E, Wilson JD, Martin JB, Fauci AS, Kasper DL, eds. Harrison's principles of internal medicine. New York: McGraw-Hill 1994:1184-1191.

97. Grant JP, Custer PB, Thurlow J. Current techniques of nutritional assessment. Surg Clin North Am 1981;61:437.

98. Stiell IG, Greenberg GH, McKnight RD, Nair RC, McDowell I, Worthington JR. A study to develop clinical decision rules for the use of radiography in acute ankle injuries. Ann Emerg Med 1992;21:384-90.

99. Thompson, JM, McFarland GK, Hirsch JE, Tucker SM, Bowers AC. Mosby's manual of clinical nursing. 2nd ed. St. Louis: C.V. Mosby Company 1989:1636.

100. Bronstein KS, Popovich JM, Stewart-amidei C. Promoting stroke recovery: a research-based approach for nurses. St. Louis: Mosby. C.V. 1991:129.

101. Tucker SM, Canobibbio MM, Paquette EV, Wells MF. Patient care standards: nursing process, diagnosis, and outcome. 5th ed. St. Louis: Mosby Year Book 1992:407.

102. Lyden P, Brott T, Tilley B et al. Improved reliability of the NIH stroke scale using video training. Stroke 1994;25:2220-6.

103. Diamond GA. Limited assurances. Am J Cardiol 1989;63:99-100.

104. Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence sample size estimation for diagnostic test studies. J Clin Epidemiol 1991;44:763-770.

105. Whittlemore AS. Sample size for logistic regression with small response probability. J Am Stat Soc 1981;76:27-32.

106. McConnochie KM, Roghmann KJ, Pasternack J. Developing prediction rules and evaluating observation patterns using categorical clinical markers: two complementary procedures. Med Dec Making 1993; 13:30-42.

107. KnowledgeSEEKER User's Guide, Decision Making Software, version 2.0. Ottawa: FirstMark Technologies Ltd. 1990.

108. Biggs D, deVille B, Suen E. A method of choosing multiway partitions for classification and decision trees. J Appl Stat 1991;18:49-62.

109. Feinstein AR. Multivariable analysis: an introduction. New Haven: Yale University Press, 1996:554.

110. Fukonada K, Kessell DI. Estimation of classification error. IEEE Trans Comput 1971;20:1521-7.

111. Monsour MJ, Evans AT, Kupper LL. Confidence intervals for post-test probability. Stat Med 1991;10:443-56.

112. Hull R, Raskob G, Hirsh J. The diagnosis of clinically suspected pulmonary embolism: practical approaches. Chest 1986;89:417(S)-25(S).

113. Dennis M, O'Rourke S, Slattery J, Staniforth T, Warlow C. Evaluation of a stroke family care worker: results of a randomised controlled trial. Br Med J 1997;314:1071-7.

114. McLean S. Commentary: no consent means not treating the patient with respect. Br Med J 1997;314:1077.

115. Weiss NS. Clinical epidemiology: the study of the outcome of illness. 2nd ed. New York: Oxford University Press, 1996:27.

116. Dronfield MW, McIllmurray MB, Ferguson R, et al. A prospective, randomized study of endoscopy and radiology in acute upper-gastrointestinal-tract bleeding. Lancet 1977;1:1167-9.

117. Mandel JS, bond JH, Church TR, et al. Reducing mortality from colorectal carcinoma by screening fro fecal occult blood. N Enol J Med 1993;328:1365-71.

118. SHEP Cooperative Research Group. Prevention of stroke by antihypertensive drug treatment in older persons with isolated systolic hypertension: final results of the systolic hypertension in the elderly program. JAMA 1991;265:3255.

119. Guggenmoos-Holzmann. How reliable are chance-corrected measures of agreement? Stat Med 1993;12:2191-2205.

120. Thompson WD. Walter SD. A reappraisal of the Kappa coefficient. J Clin Epidemiol 1988; 10:949-58.

121. Kraemer HC, Bloch DA. Kappa coefficients in epidemiology: an appraisal of a reappraisal. J Clin Epidemiol 1988;41:959-68.

122. Fleiss JL, Cohen J, Everitt BS. Large sample standard errors of Kappa and weighted Kappa. Psych Bull 1969;72:323-7.

123. Fleiss JL. Statistical methods for rates and proportions. New York: John Wiley & sons, 1981.

124. Dunn G, Everitt B. Clinical biostatistics: an introduction to evidence-based medicine. London: Edward Arnold, 1995:39-44.

125. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37-46.

126. Wickens TD. Multiway contingency table analysis for the social sciences. Los Angeles: Lawrence Erlbaum. 1989:240.

127. Fleiss JL, Cicchetti DV. Inference about weighted Kappa in the non-null case. Appl Psych Measure 1978;2:113-7.

128. Cicchetti DV. Testing the normal approximation and minimal sample size requirements of weighted Kappa when the number of categories is large. Appl Psych Meas 1981;5:101-12

129. Donner A, Eliasziw M. A goodness-of-fit approach to inference procedures for the Kappa statistic: confidence interval construction, significance-testing and sample size estimation. Stat Med 1992;11:1511-9.

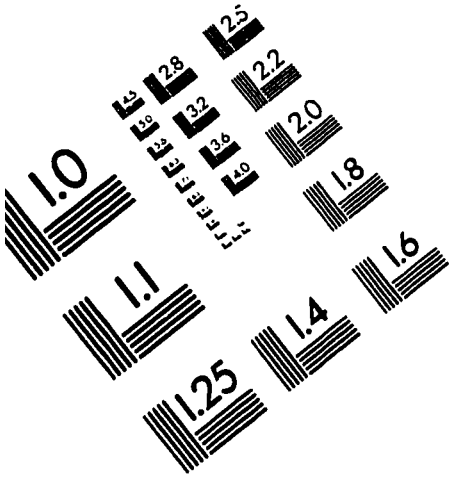
130. Bristol DR. Sample sizes for constructing confidence intervals and testing hypotheses. 1989 Stat Med 1989;8:803-11.

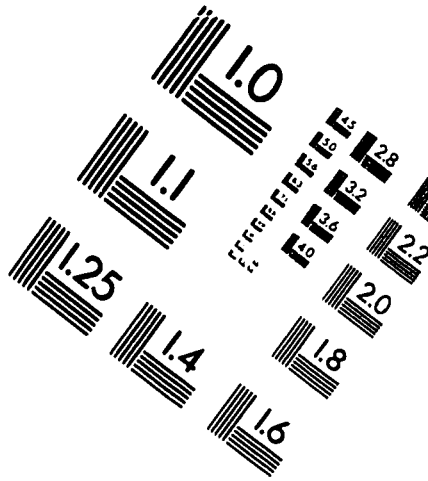
131. Manly BF. Randomization, bootstrap and Monte Carlo methods in biology. 2nd ed. London: Chapman & Hall,1997:14.

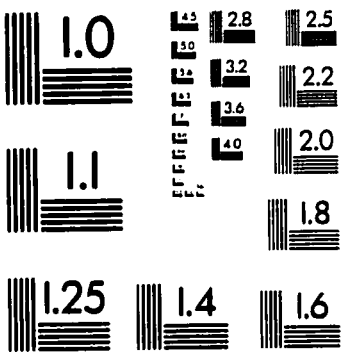
132. Feinstein AR, Cicchetti DV High agreement but low Kappa: 1. The problems of the two paradoxes. J Clin Epidemiol 1990;43:543-9.

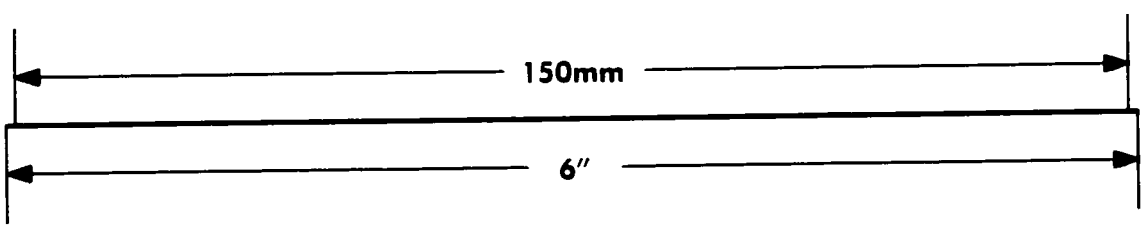
133. Bloch DA, Kraemer HC. 2x2 Kappa coefficients: measures of agreement or association. Biometrics 45:269-87.

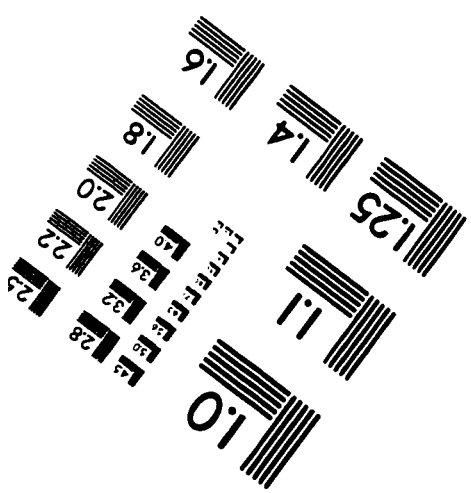
IMAGE EVALUATION
TEST TARGET (QA-3)

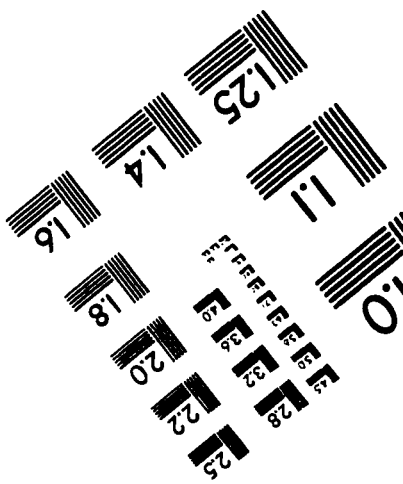












APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989