Reliability of diagnostic procedures in medical imaging: Narrow Band Imaging Endoscopy in gastric disease

Raul Marques Pereira

2010

Mestrado em Evidência e Decisão em Saúde

Faculdade de Medicina

Universidade do Porto

Orientador: Professor Doutor Mário Dinis-Ribeiro, Professor Auxiliar com Agregação da Faculdade de Medicina da Universidade do Porto

Preâmbulo

Ao longo do seu percurso profissional como Médico Interno em Medicina Geral e Familiar (MGF), o autor constatou que a literatura disponível nesta área, embora vasta, é parca em estudos de avaliação de reprodutibilidade na Medicina Familiar e na abordagem desta área como sendo fundamental para a avaliação de estudos de diagnóstico.

Um dos desafios actuais para a Investigação Clínica em MGF decorre da necessidade de avaliar a qualidade de vários dos testes de diagnóstico utilizados na prática clínica. São vários os exemplos de testes que, embora utilizados na prática clínica diária, carecem de validação e avaliação da reprodutibilidade inter-observador ou só a obtiveram vários anos depois de estarem implementados (Smilkstein, et al., 1982) (Cohrssen, et al., 2005) (Coupland, et al., 2007).

Vários destes testes decorrem de questionários que podem ser sensíveis a variações culturais, pelo que a avaliação da sua reprodutibilidade entre observadores em diferentes situações clínicas se constitui como fundamental para se ponderam considerar os resultados obtidos como fiáveis (Saameño, et al., 1996) (Lama, et al., 2010).

Com a aplicação sistemática destes e de outros protocolos de diagnóstico internacionais, torna-se fundamental desenvolver conhecimentos na área da reprodutibilidade de forma a potencialmente contribuir para a evolução da MGF nesta área.

Como aluno do Mestrado em Evidência e Decisão em Saúde (MEDS), o autor teve a oportunidade de integrar uma equipa que estava a estudar a reprodutibilidade em procedimentos diagnósticos baseados em imagem médica, nomeadamente a endoscopia de *Narrow Band Imaging* (NBI) em patologia gástrica. Desta forma, o autor pôde desenvolver aptidões nesta área com vista à posterior aplicação na MGF.

Nesta tese vão ser focados alguns pontos que se consideram relevantes para uma abordagem global ao tema. Um dos capítulos (capítulo 3) pretende explanar uma abordagem sistemática ao planeamento e condução de um estudo de reprodutibilidade. Este capítulo foi elaborado em conjunto com a Mestre Joana Oliveira Ribeiro, também aluna do MEDS, como parte de um esforço de sistematização de procedimentos nesta área, uma vez que se constatou que esta é uma lacuna da literatura actual nesta área.

O capítulo 4 refere-se a uma revisão sistemática da endoscopia de NBI em patologia gástrica. Este capítulo foi enviado como manuscrito para publicação.

No capítulo 5 discute-se a derivação de uma classificação para a endoscopia de NBI que foi desenvolvida a partir dos resultados obtidos na revisão sistemática descrita no capítulo 4. Este trabalho foi levado a cabo através do planeamento e implementação de um estudo que avaliou a reprodutibilidade desta nova classificação por diferentes observadores (endoscopistas).

Nesta Tese apresentam-se resultados parciais deste trabalho, uma vez que se discute apenas a derivação desta nova classificação de NBI e não a sua validação para a prática clínica. Este estudo será completado posteriormente através dessa validação e da inclusão de um leque mais alargado de endoscopistas. O capítulo 5 esteve na origem de dois trabalhos aceites, como comunicação oral, nos Congressos Nacional e Europeu de Gastrenterologia.

Acknowledgments

To my supervisor, Professor Mário Dinis-Ribeiro, PhD for his support, availability and for never letting me feel discouraged.

To Matilde Soares, MSc and Joana Ribeiro, MSc, my colleagues in MEDS, for their support and company in this journey.

To Juliana Rocha, MSc for her daily support, encouragement and patience.

To my parents for their continuous motivation to do better and go beyond the average.

Table of Contents

Abreviations List

AFI – Autofluorescence Endoscopy

AUC – Area under the ROC curve

BE – Barrett's Esophagus

CE – Chromoendoscopy

CI – Confidence Interval

CT – Computed Tomography

DOR – Diagnostic Odds Ratio

GC - Gastric Cancer

H. Pylori – Helicobacter Pylori

HR – High-Resolution

ICC – Intraclass Correlation Coefficient

IM - Intestinal Metaplasia

LBC – Light Blue Crest

ME – Magnifying endoscopy

MEDS – Mestrado em Evidência e Decisão em Saúde

MGF – Medicina Geral e Familiar

MVP – Microvascular Pattern

NBI – Narrow Band Imaging

QUADAS – Quality Assessment of Diagnostic Accuracy Studies

ROC – Receiver Operating Characteristic Curve

TME – Trimodal Imaging Endoscopy

WLI – White-Light Imaging Endoscopy

WOS – White Opaque Substance

Table Index

Figure Index

Summary

The accuracy of measurements is very important in clinical practice since most decisions are based in registered values and diagnostic tests. Reliability is the extent to which a measure provides the same results repeatedly and under the same circumstance. We should assess both reliability and validity of diagnostic tests before their integration in clinical daily practice or their use in clinical research.

Gastric cancer remains a major health issue as it is the fourth most common cancer and the second cause of oncological deaths worldwide. Narrow-band imaging (NBI), a new endoscopic modality, may improve the detection of early neoplastic lesions, improving the effectiveness of endoscopic surveillance and screening.

Using medical imaging such as digestive endoscopy as an example, we present a stepwise approach on how to plan and conduct a reliability study to assess inter-observer reliability.

When using ordinal variables, we can consider the classification proposed by Uedo (Uedo, et al., 2006) for the appearance of a Light Blue Crest (LBC) on the epithelial surface of gastric mucosa, as an example. This classification considered 4 degrees of LBC: non-LBC, LBC +, LBC ++, LBC +++. In this case, LBC is an ordinal variable, and we should use the Intraclass Correlation Coefficient (ICC) to assess reliability, allowing in this specific case to weight the differences between ratings.

When considering dichotomous data, for example the *H. pylori* infection status (positive or negative) referred in the paper by Tahara (Tahara, et al., 2009), we may report descriptive agreement measures as the observed agreement and specific agreement on each grade. In this case Kappa is the correct method for assessing reliability.

When designing a reliability study we should keep in mind that if the sample size is too small the study may produce an inaccurate estimate of the reliability coefficient, whereas if it is too large it may constitute a misuse of resources.

No definite classification has been consistently proposed to use NBI endoscopy for the diagnosis of gastric precancerous and cancer lesions, and consequently we conducted a systematic review of the existing studies reporting NBI use in stomach.

We summarized the reliability and accuracy of existing descriptions, and conducted the first meta-analysis for NBI endoscopy for gastric pre-cancerous and cancer diagnosis.

Eleven studies including 777 patients were evaluated in this study. Most studies addressed cancer as the main outcome while normal tissue, dysplasia and positive *H. pylori* were described only in two studies. None of the studies described all the precancerous lesions in a single classification. Quality score varied from 8 to 12 items. Only one study assessed inter-observer reliability. On a patient level analysis, NBI pooled sensitivity, specificity and DOR for gastric metaplasia were 0.82 (95% CI 0.71-0.90), 0.93 (0.89-0.96) and 53.5 (95% CI 21.4-133.9) respectively; for dysplasia, 0.90 (CI 95% 0.70-0.90), 0.95 (95% CI 0.82-0.99) and 116.9 (CI 95% 12.6-1089.2) respectively; and for differentiated cancer, 0.77 (95% CI 0.72-0.82), 0.87 (CI 95% 0.81-0.91) and 39.9 (CI 95% 12.1-130.8) respectively.

We concluded that lesions definitions vary between studies and reliability is a neglected topic, and that NBI may be an extremely useful diagnostic tool for the different gastric lesions with a valuable impact in medical decision making.

A derivation study, concerning the application of a classification derived from the systematic review previously mentioned was conducted to assess the inter-observer reliability of this classification.

The identification of different vascular and mucosal patterns was associated with high reproducibility (k=0.82 and k=0.91, respectively). The observers proposed a histological diagnosis with high agreement (k=0.81). Sensitivity, specificity and validity for normal pattern were 84 (CI 95% 81-88), 81 (CI 95% 77-84) and 83 (CI 95% 79-86) respectively; for metaplasia, 83 (CI 95% 79-87), 89 (CI 95% 86-92) and 87 (CI 95% 84-90) respectively; for *H. Pylori* gastritis, 71 (CI 95% 67-76), 71 (CI 95% 66-76) and 71 (CI 95% 66-76) respectively; for dysplasia, 92 (CI 95% 89-95), 99 (CI 95% 98-100) and 97 (CI 95% 96-99) respectively. Variable vascular density was the best parameter for identification of HP gastritis, however with a low inter-observer agreement (k=0.4).

We concluded that new mucosal and vascular patterns derived from histology results were highly valid for metaplasia and dysplasia. This technique may show efficacy in the detection of gastric metaplasia and dysplasia, and High Resolution NBI endoscopy may be an important tool for early diagnosis and for therapeutic procedures.

A validation study that should complete the derivation study described above will help clarify the role of NBI in gastric disease, especially in the diagnosis of gastric precancerous lesions and cancer.

New studies to assess reliability and validate various international questionnaires and other diagnostic tools that are applied in our national clinical setting are important for the development of Family Medicine. It is the author's objective to collaborate in this effort.

Sumário

A precisão das medições é muito importante na prática clínica, uma vez que a maioria das decisões é baseada em valores registados e em testes de diagnóstico. A Reprodutibilidade reporta-se à capacidade de uma medição fornecer os mesmos resultados repetidamente e sob a mesma circunstância. A reprodutibilidade e a validade dos testes de diagnóstico devem ser avaliadas antes da sua integração na prática clínica diária ou da sua utilização em investigação clínica.

O cancro gástrico continua a ser um problema de saúde de grande importância, uma vez que é o quarto cancro mais comum e a segunda causa de morte oncológica em todo o mundo. A endoscopia de Narrow-Band Imaging (NBI) é uma nova modalidade endoscópica que pode melhorar a detecção precoce de lesões neoplásicas, melhorando a eficácia da vigilância e do rastreio endoscópico.

Usando um procedimento de imagem médica, a endoscopia digestiva alta, como exemplo, apresenta-se neste trabalho uma abordagem sistematizada para o planeamento e execução de um estudo de reprodutibilidade para a avaliação da reprodutibilidade inter-observador.

No que diz respeito à utilização de variáveis ordinais, pode-se considerar como exemplo a classificação proposta por Uedo (Uedo, et al., 2006) para o aparecimento de *Light Blue Crest* (LBC) na superfície epitelial da mucosa gástrica, como um exemplo. Esta classificação considera 4 graus de LBC: não-LBC, LBC +, LBC ++, LBC +++. Neste caso, LBC é uma variável ordinal, e devemos usar o ICC para avaliar a reprodutibilidade, admitindo neste caso concreto, o peso das diferenças entre as classificações.

Quando se considera uma variável dicotómica, como por exemplo o estado da infecção por *H. Pylori* (positivo ou negativo) referido no artigo de Tahara (Tahara et al. 2009), podemos avaliar a concordância observada e a concordância específica para cada grau. Neste caso, *Kappa* é o método correcto para avaliar a reprodutibilidade.

Quando se desenha um estudo de reprodutibilidade, deve ser tido em conta que, se o tamanho da amostra for muito pequeno, o estudo pode produzir uma estimativa imprecisa do coeficiente de reprodutibilidade. Se o tamanho da amostra for demasiado grande, pode ocorrer um uso desproporcionado de recursos.

Não há nenhuma classificação definitiva, consistentemente proposta, para o uso de endoscopia de NBI no diagnóstico de lesões gástricas pré-neoplásicas e neoplásicas. Nesse sentido, foi conduzida uma revisão sistemática dos estudos que se referem à utilização de NBI em patologia gástrica.

A reprodutibilidade e a precisão das descrições existentes foram sistematizadas, e conduziu-se a primeira meta-análise do uso de endoscopia de NBI para o diagnóstico de lesões gástricas pré-neoplásicas e neoplásicas.

Onze estudos incluindo 777 pacientes foram avaliados neste estudo. A maioria dos estudos abordaram o cancro como o principal resultado enquanto que os padrões normal, displasia e positivo para *H. Pylori* foram descritos em apenas dois estudos.

Nenhum dos estudos descreveu as lesões pré-cancerosas numa classificação única. A avaliação da qualidade da pontuação (QUADAS) variou de 8 a 12 itens. Apenas um estudo avaliou a reprodutibilidade inter-observador.

Numa análise ao nível do doente, a sensibilidade, a especificidade e o DOR de NBI para metaplasia gástrica foram 0,82 (IC 95% 0,71-0,90), 0,93 (0,89-0,96) e 53,5 (IC 95% 21,4, 133,9) respectivamente; para a displasia, 0,90 (IC 95% 0,70-0,90), 0,95 (IC 95% 0,82-0,99) e 116,9 (IC 95% 12,6-1.089,2) respectivamente; e para o cancro diferenciado, de 0,77 (IC 95% 0,72, 0,82), 0,87 (IC 95% 0,81-0,91) e 39,9 (IC 95% 12,1-130,8) respectivamente.
Concluiu-se que as definições das lesões variam entre os estudos, que a reprodutibilidade é um tema negligenciado e que a endoscopia de NBI pode ser uma ferramenta extremamente útil para o diagnóstico das diferentes lesões gástricas com um impacto importante nas decisões clínicas.

Foi realizado um estudo de derivação, relativo à aplicação de uma classificação derivada da revisão sistemática mencionada anteriormente, para avaliar a reprodutibilidade inter-observador dessa classificação.

A identificação de diferentes padrões mucosos e vasculares foi associada com alta reprodutibilidade (k = 0,82 e k = 0,91, respectivamente). Os observadores propuseram um diagnóstico histológico com alta concordância (k = 0,81).

As sensibilidade, especificidade e validade para o padrão normal foram 84 (95% CI 81-88), 81 (IC 95% 77-84) e 83 (IC 95% 79-86) respectivamente; para metaplasia, 83 (95% CI 79 - 87), 89 (IC 95% 86-92) e 87 (IC 95% 84-90) respectivamente; para gastrite por *H. Pylori*, 71 (IC 95% 67-76), 71 (IC 95% 66-76) e 71 (IC 95% 66-76), respectivamente; para a displasia 92 (IC 95% 89-95), 99 (IC 95% 98-100) e 97 (IC 95% 96-99) respectivamente.

A densidade vascular variável foi o melhor parâmetro para a identificação da gastrite por *H. Pylori*, apresentando, no entanto, uma baixa concordância inter-observador (k = 0,4).

Concluiu-se que os novos padrões mucosos e vasculares derivados a partir dos resultados histológicos foram muito válidos para metaplasia e displasia. A endoscopia de alta resolução com NBI pode mostrar eficácia na detecção de metaplasia gástrica e displasia e pode ser uma importante ferramenta para o diagnóstico e procedimentos terapêuticos precoces.

Um estudo de validação que deverá completar o estudo de derivação descrito acima irá ajudar a esclarecer o papel da endoscopia de NBI na doença gástrica, especialmente no diagnóstico de lesões gástricas pré-neoplásicas e neoplásicas.

Para o desenvolvimento da MGF é importante levar a cabo novos estudos para avaliar a reprodutibilidade e validar questionários internacionais e outras ferramentas de diagnóstico que são aplicados no contexto clínico nacional. O autor tem o objectivo de colaborar neste esforço.

1. RATIONALE

RELIABILITY

The accuracy of measurements is very important in clinical practice since most decisions are based in registered values and diagnostic tests (Shoukri, et al., 2006). Reliability is the extent to which a measure provides the same results repeatedly and under the same circumstances (Lachin, 2004) (intra-rater reliability) or two or more raters (clinicians) agree in their diagnostic test results ratings (inter-rater reliability) (Sim, et al., 2005). Thus, the assessment of both reliability and validity of diagnostic tests is required previously to their incorporation in clinical daily practice or their use in clinical research (Shoukri, et al., 2006) (Sim, et al., 2005) (Ottenbacher, et al., 1997) (Gilchrist, 2009).

GASTRIC CANCER

Even though the incidence of gastric cancer (GC) has been declining over the years it remains a major health issue. It is the fourth most common cancer and the second cause of oncological deaths worldwide (Crew, et al., 2006). 21.500 new cases and 11.000 deaths were expected for 2008 in the USA (Quiros, et al., 2009) GC is the fourth most common cause of death by cancer in Europe, being responsible for 159 900 new cases of GC and 118 200 deaths each year (Ferlay, et al., 2007). Portugal has a high incidence of GC with roughly 3700 new cases each year (Pinheiro, et al., 2003).

When GC is diagnosed it is generally encountered at an advanced stage. Western countries have a 5-year survival rate of about 20%, whereas Japan, where intense screening for GC is standard, tumors are found earlier, which leads to a higher survival. (Quiros, et al., 2009)

H. pylori infection, high ingestion of salty or smoked foods and nitrates, pernicious anemia and smoking are all considered risk factors for gastric cancer (Quiros, et al., 2009). An inherited component is present in up to 3% of all gastric cancers, due to a mutation in the E-cadherin (CDH1) gene which is responsible for cellular adhesion and epithelial integrity (Cohen, et al., 2009).

Suspected cancer is usually evaluated by Upper Gastrointestinal Endoscopy through visualization and biopsies. A computed tomography (CT) scan of the abdomen and pelvis is often done for staging (Quiros, et al., 2009). The extension of the dissemination of the disease (localized versus systemic) will determine the treatment strategy but surgery remains the treatment of choice for gastric cancer. A therapeutic approach that includes chemotherapy, radiation or a combination of both has been used to augment the probability of success, as the disease may recur locally and distally. (Quiros, et al., 2009)

ENDOSCOPY

The standpoint of diagnostic endoscopy has been changing in the last decades from the detection of unmistakably visible abnormalities to the diagnosis of more subtle abnormalities. Furthermore, endoscopic surveillance is being progressively more used in patients who are considered to be at risk for development of gastric cancers (van Sandick, et al., 1998) (Whiting, et al., 2002) (Curvers, et al., 2009).

The first concepts of image transmission using flexible fibers date from the 1920s, although only in the 1950s were the first models of flexible fibreoptic endoscopes developed by Hopkins (Hopkins, et al., 1954), Curtiss and Hirschowitz (Hirschowitz, et al., 1958) (Sivak, 2006). In 1961 Hirschowitz published the first description of a flexible endoscopy of the stomach and duodenal bulb (Telleman, et al., 2009) (Hirschowitz, 1961). The invention of the charged coupe device in the 1960s and its incorporation in the endoscope allowed electronic images visualization in the 1970s (Sivak, et al., 1983).

Endoscopy is the method of choice for detecting early lesions of the gastro-intestinal tract. When using conventional endoscopy the detection of dysplasia depends largely on the experience of the endoscopist, since it is necessary to identify small mucosal changes (Schmid, 2008). The use of other endoscopic techniques led to diagnostic accuracy enhancement of occult cancer in patients with less than 1 cm. These techniques include chromoendoscopy (CE), magnification endoscopy (ME), narrow-band imaging (NBI), autofluorescence endoscopy (AFI), and confocal laser microscopic endoscopy (Schmid, 2008). Combining different techniques is likely to improve specificity and sensitivity of these tests (Schmid, 2008).

NBI

New imaging modalities, such as NBI, may allow for better detection of these early neoplastic lesions and may thus improve the effectiveness of endoscopic surveillance and screening. NBI uses short wavelength light (essentially blue light) for tissue excitation. When combined with magnifying endoscopy, NBI allows for clearer images of superficial structures and microvessels in the gastrointestinal mucosa (Tamai, et al., 2006).

Several studies have established the usefulness of NBI for the diagnosis of esophageal neoplasias (Curvers, et al., 2009), Barrett's esophagus (Curvers, et al., 2009) (Hamamoto, et al., 2004) and colorectal mucosal lesions. (Tamai, et al., 2006). By using NBI for the observation of esophagus we may visualize the microarchitecture of the columnar epithelium. This may be helpful to identify biopsy sites to areas of intestinal metaplasia, dysplasia and neoplasia. (Sharma, et al., 2003). (Gheorghe, 2006) (Paris Workshop on Columnar Metaplasia in the Esophagus and the Esophagogastric Junction, 2005)

The role of NBI in stomach is yet to be completely described as there are only a small number studies published (Curvers, et al., 2009). Nevertheless it has been suggested that NBI may be useful for the diagnosis of gastric carcinomas (Curvers, et al., 2009) (Tamai, et al., 2006) (Endo, et al., 2005). Using NBI we can analyze the surface of the epithelium (pit pattern) and the vascular pattern (Kiesslich, et al., 2002). NBI may reveal epithelial disorganization and of the vascular network when inflammatory and neoplastic (including premalignant) lesions of the gastrointestinal tract are present.

MOTIVATION

The author was offered the opportunity to join the group studying the reliability of NBI endoscopy as a diagnostic tool in gastric disease.

Therefore, the planning of this Thesis reflects an interest in the conceptual view of reliability as well as the study of this specific clinical setting. This enabled the author to integrate the concept of reliability with a clear understanding of gastric cancer endoscopic diagnosis.

Afterwards, the knowledge obtained through the production of this Thesis will be applied to the author's main clinical field, Family Medicine, in order to assess various concerns that affect the daily clinical decisions of a Family Physician.

2. AIM

This study comprises three purposes: the first is to present a step-wise methodological approach for the implementation of a nonspecific reliability study; the second is to present a systematic review of NBI endoscopy as a diagnostic procedure in gastric disease; the third is to conduct a validation study evaluating the reliability of a new classification which was derived from the systematic review.

3. A STEP-WISE APPROACH FOR PLANNING AND CONDUCTING RELIABILITY STUDIES: BRIEF PRESENTATION

1. Introduction

Accurate measurements are of major importance in clinical practice since most decisions are based in registered values and diagnostic tests (Shoukri, et al., 2006). Thus, both reliability and validity of diagnostic tests should be assessed previously to their integration in clinical daily practice or their use in clinical research (Ottenbacher, et al., 1997) (Sim, et al., 2005) (Shoukri, et al., 2006) (Gilchrist, 2009).

Reliability is the extent to which a measure provides the same results repeatedly and under the same circumstances (intra-rater reliability) (Lachin, 2004) or two or more raters (clinicians) agree in their diagnostic test results ratings (inter-rater reliability) (Sim, et al., 2005). In other words, if a specific procedure which result is expressed as a certain type of variable or rate is used by a rater, a classifier, an observer, a certain user in separate occasions or by different observers, the same classes' or values should tend to be similar, considering the internal variability of the phenomena under measurement that, for most currently used classifications, may be considered as the disease spectrum.

Several reliability coefficients exist to allow the user of a certain diagnostic procedure to assess the extent of measurement errors (Shoukri, et al., 2006). There is extensive literature discussing and determining which coefficient to use for a specific type of variable (Williamson, et al., 2000) (Lachin, 2004) (Sim, et al., 2005) (Gilchrist, 2009). Usually, for a specific number of observers and a certain rate with a determined number and eventually order of categories, a reliability coefficient is suggested and its advantages or disadvantages are discussed (Hripcsak, et al., 2002).

However, as far as we know, there is no manuscript focused on an important issue: how to determine the number of cases (that under this set should be considered as ratings) to be used in an inter-observer reliability study? And, how to select them? I.e., should the cases be consecutively selected and should their distribution reflect that found on clinical practice or should the distribution of cases by variables' classification classes be predetermined in a case-control faction (over or underestimating a specific class proportion)?

In clinical practice, most diagnostic procedures depend on human observation or classification. In this manuscript using medical imaging such as digestive endoscopy as an example, a stepwise approach will be presented on how to plan and conduct a reliability study to assess inter-observer reliability. A series of questions will be put and answered accordingly and a list of references will be given for further details. Medical procedures expressed as a continuous variable or studies aimed at assessing intra-rater agreement will not be developed extensively.

2. What is the importance of reliability in scientific research?

The principles of validity and reliability are the basis of the scientific method and must always be assessed before using diagnostic test results in clinical practice since the absence of reliability in results, derived from two different tests or from the same test but assessed by two different raters, may produce important medical and legal consequences (Petrie, et al., 2000). Scientific research studies are usually based on hypothesis testing whose significance may also be compromised. Scientific conclusions driven from a diagnostic test study whose reliability was not assessed may not be accurate and introduces bias and type I and II errors (Petrie, et al., 2000).

Therefore, assessing reliability and subsequently validity of a measure before using it in clinical practice, sometimes with relevant pharmacological and treatment strategies implications, is mandatory.

3. Categorical variables: concept and presentation of the study case

Variables may be considered in a continuous scale or as dichotomous, nominal or ordinal categories (Petrie, et al., 2000). Glucose or total cholesterol levels are a known and practical example of a continuous variable. Positive vs. negative are a used diagnostic test classification and an example of a dichotomous scale. An example of a classification using a categorical variable is the classification proposed by Uedo (Uedo, et al., 2006) for the appearance of a Light Blue Crest (LBC) on the epithelial surface of gastric mucosa, visualized using NBI. This classification considered 4 degrees of LBC: non-LBC, LBC +, LBC ++, LBC +++. In this case, LBC is an ordinal variable.

For statistical purpose, it is possible to transform a continuous variable in a categorical one (Bartfay, et al., 2000). This is most useful when categories have a small number of individuals or when collapsing specific categories may be needed in order to properly analyze the data. One must keep in mind that collapsing categories should not put at risk the clinical relevancy of the variables that are being studied, having sometimes repercussions in terms of larger sample size requirements (Bartfay, et al., 2000). For this reason and whenever possible, continuous or multinomial data should be kept in its original scale rather than collapsing the data into a binary variable – in other words, whenever possible data should be kept in its original appearance (Bartfay, et al., 2000).

a. Specific statistical considerations regarding balanced and unbalanced samples

Considering it is not possible to study each individual, the researcher can only select a small number of individuals that are representative of the population.

However, the disease may not be equally distributed in the sample as it is in the population so it is the researcher task to evaluate the relative proportion of cases – prevalence - in each grade and select an appropriate sample (Hripcsak, et al., 2002). If the disease prevalence is similar in the sample and in each grade then the sample is designated balanced. If not, it is considered an unbalanced sample.

A balanced sample is always preferable to an unbalanced one but most authors agree that most of the times it is not possible to achieve this (Hripcsak, et al., 2002). To avoid such situation, the researcher should be aware to separate the measurement from the demonstration study (Hripcsak, et al., 2002).

When working with balanced sample, kappa statistics can be used to assess reliability. When considering an unbalanced sample the Kappa is not suitable and should be substituted by specific agreement. This is especially relevant when more than two categories are used because Kappa statistics may not represent the real agreement between raters: it is still possible to have a high kappa value regardless poor agreement in one rare grade (Hripcsak, et al., 2002).

Statistical analysis of an unbalanced sample should include observed and specific agreement Specific agreement is therefore the most adequate method since it will allow enhancing difficulties in particular categories, identifying the problem (Hripcsak, et al., 2002). The researcher must still have in consideration that a low specific agreement in a low prevalence grade is of limited value since the rater discrimination could be different in a balanced sample (Hripcsak, et al., 2002).

Adequate statistical analysis and sampling method is therefore relevant in order to obtain a balanced sample or to adequately analyze and draw conclusion from an unbalance one.

4. How to estimate reliability of a single test conducted by different observers (inter-observer)?

Interater reliability refers to a condition where different raters classify and analyze the same object/test in a single moment (Petrie, et al., 2000). One or more raters can be enrolled in the study depending on the needed sample size. When studying categorical variables we should use kappa statistics (Petrie, et al., 2000).

As an example we can consider the study conducted by Nakayoshi (Nakayoshi, et al., 2004) in which different observers carried out NBI endoscopy on the same set of patients and classified the findings in three nominal categories: fine network, corkscrew and unclassified pattern.

In this paper, we will only considered inter-rater reliability for categorical variables. We provided a step-wise approach for conducting a reliability study, with focus on implementation, sample size calculation, measurement methods and statistical analysis for reliability studies.

The main objective was to guide the investigator and summarize the essential questions when designing and implementing reliability studies, based on a review of the literature.

5. Statistical methods for the assessment of reliability

An accurate way of measuring the reliability of a measurement method is the reliability coefficient (ρ). The reliability coefficient is calculated from duplicate collections and is defined as the “proportion of total variation observed between subject values due to differences in the true values” (Shoukri, et al., 2006). ρ is the proportion of total variation observed between subject values due to differences in the true values.

The proportion of variation between observed values and attributed to error is the reliability coefficient subtracted to one (1- ρ). When ρ equals one it means that the variation among observed values and those attributed to error is zero, or in another point of view, the observed value is the real value and in consecutive collections the obtained value would be exactly the same (Shoukri, et al., 2006). Thus 1 - ρ is the proportion of variation among observed values that is due to error. When ρ = 1 it means that the reported value is the true value and that the system would yield the exact same value on duplicate collections.

a. Which statistical methods should we employ for tests using categorical variables?

Reliability of tests using categorical variables is often assessed using the Kappa statistic (K) of Cohen. The Intraclass Correlation Coefficient (ICC) is not appropriate in this situation because, although employed as a measure of reliability of categorical measurement, it cannot be used it as an estimate of the reliability coefficient (Lachin, 2004).

Several papers have shown that Kappa behaves similarly to the ICC, and that it equals the ICC approximately. (Hripcsak, et al., 2002). When considering two raters with the same probability of a positive rating Kappa equals the ICC, and so, it can be considered a measure of reliability.

Kappa values vary from –1 to 1, but they are usually included between 0 and 1. When Kappa is equal to 1 this represents perfect agreement, indicating that the raters agree in their classification of every case. When Kappa equals zero, agreement is the same as that expected by chance. This would be as if the raters had simply speculated their rating. Negative values of Kappa indicate an agreement inferior to that expected by chance (Sim, et al., 2005).

One should only compare values of Kappa form studies with similar prevalence of the observed values as well as design and categories (Hripcsak, et al., 2002).

We will now present a summary of the strategies for assessing reliability in the different types of categorical variables: binary (or dichotomous), nominal and ordinal (Sim, et al., 2005) (Uebersax). We refer the reader to the references for a more detailed view of this matter.

b. Which tests are appropriate for dichotomous data?

For this type of data one should report descriptive agreement measures as the observed agreement and specific agreement on each grade (Hripcsak, et al., 2002). In this case Kappa is the most commonly used measure of reliability, but its level should not be over interpreted (Hripcsak, et al., 2002). As previously discussed, a balanced sample is needed to produce an accurate assessment of reliability (Hripcsak, et al., 2002).

An example of a dichotomous variable is presented in the study conducted by Yao (Yao, et al., 2008). In this paper, Yao investigates the morphology of a substance that may obscure the subepithelial microvascular pattern – the white opaque substance (WOS) – and divides the observed WOS in two types, according to its distribution: regular and irregular WOS.

When using dichotomous data, the adequate statistical method depends on the number of raters. Using two raters the adequate statistical test is Cohen's Kappa (or alternatively the ICC) to assess raw agreement, overall and specific to each category. When considering multiple raters, the ICC should be used after raw agreement assessment (overall and specific to each grade) and McNemar´s test should be used to evaluate marginal homogeneity.

c. Which tests are appropriate for nominal data?

When considering nominal data with 2 categories and paired ratings one should first use a 2x2 contingency table, with notation as indicated in Table 1 (Sim, et al., 2005).

Table 1: Crosstabs using nominal variables

		Rater 2		
		Normal	Abnormal	**Total**
Rater 1	Normal	19 (a)	2 (b)	21
	Abnormal	3 (c)	15 (d)	18
	Total	22	17	39

Adapted from Peat et al, 2005

In this table cells (a) and (d) designate correspondingly the number of patients for whom both clinicians agree on the normality or abnormality of a determined test, representing agreement. Cells (b) and (c) indicate the numbers of patients on who they disagree.

As with dichotomous data, the observed agreement and specific agreement on each grade should be reported as descriptive agreement measures (Hripcsak, et al., 2002). Kappa is also the most commonly used measure of reliability, considering two or multiple raters (see Table 2).

An example of the use of kappa in nominal data is the study by Corley and colleagues (Corley, 2009). In this study the objective was to evaluate the accuracy of a clinical Barrett's esophagus diagnosis and the reliability of an esophageal intestinal metaplasia diagnosis. The inter-rater reliability was assessed for the metaplasia diagnosis comparing the diagnosis made by 2 different pathologists. Kappa was used as a measure of reliability in 88.3% of subjects (original vs. referral pathologist, inter-rater reliability; kappa =0.42, 95% CI, 0.34-0.48).

Once again, the magnitude of kappa must only be compared to that of similar experiments (Sim, et al., 2005) (Hripcsak, et al., 2002).

Table 2: Reliability assessment for nominal variables according to number of raters

Number of raters	Statistical test for nominal data
Two raters	• Assess raw agreement, overall and specific to each category. • Use Cohen's unweighted Kappa to assess reliability • Use the Stuart-Maxwell test or the Bhapkar test to test overall marginal homogeneity. • Use McNemar's test to test marginal homogeneity relative to individual categories.
Multiple raters	• Assess raw agreement, overall and specific to each category. • If different raters are used for different subjects, use the Fleiss Kappa statistic; again, as with nominal data/two raters, attend only to the p-value of the test unless one has a genuine basis for regarding all pairs of rating categories as equally "disparate". • Conditional tests of marginal homogeneity can be made within the context of latent class modeling. • Use graphical displays to visually compare the proportion of times raters use each category (base rates), or alternatively, consider each pair of raters individually and proceed as described for two raters.

d. Which tests are appropriate for ordinal data?

For this type of data it is more appropriate to use a weighted value of kappa (weighted kappa). The use of weighted kappa will allow retaining the hierarchical nature of the categories (Sim, et al., 2005). This will reflect the degree of differences between ratings, giving more relevance to large differences between ratings (Sim, et al., 2005).

Some authors consider that the selection of relative weights for each type of disagreement remains an arbitrary strategy, as with the choice of category limits (Hripcsak, et al., 2002). The same authors suggest the use of a polychoric correlation or latent trait model if one considers it is not logical to allocate a numeric score to each of the categories (Hripcsak, et al., 2002) .

An example of an ordinal variable is presented by Tahara et al (Tahara, et al., 2009) when investigating gastric mucosal patterns by using magnifying NBI endoscopy with the objective to identify any relationship between those patterns and H. pylori–induced gastritis. Tahara considers a serological degree of atrophic gastritis given by an ordinal variable: normal, mild, moderate and severe.

The ICC is known as a reliability index that represents the "proportion of the between-subject variance to the total variance" for continuous measurements (Shoukri, et al., 2006) (Lachin, 2004) (Dunn, 1992) (Shoukri, 2004). However, its use is also recommended for ordinal measures allowing in this specific case to weight the differences between ratings. In fact, several papers have shown that kappa behaves similarly to the ICC, and that it equals the ICC approximately (Hripcsak, et al., 2002).

As most statistical tests, using the ICC has some advantages as well as disadvantages (Shoukri, et al., 2006). Using the ICC will allow the estimation of the exact power based on a previously determined number of subjects and number of replicates (Shoukri, et al., 2006) (Quan, et al., 1996). This will lower study costs and assist in the choice of an optimal design study, facilitating the choice of the number of subjects and replicates needed to achieve the expected power (Shoukri, et al., 2006).

Despite all the advantages, the ICC also has some limitations. Its results are affected by prevalence which influences result interpretation and comparison from different studies (Shoukri, et al., 2006).

For ordinal variables, the ICC can either be used when considering two or multiple raters (see table 3). For two raters the Weighted Kappa with Fleiss-Cohen (quadratic) weights is an alternative to the ICC. Also, ordered rating levels often imply a latently continuous trait. If so, measure association between the raters with the polychoric correlation or one of its generalizations.

When considering multiple raters, latent trait models can also be used, as well as graphical examine. It may also be useful to compare rater base rates and/or thresholds for various rating categories.

Table 3: Reliability assessment for ordinal variables according to number of raters

Number of raters	**Statistical test for nominal data**
Two raters	• Estimate the ICC or alternatively use the *Weighted Kappa* with Fleiss-Cohen (quadratic) weights. • Test overall marginal homogeneity using the Stuart-Maxwell test or the Bhapkar test. • Use the McNemar's tests to test for differences in rater thresholds associated with each rating category and to test for a difference between the raters' overall bias.
Multiple raters	• Estimate the ICC. • Test for differences in rater bias using ANOVA or the Friedman test.

An example of the use of ICC is the study conducted by Irani (Irani, et al., 2009) that investigates the capability of narrow-band imaging in combination with computerized image analysis to quantitatively assess airway vascularity in lung transplant recipients. In this study three representative pictures were chosen from every site and the ICC (measure for test-retest reliability) of the three measurements was determined as a measure of reliability.

Theoretically, it is possible to use kappa in ordinal categories that result from continuous data. Nevertheless, the choice of the category limits is arbitrary which lead to the conclusion that the kappa value produced using this method is of little meaning (Sim, et al., 2005). Concurrently, there is usually a loss of statistical power when one uses kappa statistic method (Sim, et al., 2005).

e. How many subjects and observers do we need to conduct a reliability study when using categorical variables?

An essential question in reliability studies arises of the need to determine the optimal combination k subjects of the study and n raters (i.e. observations, measurements, etc) that permits an accurate estimation of the ICC (Shoukri, et al., 2006).

Sample size estimation is usually of the most difficult tasks for inexperienced clinical researchers. Unfortunately, it is one of the most important too. However, further and systematic investigation is still desired (Shoukri, et al., 2006).

This is of foremost importance as one limiting factor in designing and implementing reliability studies in many situations is the difficulty of arranging for replicated observations. For instance, in many clinical situations the number of specialists qualified or willing to participate in a study may be inferior to the number dictated by the study original design. One must also keep in mind funding constrains that may undermine the recruitment of subjects for a reliability study (Shoukri, et al., 2006).

Confidence intervals may have an important function in the determination of the sample size of a reliability study. In the planning stages of a reliability study one should determine the size of the confidence interval (Shoukri, et al., 2006). In the planning stages of a reliability study one should determine the size of the confidence interval (Shoukri, et al., 2006).

Donner and Eliasziw (Walter, et al., 1998) concluded that the increase in the number of subjects will increase power up to a maximum or optimal level beyond which no increase in power can be achieved by adding more subjects.

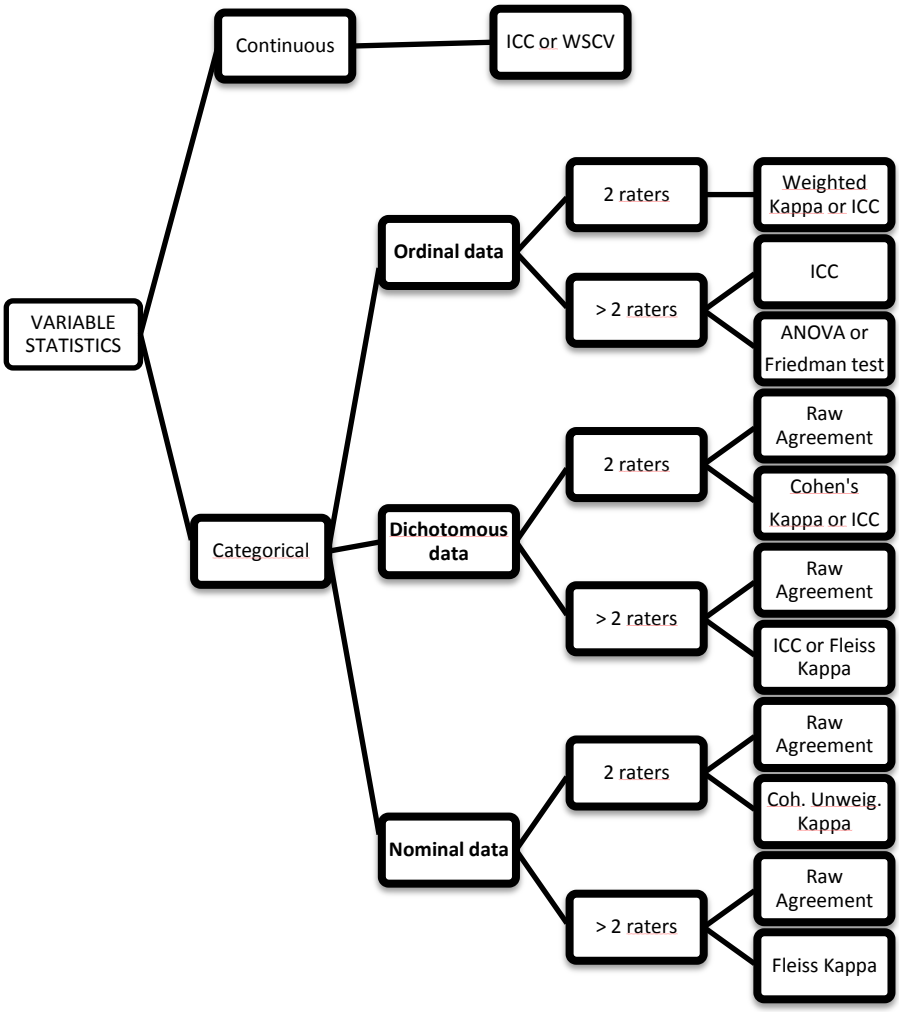
When choosing the power of the test, this is often settled by choosing a sample size that will provide power of 80% (type II error of 0.2) (Shoukri, et al., 2004). The precision of the test refers to the reliability of the power estimate, or the expected effect size. The narrower the confidence intervals that contain the desired effect size, the more reliable or reproducible the estimate (Shoukri, et al., 2004).

When designing a reliability study we should keep in mind that if the sample size is too small the study may produce an inaccurate estimate of the reliability coefficient, whereas if it is too large it may constitute a misuse of resources. This balance is critical to achieving a good ratio between the precision of the reliability coefficient and the costs involved (Shoukri, et al., 2006).

It is desirable to maximize the cost of a reliability study with an optimal balance between the required numbers of independent raters and the number of subjects they are required to rate (Shoukri, et al., 2004).

6. Development of a reliability study – a decision pathway

Figure 1: Flowchart describing a decision pathway for the development of a reliability study concerning selection of statistical analysis method for each type of variable.



4. SYSTEMATIC REVIEW AND META-ANALYSIS OF NARROW-BAND IMAGING ENDOSCOPY FOR DIAGNOSIS OF GASTRIC PRECANCEROUS LESIONS AND CANCER

1. Introduction

The incidence of gastric cancer (GC) has been declining over the years, but it remains a major health issue as it is the fourth most common cancer and the second cause of cancer deaths worldwide (Crew, et al., 2006). GC incidence is particularly high in East Asia, Eastern Europe, and parts of Central and South America, and it is about twice as high among men than among women (Brenner, et al., 2009). In the USA 21.500 new cases and 11.000 deaths were expected for 2008 (Quiros, et al., 2009). In Europe, the proportion of cured cases varies from about 9% to 27%, depending on the country and therapeutic approach considered (Francisci, et al., 2009).

Due to the fact that GC is generally diagnosed at an advanced stage, at least in Western countries, 5-year survival averages 20%. In Japan, where intense screening for GC is standard, tumors are found at an earlier stage, which leads to a higher survival (WHO, 2001) (Verdecchia, et al., 2004) (Quiros, et al., 2009).

Although there has been a strong improvement in endoscopic image quality, the early visualization of neoplasia in early stages is still a difficulty with important consequences in its misdiagnosis (Yalamarthi, et al., 2004) (van Rijn, et al., 2006) (Curvers, et al., 2009). New imaging modalities, such as NBI may allow for better detection of these early neoplastic lesions and may thus improve the effectiveness of endoscopic surveillance and screening.

The evaluation of gastric mucosa with NBI has demonstrated that it is possible to visualize distinct mucosal and vascular patterns in the fundus and in the antrum (Gheorghe, 2006). The fundus is characterized by a mucosa with regular small pit openings and subepithelial capillaries surrounding the pits, with a regular honeycomb pattern (Yao, et al., 2002) (Nakayoshi, et al., 2004) (Kwon, et al., 2005) (Gheorghe, 2006). The antrum presents epithelial crests separated by narrow grooves, with elongated subepithelial capillaries in the centre of the epithelial crests (Yao, et al., 2002) (Nakayoshi, et al., 2004) (Kwon, et al., 2005) (Gheorghe, 2006) .

Using NBI we can analyze the surface of the epithelium (pit pattern) and the vascular pattern (Kiesslich, et al., 2002). NBI may reveal epithelial disorganization and of the vascular network when inflammatory and neoplastic (including premalignant) lesions of the gastrointestinal tract are present. Pit pattern is easier to identify in large bowel when compared to stomach, because of gastric inflammation related to the prevalence of *H. Pylori* (The Paris endoscopic classification of superficial neoplastic lesions: esophagus, stomach, and colon, 2003).

Regarding to Barrett's esophagus (BE), Sharma (Sharma, et al., 2006) concluded that the correlation between the various mucosal and vascular patterns and the histology. In this study (Sharma, et al., 2006) NBI is considered a diagnostic tool with a high degree of accuracy for the detection of metaplastic and dysplastic tissue within the BE segment.

For all this, the aims of this systematic review and meta-analysis were to evaluate the 1) existing definitions, 2) reporting quality, 3) reliability assessment and 4) calculate the pooled diagnostic accuracy measures of the NBI with magnification value in characterizing gastric precancerous and cancerous lesions.

2. Materials and methods

Search strategy

A sensible search in the PubMed database was performed in order to retrieve published research articles, up to April 2010, using the NBI for the pre-malignant gastric lesions' detection running the following query:

(((“atrophy” [All Fields] AND “stomach” [All Fields]) OR ("gastritis"[MeSH Terms] OR "gastritis"[All Fields]) OR ("helicobacter pylori"[MeSH Terms]) OR "helicobacter pylori"[All Fields]) OR ("intestinal metaplasia"[All Fields] AND “stomach” [All Fields]) OR (("dysplasia"[All Fields] OR "cancer"[All Fields]) AND “stomach” [All Fields]) AND ("narrow band imaging"[All Fields] OR NBI[All Fields] OR "high resolution"[All Fields])) AND ((sensitiv*[Title/Abstract] OR sensitivity and specificity [MeSH Terms] OR diagnos*[Title/Abstract] OR diagnosis [MeSH:noexp] OR diagnostic * [MeSH:noexp] OR diagnosis,differential[MeSH:noexp] OR diagnosis[Subheading:noexp])))

Study selection

The retrieved manuscripts were selected if they fulfilled the following criteria:

Inclusion criteria: Studies published until April 2010 evaluating the NBI value to detect pre-malignant gastric lesions.

Exclusion criteria: Languages other than English, French, Spanish or Portuguese; non gastric lesions analysis; other technologies' assessment besides NBI; review articles (including meta-analysis and systematic reviews). Although the last were not included its reference list was verified in order to retrieve other pertinent articles that were not found through the query.

The first step was the article's title and abstract analysis. Those considered pertinent were assessed in their integral form and its reference list analyzed (figure 2). Each step was performed by two investigators (RP and MMS) independent and blind to each other. Disagreement cases were resolved by consensus.

Data extraction and assessment of study quality

For all the included studies in this meta-analysis, data were extracted using a form previously created for this study, by three investigators (RP, MMS and PPN). The following variables were collected: author(s), year of publication, type of study, participants' characteristics, intervention assessed, reference standard, mucosal and vascular patterns definitions [normal, metaplasia, dysplasia/ adenoma, cancer and H. pylori positive] and reliability assessment performed.

Article's quality assessment was performed using the QUADAS checklist (table 4) by three investigators (RP, MMS and MA) independent and blind to each other. Once again disagreement cases were resolved by consensus.

Table 4: Quality Assessment of Diagnostic Accuracy Studies (QUADAS) checklist

1. Was the spectrum of patients' representative of the patients who will receive the test in practice?
2. Were selection criteria clearly described?
3. Is the reference standard likely to classify the target condition correctly?
4. Is the period between reference standard and index text short enough to be reasonably sure that the target condition did not change between the two tests?
5. Did the whole sample or a random selection of the sample received verification using a reference standard?
6. Did patients receive the same standard reference regardless of the index test results?
7. Was the reference standard independent of the index text (i. e., the index test did not form part of the reference standard)?
8. Was the execution of the index test described in sufficient detail to permit replication of the test?
9. Was the execution of the reference standard described in sufficient detail to permit its replication?
10. Were the index test results interpreted without knowledge of the results of the reference standard?
11. Were the reference standard results interpreted without knowledge of the results of the index test?
12. Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?
13. Were uninterpretable/intermediate test results reported?
14. Were withdrawals from the study explained?

For the selection process and for each QUADAS' item inter-observer agreement was evaluated using the proportion of agreement (and respective 95% confidence intervals).

STATISTICAL ANALYSIS

The data analysis was performed using the Meta-DiSc (version 1.4) software in order to calculate the pooled diagnostic accuracy measures [sensitivity, specificity, diagnostic odds ratio (DOR) and area under the systematic receiver operating curve (SROC)] and respective 95% confidence intervals (CI) and heterogeneity statistic (I^2). All measures were calculated using a random-effects model.

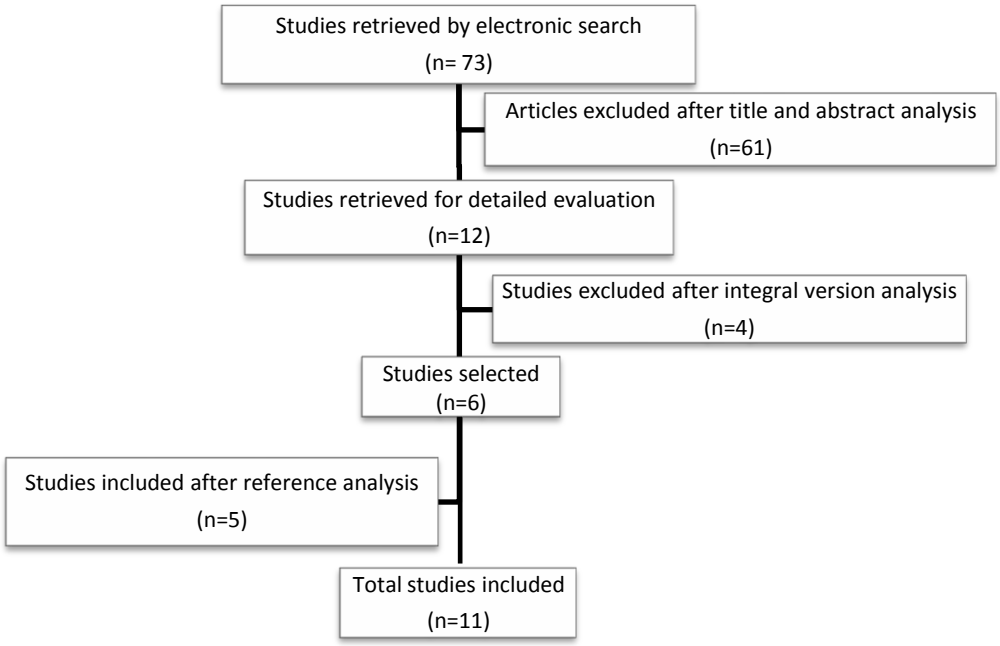
3. Results

Study Selection and quality

The study selection process is summarized in figure 2. From the 73 articles retrieved by the query only 6 were included. Through the included articles and reviews reference list analysis 5 additional articles were included, performing a total of 11 studies included in this meta-analysis.

The inter-observer agreement proportion for the study selection was excellent (AP of 0.96, 95% CI 0.92-1) for the article's selection through the title and abstract analysis and perfect for the integral version analysis (AP of 1).

Figure 2: Study selection flow diagram



A total of 10 inception cohorts and 1 case-control study were selected. The evaluation using the QUADAS checklist by the three investigators (RP, MMS and MA) presented a perfect inter-observer agreement (AP of 1).

Table 5: Quadas checklist application to the included studies

Authors	Checklist Item number													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Nakayoshi et al, 2004	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	*U*	Y	**N**	**N**
Endo et al, 2005	Y	**N**	Y	Y	Y	Y	Y	Y	Y	Y	*U*	Y	**N**	**N**
Tamai et al, 2006	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	*U*	Y	**N**	**N**
Uedo et al, 2006	Y	**N**	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	**N**	**N**
Bansal et al, 2008	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	**N**	**N**
Yao et al, 2008	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	**N**	**N**
Kaise et al, 2009	Y	**N**	Y	Y	Y	Y	Y	Y	**N**	Y	*U*	Y	**N**	**N**
Kadowaki et al, 2009	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	*U*	Y	**N**	**N**
Kato et al, 2009	Y	Y	Y	Y	**N**	**N**	Y	Y	**N**	Y	*U*	Y	**N**	**N**
Tahara et al, 2009	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	**N**	**N**
Ezoe et al, 2010	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	*U*	Y	**N**	Y

Possible Items Responses: Y- Present, **N**- Absent, *U*- Unclear

As pointed out in table 5, in three studies (27%) the selection criteria was not clearly described (item 2), in one study (10%) the sample did not received verification using a reference standard (item 5) and in another (10%) the patients did not received the same standard reference regardless of the index test results (item 6).

Additionally, in two studies (18%) the execution of the reference standard was not described in sufficient detail. Furthermore, in seven studies (64%) it is unclear if the reference standard results were interpreted without the knowledge of the results of the index test (item 11), in all the studies (100%) uninterpretable and/or intermediate test results were not reported (item 13) and withdrawals from the study (item 14) were not explained in ten of them (91%).

Table 6 summarizes the findings of the eleven articles that were retrieved. Ten were inception cohorts and one a case-control study. There were not retrieved any validating cohorts. Nine of the articles scored, at least, 10 points (in a maximum of 14) in the QUADAS evaluation (mean of 11 items present).

Cancer was the endoscopic feature which was most studied (7 papers) followed by intestinal metaplasia (IM). We didn't find any article evaluating the whole spectrum of gastric disease (from IM to cancer). Around ¾ of the studies evaluated only one outcome, one study (Yao, et al., 2002) assessed two outcomes simultaneously and only two (Bansal, et al., 2008) (Tahara, et al., 2009) evaluated three outcomes at the same time.

Only one study (Kaise, et al., 2009) reported an inter-observer reliability evaluation, through kappa values calculation, of the following outcomes: microvascular dilation, microvascular heterogeneity, fine mucosal structure disappearance and general diagnosis (benign or malignant lesions). All the outcomes showed low to modest reliability values (k values from 0.34 to 0.54).

Table 6: Selected studies characterization

Author, year	QUADAS Score	n	Intervention	Reliability assessment	Outcome				
					Normal	IM	Dysp./Ad.	Cancer	HP
Inception cohorts									
Tahara et al, 2009	12	106	NBI	No	✓	✓			✓
Bansal et al, 2008	12	47	NBI	No	✓	✓			✓
Yao et al, 2008	12	46	NBI	No			✓	✓	
Ezoe et al, 2010	11	57	NBI + ME	No				✓	
Kadaowaki et al, 2009	11	40	NBI	No				✓	
Uedo et al, 2006	11	107	NBI	No		✓			
Tamai et al, 2006	11	32	NBI	No			✓		
Nakayoshi et al, 2004	11	165	NBI	No				✓	
Endo et al, 2005	10	15	NBI	No				✓	
Kaise et al, 2009	9	100	NBI + ME	**Yes**				✓	
Case control									
Kato et al, 2009	8	62	NBI + TME	No				✓	

IM: Intestinal Metaplasia; Disp./Ad.: Dysplasia/Adenoma; NBI – Narrow Band Imaging; ME: Magnifying endoscopy; TME: Trimodal Imaging Endoscopy

OUTCOME DEFINITIONS: NORMAL PATTERN, POSITIVE HELICOBACTER PYLORI, METAPLASIA, DYSPLASIA, DIFFERENTIATED AND UNDIFFERENTIATED CANCER

A total of eleven reports were retrieved. These studies described different patterns concerning the mucosal and vascular patterns.

Two studies described the normal pattern. Bansal (Bansal, et al., 2008) tested the feasibility of NBI to predict gastric histologic diagnosis and described normality as a regular circular mucosal pattern and a regular capillary network with a normal distribution of capillaries.

Tahara (Tahara, et al., 2009) investigated gastric mucosal patterns using NBI to identify any relationship between those patterns and *H. pylori*–induced gastritis. The presence of small, round pits surrounded by regular subepithelial capillary networks, which are regularly interspersed with collecting venules was considered as a predefined indicator of normality.

Three studies evaluated intestinal metaplasia using NBI endoscopy. The paper by Uedo (Uedo, et al., 2006) investigated the value of NBI with magnifying endoscopy (ME-NBI) for diagnosing gastric intestinal metaplasia. Evidence was found that a Light Blue Crest (LBC), which is a fine, blue-white line on the crests of the epithelial surface/gyri, correlated with histological evidence of intestinal metaplasia. Bansal et al (Bansal, et al., 2008) were able to visualize a ridge or villous mucosal pattern with variable changes in the vascularity.

Tahara et al (Tahara, et al., 2009) observed two mucosal patterns that related to IM. They concluded that the degree of chronic inflammation was higher in mucosal patterns with obviously enlarged oval or prolonged pit with increased density of irregular vessels. In this study, the visualization of well-demarcated, oval or tubulovillous pit with clearly visible coiled or wavy vessels was a predictor of IM.

The endoscopic features of dysplasia and adenoma were described by Yao (Yao, et al., 2008) and Tamai et al (Tamai, et al., 2006). The work by Yao (Yao, et al., 2008) evaluated the presence and morphology of White Opaque Substance (WOS, a white substance within the neoplastic epithelium that may obscure the subepithelial microvascular pattern (MVP), resulting in its difficult visualization). They found that WOS is more frequently present in adenomas than in carcinomas and that 100% of adenomas demonstrated a regular distribution of WOS. Among the lesions in which the MVP could be visualized, 86% of adenomas demonstrated a regular MVP.

Tamai et al (Tamai, et al., 2006) conducted a study to elucidate clinical and endoscopic characteristics of depressed gastric adenoma. In this study, comparing depressed and protruding adenomas, depressed adenomas appeared endoscopically as reddish and large. 71% of depressed adenomas presented a regular ultrafine network of microvessels (UNM), whereas no protruding adenoma showed UNM.

Six articles referred to the visualization of cancer using NBI endoscopy. Nakayoshi et al (Nakayoshi, et al., 2004) measured the correlation between the images obtained with NBI and histological findings, especially with regard to vascular pattern in early gastric cancer. They characterized differentiated-type depressed early gastric cancer as more likely to present a relatively regular fine network pattern. Undifferentiated-type presented a relatively irregular, twisting or corkscrew pattern, with a relatively low density of microvessels. In this study Nakayoshi et al (Nakayoshi, et al., 2004) described atrophic gastritis as a series of various types of capillary patterns, such as thinning, stretching, and a lack of regular capillary patterns.

Endo et al (Endo, et al., 2005) conducted a study of the tumor vessels in depressed-type early gastric cancers using NBI. They observed a grid network pattern with hypervascularity associated with differentiated early-type gastric cancer. Undifferentiated early-type gastric cancer was associated with a short twig or branch-like pattern with hypovascularity.

The research by Kaise (Kaise, et al., 2009) focused on evaluating ME-NBI criteria for cancer diagnosis in superficial depressed gastric lesions in comparison to conventional white light endoscopy. They concluded that the absence of fine mucosal structure and dilation, tortuousness, heterogeneity and abrupt vessel caliber alteration were significantly linked to gastric cancer.

Kadaowaki (Kadowaki, et al., 2009) studied four magnifying endoscopy methods to determine which is most effective in enhancing the recognition of early gastric cancer demarcations. They concluded that NBI combined with enhanced-magnification with acetic acid at the mucosal surface was the most useful method for identifying early gastric cancer demarcations. NBI associated to magnifying endoscopy improved recognition of early gastric cancer demarcations by showing differences in capillary structure.

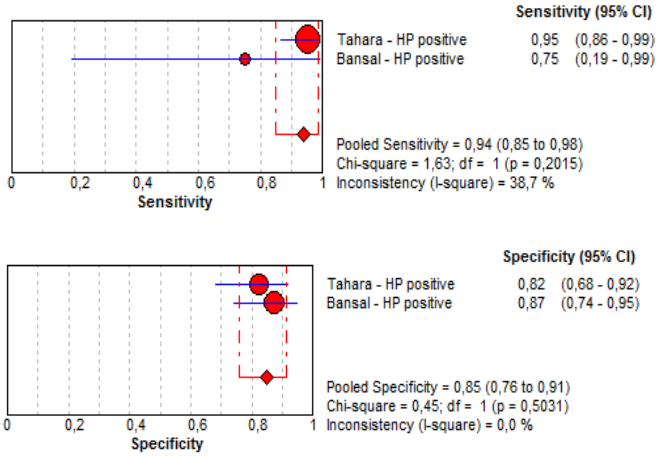
The article by Ezoe et al (Ezoe, et al., 2010) aimed to compare the real-time diagnostic accuracy of magnifying white-light imaging (WLI) and magnifying NBI for gastric small depressive lesions. In this study NBI showed demarcation lines between cancer and normal mucosa in 83% of the cancerous lesions and irregular micro-vascular pattern in 73% of the cancerous lesions.

Kato et al (Kato, et al., 2009) conducted a study to investigate the diagnostic potential of trimodal imaging endoscopy (TME), which combines WLI, autofluorescence imaging (AFI), and NBI, for the diagnosis of superficial gastric neoplasia. In this study the criteria for superficial gastric neoplasia as visualized by NBI was the disappearance of the fine mucosal structure and microvascular irregularities showing dilation, abrupt caliber alteration, heterogeneity in shape, and tortousness. In the study by Yao et al (Yao, et al., 2002) 83% of carcinomas showed an irregular distribution of WOS. Among the lesions in which the MVP could be visualized, 96% of carcinomas demonstrated an irregular MVP.

Diagnostic Accuracy

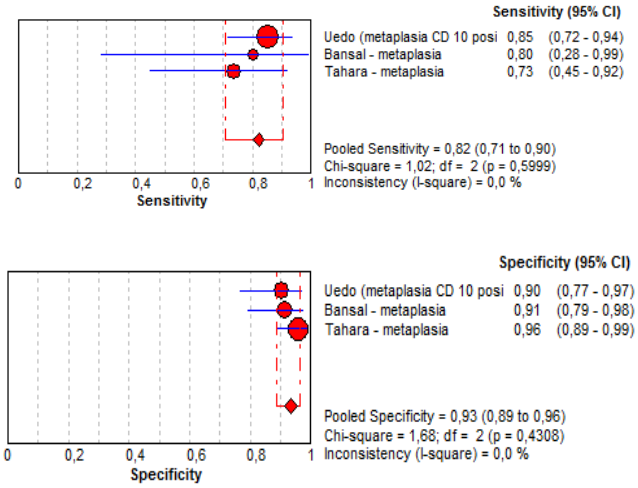
Eleven studies including 777 patients were analyzed for detection of positive *H. pilory*, metaplasia, adenoma / dysplasia, differentiated and undifferentiated gastric cancer on per-patient analysis.

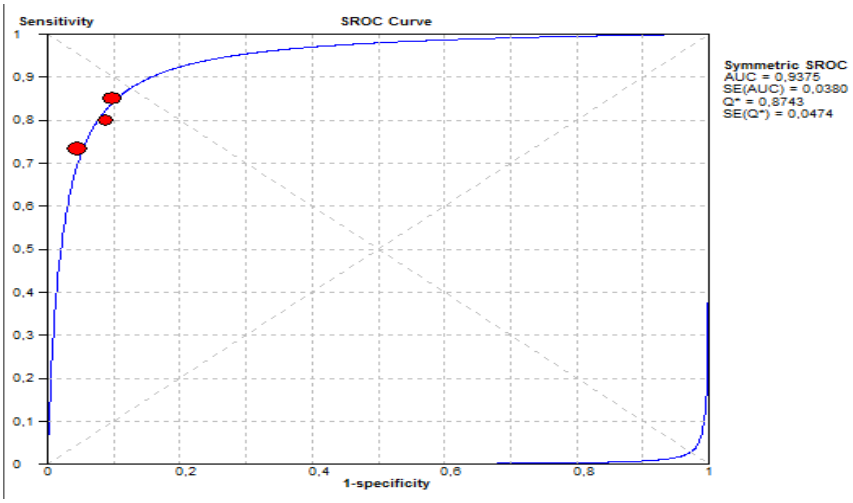
Figure 3: Positive Helicobacter pylori detection by narrow band imaging: sensitivity and specificity



The pooled sensitivity, specificity and diagnostic odds ratio (DOR) for the NBI *H. pylori* presence detection were 0.94 (95% CI 0.85-0.98), 0.85 (CI 95% 0.76-0.91) and 60.4 (CI 95% 16.8-217.6) respectively (figure 3). The SROC was not performed as only two studies evaluated this outcome.

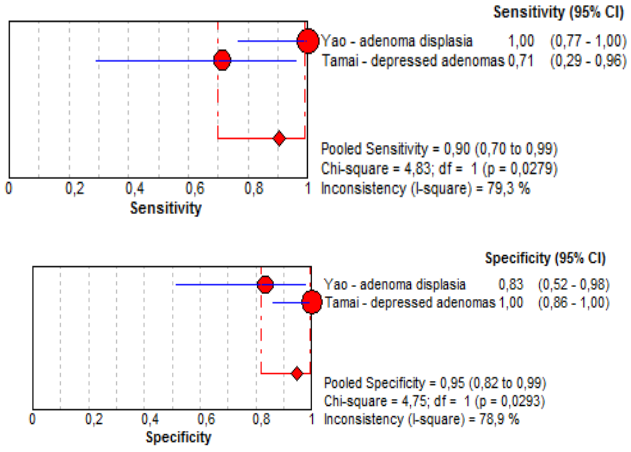
Figure 4: Metaplasia detection by narrow band imaging: sensitivity, specificity and respective systematic receiver operating curve (SROC)





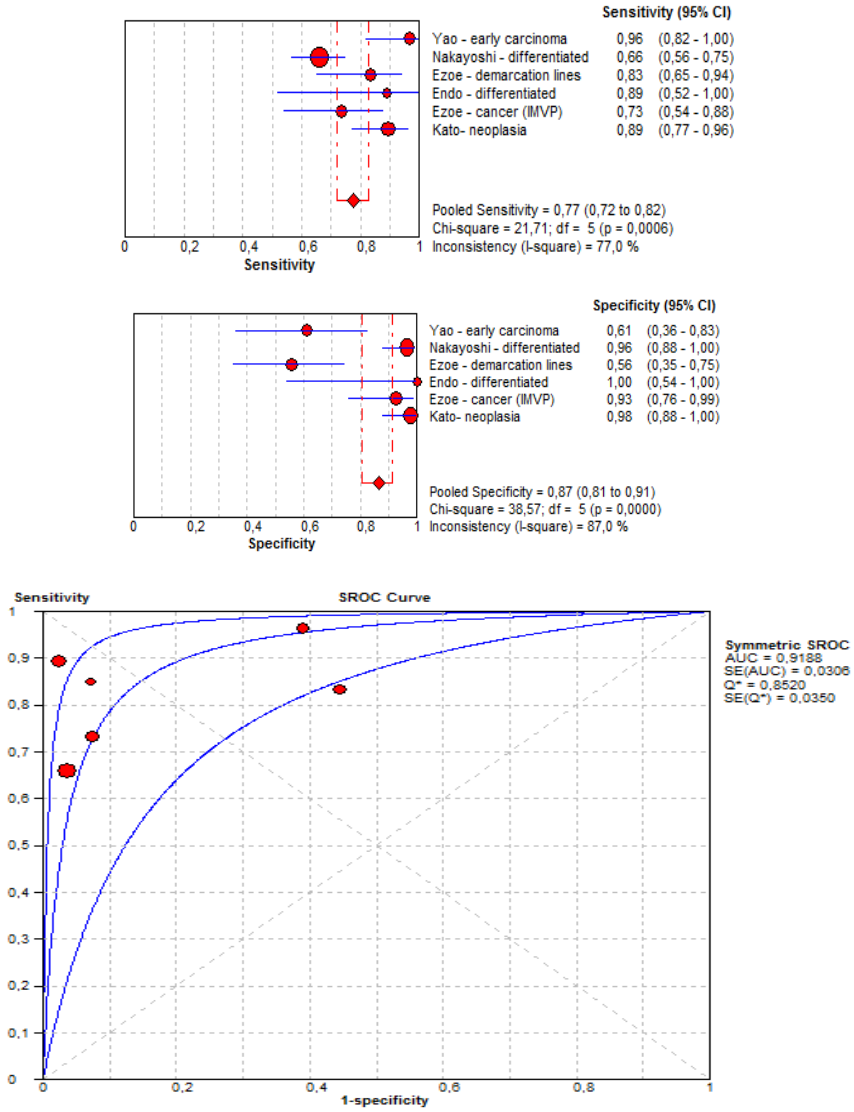
For metaplasia detection NBI presented the following pooled sensitivity and specificity of 0.82 (95% CI 0.71-0.90) and 0.93 (0.89-0.96) respectively. The pooled DOR was 53.5 (95% CI 21.4-133.9). The AUC was 0.94 with a Q* of 0.87 (SE 0.05) (figure 4).

Figure 5: Adenoma / dysplasia detection by narrow band imaging: sensitivity and specificity



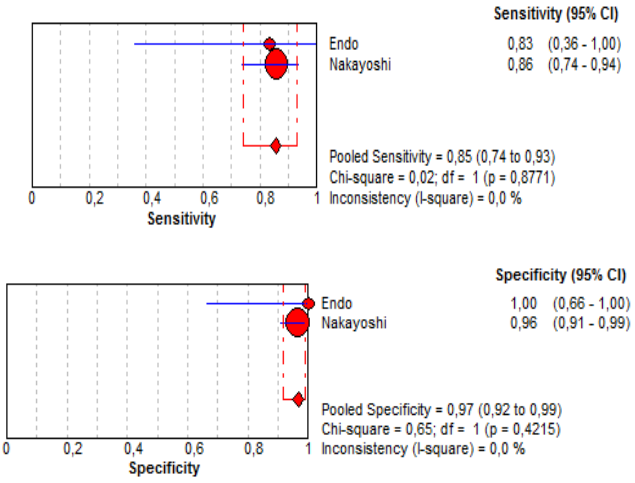
The pooled sensitivity and specificity of NBI for the detection of adenoma / dysplasia were 0.90 (CI 95% 0.70-0.90) and 0.95 (95% CI 0.82-0.99) respectively, which can be translated into a DOR of 116.9 (CI 95% 12.6-1089.2) (figure 5). Once again the SROC was not performed due to the fact that only two studies addressed this outcome.

Figure 6: Differentiated cancer detection by narrow band imaging: sensitivity, specificity and respective SROC



The NBI pooled sensitivity, specificity and diagnostic odds ratio (DOR) for differentiated cancer detection were 0.77 (95% CI 0.72-0.82), 0.87 (CI 95% 0.81-0.91) and 39.9 (CI 95% 12.1-130.8) respectively. The AUC was 0.92 with a Q* of 0.85 (SE 0.04) (figure 6).

Figure 7: Undifferentiated cancer detection by narrow band imaging: sensitivity and specificity



The NBI pooled sensitivity, specificity and DOR for the undifferentiated cancer detection were 0.85 (CI 95% 0.74-0.93), 0.97 (CI 95% 0.92-0.99) and 142.7 (CI 95% 44.3-459.9) respectively. As only two studies addressed this outcome the SROC was not performed.

For all the calculated pooled diagnostic accuracy measures for positive *H. pylori*, metaplasia and undifferentiated cancer the heterogeneity was low (I^2 inferior to 50%). On the other hand, for adenoma / dysplasia and differentiated cancer the heterogeneity was high (I^2 superior to 75%).

4. Discussion

Principle findings

The role of NBI in stomach is yet to be completely described as there are only a small number studies published, when compared with other endoscopic technologies (Areia, et al., 2010). In this meta-analysis only eleven studies met the inclusion criteria including 777 patients and presenting definitions of the following gastric lesions: presence of *H. pylori*, metaplasia, dysplasia/ adenoma, differentiated and undifferentiated cancer.

The most defined and evaluated outcome was gastric cancer (7 studies) while normal tissue, dysplasia and positive *H. pylori* were described only in two studies. Additionally, for each outcome definitions varied. A standardization of this outcomes' definition is a vital step yet to perform.

We have observed that in more than half of the studies it was unclear if the reference standard results were blind to the index test and that in almost all of them uninterpretable and / or intermediate results were not reported and withdrawals were not explained. This represents some important limitations of the available evidence. On the other hand, the QUADAS score varied from 8 to 12 items (out of 14), with a mean of 11 items present, which represents a reasonably good reporting methodology.

Sample sizes were from 15 (Endo, et al., 2005) to 165 (Nakayoshi, et al., 2004) (mean 71) indicating the difficulty to enroll pertinent patients and greatly affecting the confidence interval range. Furthermore, reliability was assessed only by one study (Kaise, et al., 2009) using different outcomes from those described in our study.
The *H. pylori* detection presented the higher sensitivity (0.94; 95% CI 0.85-0.98) while the undifferentiated cancer diagnostic showed the highest specificity (0.97; CI 95% 0.92-0.99) and DOR (142.7; CI 95% 44.3-459.9). The late is an extremely important result, since it represents a major alteration in the post-test odds and, also due to its consequence, can affect medical decision making.

The AUC was only viable to perform for two outcomes: metaplasia (0.94) and differentiated cancer (0.92). Additionally, for all outcomes sensitivity was equal or superior to 0.75, specificity to 0.85 and DOR to 40 which indicates the excellent NBI accuracy for all the different gastric lesions. Only for the adenoma / dysplasia and differentiated cancer diagnostic accuracy measures the statistical heterogeneity was high. This fact highlights the necessity of performing further studies for this outcomes detection using the NBI.

Clinical implications

Our results related to the accuracy of NBI provide valuable data in support of its use on a daily basis as a diagnostic tool for the different gastric lesions, but particularly metaplasia and differentiated cancer. The clinical consequences of this distinction need to be properly evaluated in matters such as early cancer recognition and changing of therapeutic approaches.

Before we disseminate NBI endoscopy, we also need to assess its interobserver reliability. For this we should consider validating an endoscopic classification for NBI endoscopy that may be used by different observers in different clinical settings.

Limitations of the study

Although we have made an effort to minimize it, all meta-analyses intrinsically comprise limitations.

Our query presents some limitations since it only detected 73 pertinent studies and from the 11 included studies almost half were retrieved by reference analysis. On the other hand, it was very difficult to refine it in order to equilibrate the detection of all the pertinent studies and a reasonable number of retrieved studies.

The study selection process was performed according to QUADAS guidelines and presented excellent to perfect agreement between the investigators. We have used only the agreement proportion due to the impact of the low number of included studies in the kappa value, leading to a misguiding value.

Although the reporting methodology of the retrieved studies, according to QUADAS, was reasonably good sample sizes varied from 15 to 165 (mean 71), which may partly explain the high heterogeneity found for some outcomes (namely dysplasia and differentiated cancer). Additionally, we have verified that outcomes' definition and techniques varied from study to study and that in more than half of the studies it was unclear the level of blinding for the reference test interpretation.

The definition of normal tissue, dysplasia and positive *H. pylori* were described only in two studies and the NBI diagnostic accuracy for *H. pylori*, dysplasia and undifferentiated cancer were evaluated again only in two studies. Furthermore, no single validation study was retrieved.

5. Conclusions

To the best of the author's knowledge, this is the first ever reported assessment of the QUADAS statements and meta-analysis in the field of NBI endoscopy for gastric pre-cancerous and cancerous lesions. We have found that, for all outcomes NBI showed excellent sensitivity, specificity, DOR and AUC values indicating that it is an extremely useful for the different gastric lesions detection.

Although there are promising results (Uedo, et al., 2006) (Tamai, et al., 2006) for different type of uses for NBI we are still in an initial stage of use and comparison with WLI. Additional investigation is needed to determine if the uses suggested by these studies for NBI (namely for the detection of early neoplasia and differentiation of grades of histological diagnosis) are applicable methods for the future. As so, the daily practice use of NBI in the stomach needs further validation before being implemented.

For all this, a future validation research with standardized outcomes definition, adequate sample size, including the complete spectrum of gastric lesions, reporting the uninterpretable and / or intermediate results and explaining withdrawals is required in order to truly assess the NBI value for gastric lesions diagnostic and therefore improve its paper in medical decision making.

5. RELIABILITY OF HIGH RESOLUTION NBI ENDOSCOPY FOR DIAGNOSIS OF INTESTINAL METAPLASIA AND GASTRIC DYSPLASIA – A DERIVATION STUDY

1. Introduction

Gastric adenocarcinoma is the second most lethal cancer worldwide (Black, et al., 1997) with only a minority of gastric adenocarcinomas diagnosed at a curable and ressectable form (Hundahl, et al., 1997). Even though modifications in risk factors, such as in diet and in the prevalence of *H. pylori* infection (Pinheiro, et al., 2003) have been pointed as important reasons for a reduction in stomach cancer related incidence rate, secondary prevention through diagnosis of pre-malignant lesions and early gastric cancer (by screening or by follow-up of at high-risk individuals) would probably be the most immediate strategy for improving survival (Zera, et al., 1993) (Stemmermann, et al., 2002).

Diverse descriptions of new methods of high-resolution with NBI (HR-NBI), have been published, and seem to present good results for intestinal metaplasia and cancer, but some methodological weaknesses have been found: reliability was seldom evaluated, no single description for all the spectrum of lesions was made and no external validation of any derivative description is known.

We aimed at assessing the reliability of the previously described NBI features (Chapter 4), to simplify them under a new classification. Therefore, we conducted a derivation study, concerning the application of a classification derived from the systematic review previously mentioned in order to assess the inter-observer reliability of this classification.

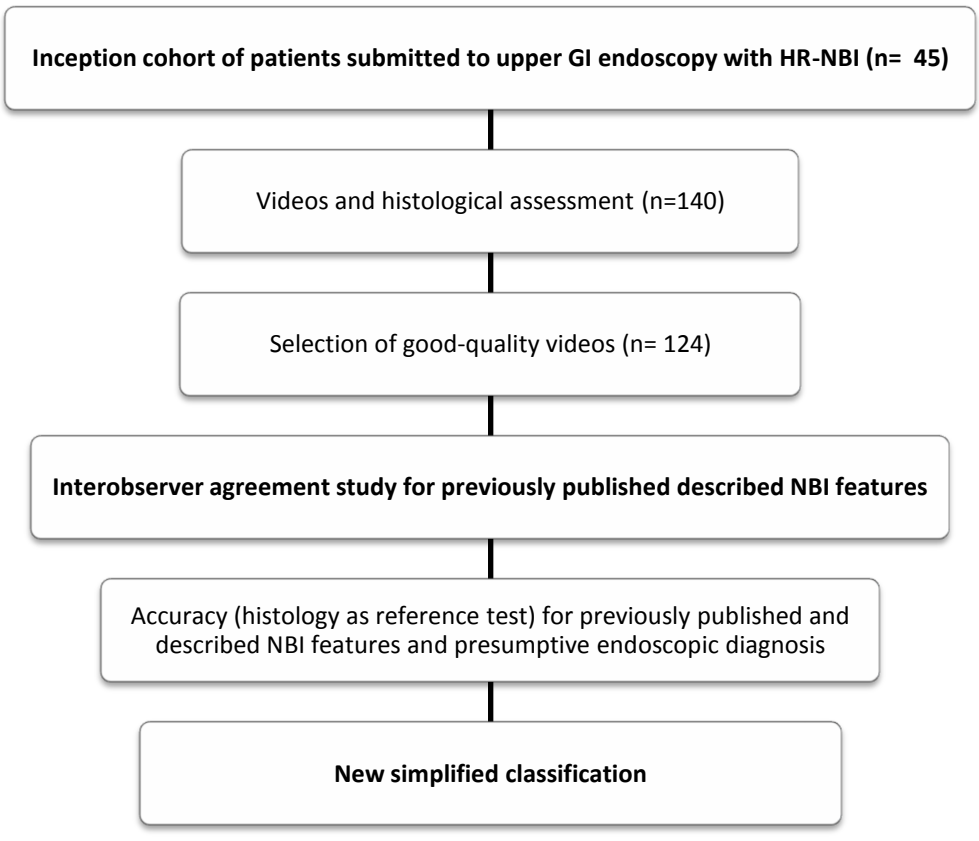
2. Materials and methods

Type of study and selection of patients

Patients submitted to upper gastrointestinal endoscopy observed at routine practice at two hospitals in the North of Portugal, were consecutively considered and included in this study after informed consent. Patients with chronic liver disease, anticoagulant therapy or coagulation disorders, psychiatric conditions were excluded. This study was approved by Ethical Committee of both Hospitals.

The study plan is presented in Figure 8. The first 45 patients submitted to upper GI endoscopy with HR-NBI constituted an inception cohort (n= 45) providing data for a reliability study of the previously described mucosal and vascular features using NBI in gastric mucosa (Chapter 4) and derivation of a simplified NBI classification, considering also validity measures (using histology as reference test).

Figure 8: Study plan flow diagram



Endoscopic procedures and selection of videos

Under pharyngeal anesthesia, all patients were submitted to upper GI endoscopy using a high-resolution Olympus® endoscope with NBI. Detailed observation of esophageal, gastric and duodenal mucosa was performed and all endoscopic lesions were described accordingly.

High-resolution videos of magnified (1.5x) NBI endoscopy were recorded for further analysis (reliability analysis) from, at least, five areas of antral, incisura and corpus mucosa (as for histology assessment according to Sydney-Houston); and from those areas with endoscopic changes, both at 'conventional' HR or at NBI (n= 140).

A shift between the ME-NBI and overview-white light modes was used to ensure position and precision of biopsies taken. Nevertheless, NBI description using previously reported mucosal and vascular features was performed by three endoscopists without the knowledge of histology results.

From these, 124 videos were cut for approximately 10 seconds and converted into AVI files. Each video was labeled with a random number and transferred into a computerized database. Only those of good mucosal morphology quality in which the subsequent video observation confirmed the targeting of the biopsies were selected. Quality of the videos was assessed by 2 experienced endoscopists.

All the potential spectrum of histological lesions was considered for videos selection: no intestinal metaplasia or dysplasia, presence of intestinal metaplasia, presence of dysplasia or carcinoma, presence of H. pylori irrespective of histology.

Histopathological procedures

All gastric mucosa specimens were obtained by endoscopic biopsies. Specimens were fixed in formalin, paraffin embedded, sectioned and stained with hematoxylin-eosin and reviewed by two pathologists blind to clinical and endoscopic observations.

Gastric neoplasias were classified as low (LGD) or high-grade intra-epithelial neoplasias or dysplasia (HGD), and invasive carcinoma according to Vienna classification (Schlemper, et al., 2000).

Lesions such as chronic gastritis, atrophy and intestinal metaplasia were considered as 'negative for dysplasia' but associated with gastric cancer, according to Vienna classification (Schlemper, et al., 2000) and Sydney-Houston classification (Dixon, et al., 1999). Chronic gastritis was considered as a chronic diffuse inflammatory infiltrate with lymphocytes and plasmocytes, expanding the *lamina propria* and epithelium, with no atypical cellular nuclei; atrophy as the disappearance of the normal glands in a certain area of the stomach; intestinal metaplasia classified as complete and incomplete. Paneth cells were considered as marker of complete IM. Gastric specimens were also evaluated for *H. pylori* infection using both modified Giemsa (2%) in hematoxicilin-eosin (H&E) biopsies stain.

Selection of raters

For reliability assessment of previous described NBI features, three endoscopists with previous clinical practice of NBI assessed all videos blinded to histology and for each other classification.

Variables

For reliability assessment, each video was classified by all the observers considering NBI features and grade of certainty. Also, each observer was asked to presume a histological diagnosis based on NBI features and a grade of certainty.

Histopathological assessment was considered the gold standard or reference test for accuracy estimates.

Statistical analysis

Statistical Package for Social Sciences (SPSS 9.0 Package Facility, SPSS Inc, IL, USA) was used for data support and analysis.

Cohen's kappa coefficient (k) and proportion of agreement (Pa) was calculated as measures for agreement between observers in the classification of endoscopic images. Strength of agreement was considered as follows: 0-0.2 slight, 0.2-0.4 fair, 0.4-0.6 moderate, 0.6-0.8 substantial, 0.8-1 almost perfect.

For sample size estimation we considered 3 groups of 3 observers and 3 classifications with 3 possible outcomes (gastric type mucosa, non-neoplastic SIM and neoplastic SIM). For a power of 80% and p value of 0.05, different possible sample size values to estimate differences higher than 0.2 in kappa values. This would represent a change in strength of agreement (for instance from substantial 0.6-0.8 to excellent 0.81-1.00). With 80 to 90 videos differences between kappa values of 0.4 and 0.578 or 0.568, respectively, would be observed; or between a kappa value of 0.6 and 0.767 or 0.758, respectively; or even between 0.8 and 0.932 or 0.925 respectively.

Each video classification was compared with the histological diagnosis of the corresponding specimens (gold standard or reference test). Sensitivity, specificity and predictive values were calculated. Global accuracy was estimated based on the proportion of true positive and true negative results. The level of agreement (CI %) and Kappa statistics were used to measure reliability in evaluation of mucosal and vascular patterns and underlying histological features.

3. Results

Description of participants and videos

The characteristics of patients and of the endoscopic procedures included in the study are shown in table 7. According to the specificity of the institution, a great number of endoscopic exams were done because of dysplasia or for mucosectomy (36%). This allowed a significant number of histological samples of dysplasia (16%) and metaplasia (30%). A total number of 124 videos were made from 140 histological samples.

Table 7: Characteristics of patients and endoscopic procedures included in the derivation study

Patients	n = 45
Gender (M/F)	28/17
Age	Mean: 61 Median (min-max): 60 (21-91)
Number of biopsies taken	Mean: 3 Median (min-max): 3 (1-5)
Indication for Upper GI	
Dyspepsia without known lesions	12
Follow up of metaplasia	6
Previously detected dysplasia without known lesions	11
Previously detected dysplasia for mucosectomy	5
Follow up of gastric mucosectomy	8
Other (e.g. GERD)	3
Main endoscopic findings	
Normal	13
Gastric superficial lesions	11
Papular-erythematous gastritis	8
Gastric scar	6
Gastric irregularity	5
Erosive gastritis	2
Histological diagnosis (n=140)	
Normal mucosa	76
Intestinal metaplasia antrum	27
Intestinal metaplasia corpus	15
Dysplasia or more severe	22
H. Pylori infection	68

Reliability of NBI features

In table 8 we present the reliability of the different NBI features evaluated by the 3 different observers. The identification of different mucosal and vascular patterns was associated with high reliability (k=0.82 and k=0.91, respectively). The identification of LBC or WOS was associated with substantial reliability (k=0.62 and k=0.61, respectively). On the other hand, other vascular features like density or thickness had a weak reliability.

Table 8: Reliability of the different NBI features evaluated

Videos	**Reliability**	
	Kappa	**Pc**
Mucosal pattern	0.82 (0.77-0.87)	Almost perfect
Light Blue Crest (LBC)	0.62 (0.54-0.70)	Substantial
White Opaque Substance (WOS)	0.61 (0.52-0.69)	Substantial
Vascular pattern	0.91 (0.88-0.94)	Almost perfect
Vascular thickness	0.10 (0.00-0.20)	Slight
Vascular density	0.23 (0.12-0.35)	Fair
Variable vascular density	0.42 (0.31-0.53)	Moderate
Histologic diagnosis	0.81(0.76-0.86)	Almost perfect
H. Pylori diagnosis	0.60 (0.59-0.71)	Moderate

Variable vascular density was the best parameter for identification of HP gastritis. However, it achieved only a moderate inter-observer agreement (k=0.42). The observers proposed a histological diagnosis with high agreement for histological diagnosis (k=0.81) and moderate agreement for HP infection (k=0.6). New mucosal and vascular patterns derived from histology results were highly valid for metaplasia and dysplasia (table 9).

Table 9: Sensitivity, specificity and validity of new mucosal and vascular patterns derived from histology

Mucosa	Vessels	Diagnosis	Sensitivity (CI 95%)	Specificity (CI 95%)	Validity (CI 95%)
Regular, circular	Regular	Normal	84(81-88)	81(77-84)	83(79-86)
Regular, villous or tubulous-villous,	Regular	Metaplasia	83(79-87)	89(86-92)	87(84-90)
Regular	Variable density	HP gastritis	71(67-76)	71(66-76)	71(66-76)
Irregular or absent	Irregular	Dysplasia	92 (89-95	99 (98-100)	97 (96-99)
Light blue crest		Metaplasia	57(52-61)	96(93-98)	86(83-89)

4. Discussion

This study was conducted in order to assess the inter-observer reliability of a classification derived from the systematic review previously presented (Chapter 4). We believe that the assessment of reliability of these mucosal and vascular features as observed by NBI endoscopy (Chapter 4) is indispensable to an effective clinical approach to this gastric endoscopic technique.

One of the limitations of this study is that reliability of the different NBI features was assessed through the evaluation of the data by only 3 observers. Therefore, a further validation study will be conducted, in order to evaluate this new NBI classification by assessing inter-observer reliability within a larger and more diverse group of endoscopists, in different clinical settings. This will allow us to assess the limitations of the classification and give indication of its clinical efficiency.

5. Conclusions

To the best of the author's knowledge, this is the first assessment of reliability using HR-NBI endoscopy for the identification of gastric lesions. Our results suggest the efficacy of this technique in the detection of gastric metaplasia and dysplasia. Irregularity of vascular/mucosal pattern is identified in a reproducible manner and it is consistently associated with gastric dysplasia.

If further validation studies confirm our results, HR-NBI endoscopy may establish itself as an important tool for early diagnosis and therapeutic procedures concerning gastric precancerous and cancer lesions.

6. FURTHER STUDIES

Reliability assessment influences our daily clinical decision making. Therefore, it is the author's objective to pursue scientific activity with significant impact comprising the assessment of reliability as a tool for the clarification of clinical concerns.

The author will continue to participate in the evaluation of diagnostic procedures in the gastric disease by participating in the validation study that should complete the derivation study described in chapter 5 of this thesis. This study will help clarify the role of NBI in gastric disease, especially in the diagnosis of gastric precancerous lesions and cancer.

When running a PubMed query of "Family Practice"[Mesh] AND "Reproducibility of Results"[Mesh] one can find less than 600 articles. The author's experience, acquired through this Master's degree and the production of this thesis, as well as his clinical background as a Family Medicine Resident creates an opportunity to develop the application of reliability assessment in Family Medicine.

The author considers that it is important for the development of Family Medicine the conduction of studies that assess reliability and validate various international questionnaires and other diagnostic tools that are applied in our national clinical setting. This may include tests that are used since many years, like the family APGAR test (Smilkstein, 1978) or the genogram (Coupland, et al., 2007), or other more recent diagnostic tools like the Patient Enablement Instrument (Lama, et al., 2010).

7. REFERENCES

Areia, M, Soares, M e Dinis-Ribeiro, M. 2010. Quality reporting of endoscopic diagnostic studies in gastrointestinal journals: where do we stand on the use of the STARD and CONSORT statements? *Endoscopy.* 2010, Vol. 42 (2), pp. 138-147.

Bansal, A, Ulusarac, O e Mathur, S et al. 2008. Correlation between narrow band imaging and nonneoplastic gastric pathology: a pilot feasibility trial. *Gastrointest Endosc.* 2008, Vol. 67, pp. 210-216.

Bartfay, E e Donner, A. 2000. The effect of collapsing multinominal data when assessing agreement. *International Journal of Epidemiology.* 2000, Vol. 29, pp. 1070-1075.

Black, RJ, Bray, F e Ferlay, J et al. 1997 . Cancer incidence and mortality in the uropean Union: cancer registry data and estimates of national incidence for 1990. *Eur J Cancer.* 1997 , Vol. 33 (7), pp. 1075-1107.

Brenner, H, Rothenbacher, D e Arndt, V. 2009. Epidemiology of stomach cancer. *Methods Mol Biol.* 2009, Vol. 472, pp. 467-477.

Cohen, S, Sweed, M e Edmonson, D. 2009. Tumours of the esophagus, gastroesophageal junction and stomach. *Seminars in Oncology Nursing.* 2009, Vol. 25, pp. 61-75.

Cohrssen, A, Anderson, M e Merrill, A et al. 2005. Reliability of the Whiff Test in Clinical Practice. *The Journal of the American Board of Family Practice.* 2005, Vol. 18, pp. 561-562.

Corley, D et al. 2009. Diagnosing Barrett's esophagus: reliability of clinical and pathologic diagnoses. *Gastrointest Endosc.* 2009, Vol. 69 (6), pp. 1004-1010.

Coupland, SK, Serovich, J e Glenn, J. 2007. Reliability in constructing genograms: a study among marriage and family therapy doctoral students. *Journal of Marital and Family Therapy.* 2007, Vol. 21, pp. 251-263.

Crew, K e Neugut, AI et al. 2006. Epidemiology of gastric cancer. *World J Gastroenterol.* 2006, Vol. 12 (3), pp. 354-362.

Curvers, WL e van den Broek, F et al. 2009. Systematic review of narrow-band imaging for the detection and differentiation of abnormalities in the esophagus and stomach. *Gastrointestinal Endoscopy.* 2009, Vols. 69, No. 2, pp. 307-317.

Dunn, G. 1992. Design and analysis of reliability studies. *Statistical Methods in Medical Research.* 1992, Vol. 1, p. 123.

Endo, Y, Nosho, K e Arimura, Y et al. 2005. Study of the tumor vessels in depressed-type early gastric cancers using narrow band imaging magnfying endoscopy and cDNA array analysis. *Dig Endosc.* 2005, Vol. 17, pp. 210-217.

Ezoe, Y, Muto, M e Horimatsu, T et al. 2010. Magnifying narrow-band imaging versus magnifying white-light imaging forthe differential diagnosis of gastric smal ldepressive lesions: a prospective study. *Gastrointestinal Endoscopy.* 2010, Vols. Volume71, No.3.

Ferlay, J, Autier, P e Boniol, M et al. 2007. Estimates of the cancer incidence and mortality in Europe in 2006. *Annals of Oncology.* 2007, Vol. 18, pp. 581–592.

Francisci, S, Capocaccia, R e Grande, E et al. 2009. The cure of cancer: A European perspective. *European Journal of Cancer.* 2009, Vol. 45, pp. 1067-1079 .

Gheorghe, C. 2006. Narrow-Band Imaging Endoscopy for Diagnosis of Malignant and Premalignant Gastrointestinal Lesions. *J Gastrointest Liver Dis.* 2006, Vol. 15 No.1, pp. 77-82.

Gilchrist, J. 2009. Weighted 2x2 kappa coefficients:recomended indices of diagnostic accuracy for evidence-based practice. *Journal of Clinical Epidemiology.* 2009.

Hamamoto, Y e Endo, T et al. 2004. Usefulness of Narrow Band Imaging endoscopy for diagnosis of Barrett's esophagus. *J Gastroenterology.* 2004, Vol. 39, pp. 14-20.

Hirschowitz, BI. 1961. Endoscopic examination of the stomach and duodenal cap with the fiberscope. *Lancet.* 1961, Vol. 1, pp. 1074-1078.

Hirschowitz, BI, Curtiss, LE e Peters, CW et al. 1958. Demonstration of a new gastroscope, the ‘‘fiberscope’’. *Gastroenterology.* 35, 1958, pp. 50–53.

Hopkins, HH e Kapany, N. 1954. A flexible fiberscope, using static scanning. *Nature.* 1954, Vol. 76, pp. 864–869.

Hripcsak, G e Heitjan, D. 2002. Measuring agreement in medical informatics reliability studies. *Journal of Biomedical Informatics.* 2002, Vol. 35, pp. 99-110.

Hundahl, SA, et al. 1997. The National Cancer Data Base report on gastric carcinoma. *Cancer.* 1997, Vol. 80 (12), pp. 2333-2341.

Irani, S e Thuer, I et al. 2009. Endoscopic narrow-band imaging-quantitative assessment of airway vascularity after lung transplantation. *J Biomed Opt.* 2009, Vol. 14 (1).

Kadowaki, S e Tanaka, K et al. 2009. Ease of early gastric cancer demarcation recognition: A comparison of four magnifying endoscopy methods. *Journal of Gastroenterology and Hepatology.* 2009, Vol. 24, pp. 1625–1630.

Kaise, M, Kato, M e al, Urashima M et. 2009. Magnifying endoscopy combined with narrow-bandimaging for differential diagnosis of superficial depressed gastric lesions. *Endoscopy.* 2009, Vol. 41, pp. 310–315.

Kato, M, Kaise, M e Yonezawa, J et al. 2009. Trimodal imaging endoscopy may improve diagnostic accuracy of early gastric neoplasia: a feasibility study. *Gastrointestinal Endoscopy.* 2009.

Kiesslich, R e Jung, M et al. 2002. Magnification endoscopy: does it improve mucosal surface analysis for the diagnosis of gastrointestinal neoplasias? *Endoscopy.* 2002, Vol. 34, pp. 819-822.

Kwon, RS e Sahani DV, et al. 2005. Gastrointestinal cancer imaging: deeper than the eye can see. *Gastroenterology.* 2005, Vol. 128, pp. 1538-1553.

Lachin, J. 2004. The role of measurement reliability in clinical trials. *Clinical Trials.* 2004, Vol. 1, p. 553.

Lama, C, Yuena, N e Mercerb, S et al. 2010. A pilot study on the validity and reliability of the Patient Enablement Instrument (PEI) in a Chinese population. *Family Practice.* Apr de 2010, p. [Epub ahead of print].

Nakayoshi, T, Tajiri, H e Matsuda, K et al. 2004. Magnifying endoscopy combined with narrow band imaging system for early gastric cancer: correlation of vascular pattern with histopathology. *Endoscopy.* 2004, Vol. 36, pp. 1080-1084.

Ottenbacher, K, Msall, M e Lyon, N et al. 1997. Interrater Agreement and Stability of the Functional Independence Measure for Children (WeeFIMTM): Use in Children

with Development Disabilities. *Archives of Physical Medicine and Rehabilitation.* 1997, Vol. 78, pp. 1309-1315.

Paris Workshop on Columnar Metaplasia in the Esophagus and the Esophagogastric Junction. **Workshop, Paris. 2005.** Paris, France : Endoscopy, 2005. Vol. 37, pp. 879-920.

Petrie, A e Sabin, C. 2000. *Medical Statistics at a Glance.* s.l. : Blackwell Science, 2000.

Pinheiro, PS e Tyczyński, JE, Bray F et al. 2003. Cancer incidence and mortality in Portugal. *Eur J Cancer.* 2003, Vol. 39 (17), pp. 2507-2520.

Quan, H e WJ, Shih. 1996. Assessing reproducibility by the within-subject coefficient of variation with random effects models. *Biometrics.* 1996, Vol. 52, pp. 1195-1203.

Quiros, RM e Bui, CL et al. 2009. Multidisciplinary Approach to Esophageal and Gastric Cancer. *Surg Clin N Am.* 2009, Vol. 89, pp. 79–96.

Saameño, JA, Sánchez, AD e Castillo, J et al. 1996. Validity and reliability of the family Apgar family function test. *Aten Primaria.* Oct de 1996, Vol. 18(6), pp. 2892-96.

Schlemper, RJ, Riddell, RH e Kato, Y et al. 2000. The Vienna classification of gastrointestinal epithelial neoplasia. *Gut.* 2000, Vol. 47 (2), pp. 251-255.

Schmid, RM. 2008. The Future for Endoscopy Is Bright. *Gastroenterology.* 2008, Vol. 135, pp. 1032–1034.

Sharma, P, Bansal, A e Mathur, S et al. 2006. The utility of a novel narrow band imaging endoscopy system in patients with Barrett's esophagus. *Gastrointest Endosc.* 2006, Vol. 64 (2), pp. 167-175.

Sharma, P, Weston, AP e Topalovski, M et al. 2003. Magnification chromoendoscopy for the detection of intestinal metaplasia and dysplasia in Barret's oesophagus. *Gut.* 2003, Vol. 52, pp. 24-27.

Shoukri, M. 2004. *Measures of inter-observer agreement.* Florida : Chapman & Hall/CRC Press, 2004.

Shoukri, M, Asyali, M e Donner, A. 2004. Sample size requirements for the design of reliability study: review and new results. *Statistical Methods in Medical Research.* 2004, Vol. 13, p. 251.

Shoukri, M, Elkum, N e Walter, S. 2006. Interval Estimation and optimal design for the within-subject coefficient of variation for continuos and binary variables. 2006, Vol. 6, pp. 6-24.

Sim, J e Wright, C. 2005. The kappa Statistics in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical Therapy Journal.* 2005, Vol. 85, pp. 257-268.

Sivak, MV e Fleischer, D. 1983. Colonoscopy with a videoendoscope: Preliminary experience with a new type of endoscope. *Gastrointest Endosc.* 29, 1983, p. 187.

Sivak, MV. 2006. Gastrointestinal endoscopy: past and future. *Gut.* 55, 2006, pp. 1061–1064.

Smilkstein, G. 1978. The Family APGAR: A proposal for family function test and its use by physicians. *Journal of Family Practice.* 1978, Vol. 6(6), pp. 1231-1239.

Smilkstein, G, Ashworth, C e Montano, D. 1982. Validity and reliability of the family APGAR as a test of family function. *J Fam Pract.* Aug de 1982, Vol. 15(2), pp. 303-311.

Stemmermann, GN e Fenoglio-Preiser, C. 2002. Gastric carcinoma distal to the cardia: a review of the epidemiological pathology of the precusors to a preventable cancer. *Pathology.* 2002, Vol. 34 (6), pp. 494-503.

Tahara, T, Shibata, T e M, Nakamura. 2009. Gastric mucosal pattern by using magnifying narrow-band maging endoscopy clearly distinguishe shistologica land serological severity of chronic gastritis. *Gastrointestinal Endoscopy.* 2009, Vols. Volume70, No.2.

Tamai, N, Kaise, M e Nakayoshi, T et al. 2006. Clinical and endoscopical characterization of depressed gastric adenoma. *Endoscopy.* 2006, Vol. 38 (4), pp. 391-394.

Telleman, H e Burger, T et al. 2009. Evolution of gastroenterology training. *World J Gastroenterol.* 2009, Vol. 15, pp. 1793-1798.

The Paris endoscopic classification of superficial neoplastic lesions: esophagus, stomach, and colon. **2003.** s.l. : Gastrointest Endosc, 2003. Vol. 58 (Suppl 6), pp. S3-S43.

Uebersax, J. Statistical Methods for Rater Agreement. [Online] [Citação: 12 de Maio de 2009.] http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm.

Uedo, N e Ishihara, R et al. 2006. A new method of diagnosing gastric intestinal metaplasia: narrow-band imaging with magnifying endoscopy. *Endoscopy.* 2006, Vol. 38 (8), pp. 819-824.

van Rijn, JC, Reitsma, JB e Stoker, J et al. 2006. Polyp miss rate determined by tandem colonoscopy: a systematic review. *Am J Gastroenterol.* 2006, Vol. 101, pp. 343-50.

van Sandick, JW, van Lanschot, JJ e Kuiken, BW et al. 1998. Impact of endoscopic biopsy surveillance of Barrett's oesophagus on pathological stage and clinical outcome of Barrett's carcinoma. *Gut.* 1998, Vol. 43, pp. 216-222.

Verdecchia, A, Corazziari, I e Gatta, G et al. 2004. Explaining gastric cancer survival differences among European countries. *Int J Cancer.* 2004, Vol. 109, pp. 737-741.

Walter, DS, Eliasziw, M e Donner, A. 1998. Sample size and optimal design for reliability studies. *Statistics in Medicine.* 1998, Vol. 17, pp. 101-110.

Whiting, JL, Sigurdsson, A e Rowlands, DC et al. 2002. The long term results of endoscopic surveillance of premalignant gastric lesions. 2002, Vol. 50, pp. 378-381.

WHO. 2001. IARC Unit of Descriptive Epidemiology: WHO cancer mortality databank. 2001.

Williamson, J, Manatunga, A e Lipsitz, S. 2000. Modeling kappa for measuring dependent categorical agreement data. *Biostatistics.* 2000, pp. 191-202.

Yalamarthi, S, Witherspoon, P e McCole, D et al. 2004. Missed diagnoses in patients with upper gastrointestinal cancers. *Endoscopy.* 2004, Vol. 36, pp. 874-9.

Yao, K, Iwashita, A e Tanabe, H et al. 2008. White opaque substance within superficial elevated gastric neoplasia as visualized by magnification endoscopy with narrow-band imaging: a new optical sign for differentiating betwee nadenoma and carcinoma. *Gastrointestinal Endoscopy.* 2008, Vol. 68 (3).

Yao, K, Oishi, T e Matsui, T et al. 2002. Novel magnified endoscopic findings of microvascular architecture in intramucosal gastric cancer. *Gastrointest Endosc.* 2002, Vol. 56, pp. 279-284.

Zera, RT, Nava, HR e Fischer, JI. 1993. Percutaneous endoscopic gastrostomy (PEG) in cancer patients. *Surg Endosc.* 1993, Vol. 7 (4), pp. 304-307.