
# The diagnostic application of new technologies: A critical evaluation of current research

## La aplicación diagnóstica de nuevas tecnologías: Una valoración crítica del estado actual de la investigación

Doctoral Thesis

**Lucy Anne Parker**

Doctorado Europeo de la Universidad Miguel Hernández

Departamento de Salud Pública, Historia de la Ciencia y Ginecología

Universidad Miguel Hernández de Elche

Alicante, 2011

Director / Directora: Drª Blanca Lumbreras Lacarra

UNIVERSITAS
Miguel
Hernández

D. ENRIQUE PERDIGUERO GIL, Director del Departamento de Salud Publica, Historia de la Ciencia y Ginecologia de la Universidad Miguel Hernandez de Elche,

CERTIFICA:

Que la presente memoria titulada 'La aplicación diagnóstica de nuevas tecnologías: Una valoración crítica del estado actual de la investigación', presentada por Doña Lucy Anne Parker, para optar al grado de Doctor, ha sido realizada en este Departamento bajo la dirección de la Doctora Blanca Lumbreras Lacarra.

Sant Joan d'Alacant, marzo de 2011

Dr. D. Enrique Perdiguero Gil

UNIVERSITAS
Miguel
Hernández

La Doctora Dna. Blanca Lumbreras Lacarra, del Departamento de Salud Publica, Historia de la Ciencia y Ginecologia de la Universidad Miguel

Hernandez de Elche,

CERTIFICA:

Que la presente memoria titulada 'La aplicación diagnóstica de nuevas tecnologías: Una valoración crítica del estado actual de la investigación', presentada por Doña Lucy Anne Parker, para optar al grado de Doctor, ha sido realizada en este Departamento bajo su dirección.

Sant Joan d'Alacant, marzo de 2011

La Directora

Blanca Lumbreras Lacarra

UNIVERSITAS
Miguel
Hernández

Me gustaría expresar mi más sincero agradecimiento a todas las personas e instituciones que estuvieron implicadas en este trabajo, especialmente:

- A mi directora de tesis, Blanca Lumbreras, por su gran dedicación, excelente orientación y porque siempre me mostró su apoyo y amistad.

- A Ildefonso Hernández Aguado, por la inestimable oportunidad que me brindó para iniciarme en el camino de la investigación en Salud Pública. Su apoyo y entusiasmo han sido fundamentales.

- A los coautores de los artículos que componen esta tesis, por su colaboración y contribución científica, particularmente a Miquel Porta cuyo pensamiento crítico y numerosas aportaciones han sido claves en mi trabajo.

- A Noemi Gomez por su amistad, su constante apoyo, y porque siempre estuvo disponible para resolver mis dudas con el castellano.

- A Joaquín García Aldeguer, por su imprescindible ayuda administrativa.

- Finally, I'd like to thank my family for their support; and Manuel for his optimism and endless encouragement.

UNIVERSITAS
Miguel
Hernández

# INDEX

BOXES

# ÍNDICE

TABLAS

# PART 1

## INTRODUCTION

In recent years there have been remarkable advances in molecular technologies. Expectations are high for the development of non-invasive molecular diagnostic tests. A large number of research reports are currently published in this field yet few of the many tests proposed have been introduced to clinical practice with clearly defined benefits (Check E, 2010; Diamandis EP, 2007; Ransohoff DF, 2010). Offering guidance for the introduction of a new diagnostic test into clinical practice is complicated and given the intense promotion of molecular diagnostics, it is of the utmost importance that application of these new technologies to clinical practice is based in the best available evidence (Ransohoff DF, 2007; Porta M et al, 2007).

Part 1.1 of this doctoral thesis offers an overview of diagnostic research strategies, focusing on the evidence based transfer of diagnostic research to clinical practice. There will be a brief introduction of the sources of variation in studies of diagnostic accuracy, followed by an overview of the tools currently available to promote transparent reporting and methodological vigour in diagnostic accuracy research reports. I will discuss proposals for a formal process that involves distinct phases of research in order to guide the development and validation of a new diagnostic test. In part 1.2, I will describe the new technologies that are the topic of this doctoral thesis, and some of the challenges that have been observed in the translation of diagnostic tests based in these techniques into clinically useful tools.

The term diagnostic test is used to refer to any new procedure, marker, or other evaluation that provides new information used to establish the presence or absence of a certain disease or condition. This may be in a routine clinical setting or in a public health situation for population screening. Tests which are carried out to determine a

patient's prognosis or to determine which patients might benefit from a particular therapy are not considered. This decision was made because the research methodology involved in the validation of these latter tests involves the follow-up of patients and is therefore different from diagnostic accuracy studies.

## 1.1 The evidence based provision of diagnostic tests in clinical practice

### 1.1.1 Diagnostic accuracy studies

In a diagnostic accuracy study, the new diagnostic test (referred to as the index test) is compared to the best available method for establishing the presence or absence of the disease or condition of interest (referred to as the reference standard). Diagnostic accuracy refers to how the index test classifies the condition of interest compared to the reference standard and may be expressed in several ways –such as likelihood ratios, diagnostic odds ratios, or the area under the ROC curve (Feinstein AR, 1985; Florkowski CM, 2008). The most common way to express diagnostic accuracy is in terms of sensitivity and specificity, that is the proportion of individuals with the disease or characteristic of interest (according to the reference standard) that test positive with the index test - sensitivity, or the proportion of individuals without the disease or character of interest that test negative with the index test - specificity. Although it is desirable for a new diagnostic test to have high sensitivity and specificity, this is not the only occasion in which a new test may prove to be clinically useful. A highly specific test may be useful to rule in a disease, whereas a test with high sensitivity may be useful to rule out a disease.

### 1.1.2 Sources of error in diagnostic accuracy studies

There are a number of possible causes for variation in diagnostic accuracy. Firstly, measures of diagnostic accuracy may vary from study to study due to chance or random error. In this case, error can be minimized by increasing the study size and can be estimated with confidence intervals and statistical tests. Unfortunately many diagnostic tests are validated in a small sample and rarely report the confidence intervals for sensitivity and specificity (Bachmann L et al, 2006). In addition, exaggerated or biased estimates of diagnostic accuracy may be attributed to the design of the study, potentially

limiting the internal and external validity of the findings (Knotternus JA, 2009; Rutjes AWS et al, 2006; Whiting P et al. 2004). These may be linked to 1) the patients involved in the study, 2) the implementation of the index and reference tests, or 3) the interpretation and analysis of the test results.

*1) The patient population*: Differences in the clinical or demographic characteristics of the patients included in a diagnostic accuracy study may have considerable influence on the diagnostic accuracy achieved (Feinstein AR, 1985; Leeflang MM et al, 2009). It is therefore important that the spectrum of patients included in the study reflect the spectrum of patients who would receive the test in practice (Ransohoff DF et al, 1978; Mulherin SA et al, 2002). For example, it would be unsuitable to use patients with established disease to estimate the diagnostic accuracy of a test designed for use in a screening setting, where the individuals that would receive the test would be asymptomatic healthy individuals. Systematic differences between the test patients and the target population is referred to as spectrum bias, and can limit the external validity of the findings.

*2) The implementation of the index and reference tests:* The choice of an appropriate reference standard is fundamental when performing a valid diagnostic accuracy study. As discussed, the reference standard should be the best available method for determining the presence or absence of the disease or condition of interest. It may be a single test, or a combination of procedures including clinical follow-up. Given that diagnostic accuracy is expressed in terms of how the index test compared with the reference test, a reference standard that incorrectly classifies the target condition will produce misleading measurements of diagnostic accuracy (Feinstein AR, 1985). Furthermore, biased estimates of diagnostic accuracy may be obtained if the index test forms part of reference standard (incorporation bias) or is only carried out in a sub-group of patients (partial verification bias or work-up bias). Similarly, differential verification bias may arise when the disease is only confirmed in patients who test positive with the index test (Whiting P et al, 2004). These latter problems may be fairly common when the reference test involves invasive techniques which would be unethical or risky to perform without clinical indication. Disease progression bias may occur if there is a substantial time difference between performance of the index test and the

reference test, and if during this time period some patients have experienced a spontaneous recovery or have progressed to a more advanced disease.

*3) The interpretation and analysis of the test results:* Diagnostic accuracy may be overestimated if the investigator reading or interpreting the results of the index test is aware of the results of the reference standard, or vice versa (Lijmer JG et al, 1999). This situation is referred to as reviewer bias and can be avoided if interpretation of each test is carried out without knowledge of the results of the other, in a similar fashion to blinding in clinical trials. Finally, there is evidence that diagnostic accuracy may be biased if researchers exclude intermediate or unclear test results from their analysis. Bias occurs when intermediate or unclear results do not occur randomly in the study population but rather are correlated with disease status. Furthermore, it is essential to consider intermediate or unclear test results in the evaluation of the cost effectiveness of a new diagnostic procedure.

### 1.1.3 Guidelines for reporting diagnostic accuracy studies

As discussed, methodological shortfalls may lead to bias and cause misleading or erroneous estimates of diagnostic accuracy. Unfortunately diagnostic research is generally of poor quality when compared to therapeutic research (Reid MC et al, 1995; Smidt N et al, 2005; Lumbreras-Lacarra B et al, 2004). In order to be able to effectively evaluate the potential for bias, evidence based decision makers must rely on the transparent reporting of the study methods including the strategy used to recruit patients. There have been a number of publications set out to provide researchers with a list of essential elements to include in their research reports in order to ensure that readers are able to effectively judge the usefulness of the data and the context where the conclusions apply (McShane LM et al, 2005; von Elm E et al, 2007; Schulz KF et al, 2010). In diagnostic research, the first significant development in test reporting was the popularisation of Reid's seven methodological standards (Reid MC et al, 1995). Reid's standards paved the way for the development of today's widely accepted STARD (Standards for Reporting of Diagnostic Accuracy) initiative (Bossuyt PM et al, 2003; Bossuyt PM et al, 2003). Developed by a group of scientists and editors, the STARD statement is comprised of a list of 25 items and a flow diagram which can be used to

ensure that all essential elements are reported in the research report, therefore allowing the transparent assessment of potential threats to the validity of the study.

### 1.1.4 Guidelines for assessing the quality of diagnostic accuracy studies

Aside from initiatives to encourage transparent reporting, a number of journals or research groups have made proposals for the quality appraisal of published research reports (Guyatt GH et al, 1993). For diagnostic research the most significant of these is the QUADAS (Quality Assessment of Diagnostic Accuracy Studies) guideline, which is a tool for evaluating the quality of the diagnostic research reports included in systematic reviews or meta-analyses (Whiting P et al, 2003). It includes 14 items in the form of questions which refer to the numerous biases that may threaten the validity of diagnostic research and may help identify the potential causes of heterogeneity in the diagnostic accuracy estimates reported by studies included in systematic reviews. Both STARD and QUADAS have been integrated into the requirements for many of the major biomedical research journals, and have made a considerable impact in promoting evidence based diagnosis.

### 1.1.5 Interpretation of clinical applicability of diagnostic accuracy studies

It is possible that a tendency to overinterpret or exaggerate preliminary results as providing conclusive evidence for clinical applicability hinders the evidence based transition of new diagnostic procedures. Conceptual and methodological requirements for the validation of a new diagnostic test have been described. Nevertheless, there is still no widely applied formal structure to guide the introduction of a new test into practice. While proposals like STARD and QUADAS do help readers, physicians and other decision makers identify methodological weaknesses in studies which could potentially bias results, they do not help them identify misleading claims and interpretations made by the authors of methodologically sound studies (Segal JZ, 1993; Montori VM, 2004). For example, a well performed study carried out in a limited number of clinically relevant patients should not claim to provide conclusive evidence of clinical utility.

### 1.1.6 Research phases in the validation of a new diagnostic test

Alvan R Feinstein first proposed that the development of a new diagnostic test should follow sequential phases of research in a similar vein to clinical trials in therapeutic research (Feinstein AR, 1985). He described four phases of research culminating in the analysis of test utility in a large consecutive series of suitable patients. Similarly, David L. Sackett and Brian Haynes described four stages of test development each carried out to provide the answer to a different clinical question and in which the ultimate phase requires the demonstration of a clinical benefit in the patients undergoing the new diagnostic procedure (Sackett DL et al, 2002). Furthermore, Margaret S. Pepe et al. proposed a formal structure for the development of a clinical biomarker for population screening. It differs slightly from the others because by referring to the development of a screening tool, it includes a phase dedicated to ascertaining the utility of the biomarker for detecting the condition in a pre-clinical phase (Pepe MS et al, 2001). Despite variations in numbering or the phases included, the proposals agree that initial studies may use a case-control design and evaluate the discrimination between healthy controls and known disease cases, or diseased tissues and healthy adjacent tissues. Nevertheless, excellent results in these studies do not provide definitive evidence of clinical utility, because in real practice, there is usually a wider spectrum of disease than in a case-control study. The phases are ordered by the strength of evidence that each provides in favour of the test and, in order to achieve clinical validation, the test must be evaluated in a population that is similar to that in which the test is intended to be used eventually. Box 1 provides an overview of the different study designs used in the validation of a diagnostic test.

In this doctoral thesis randomised trials or cost-effectiveness studies will not be considered. Instead, I will focus on studies which constitute the validation of a new diagnostic or screening procedure by determining the diagnostic accuracy (the first two bullet points in box 1). The reason for this is that the thesis involves evaluating the current diagnostic research on new technologies and few of the proposed tests have reached a suitable level of validation in which it would be necessary to evaluate clinical or cost effectiveness.

**Box 1: Different study designs involved in the validation of a new diagnostic test**

- Initial studies should demonstrate that the test is able to distinguish individuals with the disease under question from those without. A **case-control** design may be used to discriminate between patients with overt disease and healthy individuals, or diseased tissue and healthy tissue. The next phases also use this design but involve an increased patient spectrum, (for example, individuals with competing diagnoses, diverse co-morbidities or varying levels of disease severity) or evaluate changes in diagnostic accuracy according to particular patient characteristics.
- In the following stages, the test is evaluated in a **prospective series** of individuals that reflect, with the maximum degree of fidelity, the clinical or public health setting where the test would be used. In a diagnostic setting, all patients would be symptomatic and it would be clinically reasonable to suspect that they have the disease in question. In a population screening setting, this would be a consecutive series of the target population and the main aim would be to estimate the false referral rate (i.e., healthy persons who test positive with the screening test and are referred for diagnostic work-up but are not finally diagnosed with the disease).
- The final stages of test validation would be an evaluation of clinical effectiveness and cost effectiveness. For this purpose, a **randomised trial** should be carried out to establish if patients who undergo the new diagnostic procedure actually fare better in their ultimate health outcome compared to those who receive the existing diagnostic procedures. In a population screening setting, it is necessary to establish if introduction of the screening programme actually leads to improved indicators of morbidity or mortality in the disease under question. Finally, in both settings, once benefits have been clearly described it should be established if the cost is acceptable.

## 1.2. New technologies

Underlying cellular and molecular changes involved in some disease processes may provide new opportunities for diagnosis. For example, molecular mutations preceding the onset of clinically detectable cancer have shown considerable potential for early diagnosis (Negm RS et al, 2002). Molecular diagnostic tests could be an attractive alternative to tissue pathology because they may be carried out in samples obtained non-invasively, such as plasma, serum, or urine, and thus could avoid the patient discomfort involved in obtaining biopsied tissue samples. While it has been possible to study individual genes or specific loci for some years, the completion of the sequencing of the human genome, has made it possible to study the genome as a unified whole - genomics. This and other technological advances in the past 20 years have spurred the '-omics' revolution, in which by adding the suffix '-omics', we can refer to the study of almost any cellular constituent as a unified whole. For example, transcriptomics refers to analysis of total mRNA expression and proteomics refers to the analysis of the proteome, the total protein content (Nature briefing, 1999; Hanash SM et al, 2002; Wild C et al, 2009). It is proposed that these high throughput technologies coupled with computer assisted discrimination systems may hold the future of clinical diagnosis, thus leading to diagnostic tests based on multi-marker patterns or biomarker profiles, rather than on single alterations.

### 1.2.1 Clarification of terms

In this doctoral thesis, I will describe the development of a tool for evaluating the quality of '-omics' technologies. Two of the articles presented in the results section (Parts 4.1 – 4.2) refer specifically to '-omics' technologies. In the final paper (Part 4.3), the evaluation is not limited to '-omics' technologies, but rather we refer more generally to molecular diagnostic research and tests based in molecular techniques. Box 2 provides some clarification of what we are referring to when using the term *'-omics' technologies*, or the term *molecular techniques*.

**Box 2: Definitions and descriptions**

- **'-Omics' technologies:** Technologies that permit large-scale parallel measurements for the comprehensive analysis of the complete, or near-complete, cellular specific constituents, such as RNAs, DNAs, proteins, or intermediary metabolites. Common techniques include microarray chips allowing the analysis of up to 80,000 genes at a time, or surface-enhanced laser desorption ionisation time-of-flight mass spectrometry (SELDI TOF MS), which is a high-throughput tool for detecting the masses of differentially expressed proteins.
- **Molecular techniques:** All techniques involved in the characterization, isolation, and manipulation of the molecular components of cells and organisms. Techniques include in situ hybridization of chromosomes for cytogenetic analysis, identification of pathogenic organisms by analysis of species-specific DNA sequences, the detection of mutations with polymerase chain reaction, the analysis of DNA methylation or other epigenetic modifications, as well as '-omics' technologies.

### 1.2.2 Problems with translation from discovery to clinical practice.

In 2002, proteomic spectra patterns in patients with ovarian cancer were shown to completely segregate patients with cancer from those without (Petrecoin EF et al, 2002). The resulting blood test appeared nearly 100% sensitive and specific for the detection of ovarian cancer. Although commercial laboratories were quick to plan the development and marketing of the new test, OvaCheck® doubts surfaced regarding its reliability and reproducibility and to date, it remains without approval from the U.S. food and drug administration (Ransohoff DF, 2005; Wagner L, 2006; Correlogic Systems, Inc, 2010). Researchers have criticized the approach and suggested that the apparent discrimination observed was in fact due to systematic differences in the experimental procedures used for the cases of ovarian cancer and the controls, or simply due to chance (Baggerly KA

et al, 2005; Ransohoff DF, 2005). Despite these setbacks, in June of 2010, OvaCheck® cleared the regulatory requirements for distribution and sale in the European Union (Correlogic Systems, Inc, 2010). The differing views of regulatory bodies perhaps highlight how challenging it is to offer guidance on the adoption of diagnostic tests based in new technologies in clinical practice.

In the next section, I will outline some of the challenges which may complicate the transition from discovery to clinical translation of molecular diagnostic research. While there are some issues fairly specific to molecular or '-omics' based tests, many challenges are common to all diagnostic research: issues such as reproducibility and bias must always be appropriately considered. Nevertheless, certain issues may carry more weight because diagnostic tests based in molecular techniques may be even more susceptible to bias than traditional diagnostic tests if they rely on biologically unstable material like RNA, or biomarker profiles that are sensitive to changes in temperature like serum proteins. Furthermore, given that the procedures involved in new molecular tests may be more complex than in traditional diagnostic tests, they may present additional opportunity for error and variation, and thus uncover new challenges and new biases. Although some of the errors or limitations mentioned in the next section are relevant for all diagnostic research, I will focus on the additional challenges posed by new technologies

### 1.2.3 Sources of error in molecular diagnostic research

*Biological variation:*

Researchers must consider the socio-demographic, clinical and physiological characteristics of the patients who have provided the biological specimens as the molecular constituents detected by the new tests may vary according to such characteristics. The serum protein profile may be influenced by factors such as stress or hormonal cycles. Aditionally, some physiological compounds in the blood such as cholesterol, immunoglobins, and testosterone can be subject to daily or seasonal variation and therefore collection of samples at different times of the day or year could influence the observed biomarker profile (Garde AH et al, 2000). Furthermore, it is important to consider any treatment or diagnostic procedures and how these may

influence the biomarker pattern. For example, surgical manipulation has been show to result in significant gene expression changes (Lin DS et al, 2006).

*Pre-analytical variation:*

Given the relative lack of stability of some of the molecular constituents detected by new tests, investigators must ensure that all samples are handled identically. Differences in specimen collection and management may influence the biomarker pattern and thus may introduce bias into the experiment. For example, changes in pre-analytical handling conditions such as tube or anticoagulant type, clotting time, transport time, storage conditions and temperature have been shown to affect serum proteins (Timms JF et al, 2007). Furthermore, experiments may be influenced by RNA degradation due to repetitive freezing cycles (Botling J et al, 2009).

*Analytical variation*:

Analytical variation refers to differences in how the experiments are carried out. It is an important concern for all diagnostic research, but molecular research is particularly sensitive to changes in the experimental protocol. Variation may be introduced when the experiments are performed by different labs, using different instruments, by different technicians and on different days, and bias will occur if these procedural differences are correlated with the disease of interest. For example, if serum samples from cancer cases are analysed on one day, and the control samples on a different day or by a different lab. Furthermore, the inability to cross-validate microarray results with studies carried out using materials from distinct manufacturers has been reported (Marshall E, 2004). While all diagnostic studies should consider the potential for analytical variation, the issue is vital in the validation of new technologies such as serum proteomics, because the resulting biomarker patterns are especially susceptible to change and variation.

*Data analysis*:

Finally, the reproducibility of '-omics' studies have been questioned, suggesting that in some cases the apparent discrimination is due to nothing more than chance. One report demonstrated the inability to replicate the pattern distinction models in two publicly available datasets which claimed to have found a proteomic pattern capable of segregating cases and controls (Baggerly KA et al, 2005). Bearing in mind '-omics' techniques may evaluate tens of thousands of parameters simultaneously; it is not

surprising that some parameters which appear to discriminate between the two diagnostic groups are actually false positives. This, and the tendency to develop or 'discover' the biomarker patterns using the available data, rather than having a predefined hypothesis as to which biomarkers are likely to be involved, make these studies susceptible to overfitting (i.e., the apparent discrimination is due to chance and results cannot be reproduced in other populations).

### 1.2.4 Current tools and guidelines for molecular research

There has been a recent surge in scientific production related to the evidence based use and interpretation of genetic association studies. A series of three articles were published in JAMA in order to serve as an introduction to clinicians wishing to read and critically appraise genetic association studies (Attia J et al, 2009; Attia J et al, 2009; Attia J et al, 2009). Furthermore, the STREGA (Strengthening the reporting of genetic association studies) guideline was published simultaneously in a number of high profile journals in 2009, in an effort to enhance the transparency of reporting of these studies (Little J et al, 2009). It is an extension of the STROBE statement (von Elm E et al, 2007) which is a guideline to promote the complete and transparent reporting of observational studies in order to enable the proper assessment of a study's strengths, weaknesses and generalisability. The STREGA guideline incorporates 12 new items relevant to genetic association studies.

There have been a number of initiatives focusing on the analytical characteristics of new molecular technologies. The MIAME (Minimum Information about a Microarray Experiment) guideline describes the minimum information necessary to enable the unambiguous interpretation of microarray experiments (Brazma A et al, 2001). It has spurred similar projects for the other new technologies such as the MIAPE (Minimum Information about a Proteomics Experiment) (Taylor CF et al, 2007) guideline or MISFISHIE (Minimum Information Specification for In Situ Hybridization and Immunohistochemistry Experiments) (Deutsch EW et al, 2008). Such initiatives are useful because they provide a comprehensive list of the analytical aspects that must be addressed in research reports using these technologies. Nevertheless, they do not address other aspects which would be vital for the diagnostic application of such technologies.

In fact, none of the above mentioned initiatives deal specifically with the diagnostic application of these new technologies. As previously discussed, there are numerous biases and limitations that must be taken into account for diagnostic research, and in addition, there appears to be a number of additional challenges posed by these new technologies which must not be overlooked. Generic diagnostic guidelines such as STARD or QUADAS are not presently suited for '-omics'-based diagnostic accuracy studies as they do not take into consideration the additional challenges presented by these new technologies, such as avoiding overfitting. A tool that addresses potential sources of bias specific to new technologies, as well as those relevant for all diagnostic research is called for.

UNIVERSITAS
Miguel
Hernández

# PART 2

## JUSTIFICATION AND HYPOTHESES

The introduction of a new diagnostic test into clinical practice does not follow the same rigorous process as the introduction of a new treatment or pharmaceutical. Nevertheless, inappropriate or premature application of diagnostic procedures may lead to incorrect clinical decisions, unnecessary patient discomfort, and adverse patient outcomes. Despite these important implications, diagnostic research remains poorly reported and diagnostic studies have been shown to be subject to methodological shortcomings biasing their results. In order to improve this situation and spur the evidence based application of new diagnostic procedures, academic groups have proposed guidelines for reporting, and for the quality appraisal of diagnostic accuracy studies. Furthermore, proposals have been made to set out distinct phases of research in the validation of a new diagnostic procedure.

In the past decade, diagnostic research has been confronted by the new challenges posed by technological advances in molecular biology and the advent of the '-omics' revolution. Reproducing the initial claims of diagnostic accuracy in this new field has proven to be even more complex and few of the proposed tests have made any impact on clinical decision making. Some of the challenges that face '-omics' based diagnostic tests are not adequately addressed by generic guidelines, such as the threat of overfitting. For this reason, existing proposals for improving the evidence based transition of diagnostic tests should be adapted in such a way that they take into account these particular aspects. Hence, in this doctoral thesis I will describe the development and validation of a tool called QUADOMICS, which is an adaptation of QUADAS to '-omics' based technologies.

Ensuring the timely and effective transfer of molecular diagnostic research results to clinical practice requires that researchers, editors, and physicians produce, publish and use only results coming from valid, reproducible research. Few '-omics' based

diagnostic tests have actually been introduced into clinical practice, despite extensive commercial support. As discussed, the apparent –but in fact artifactual– power to discriminate between diagnostic groups using '-omics' technologies may actually be due to differences in pre-analytical procedures, in clinical or physiological characteristics of the patients who provided the biological samples, or simply chance. With this in mind, it is proposed that the published research carried out in this area is of poor methodological quality and may be subject to numerous biases. Accordingly, I will explore the methodological quality of a sample of '-omics' based diagnostic accuracy studies and identify where methodological short-falls lie.

Finally, with respect to the 'hype' and commercial interest involved in the widespread dissemination of molecular diagnostic technology, it would be appropriate to evaluate how the authors of such studies interpret the clinical applicability of their findings. The tendency to exaggerate the clinical relevance of preliminary research findings is of particular importance for new technologies when, given the limited knowledge regarding potential limitations or bias, one should be cautious. It is proposed that some researchers evaluating the diagnostic accuracy of new molecular tests may not be sufficiently versed in issues related to study design and potential biases to diagnostic accuracy, and that they tend to overinterpretate or exagerate the clinical applicability of preliminary research findings. Consequently, the final part of this thesis includes an evaluation of how the authors of molecular diagnostic studies tend to interpret the clinical applicability of their research findings.

# PART 3

## OBJECTIVES

### 3.1 Overall objective:

The overall objective of this doctoral thesis is two-fold: Firstly, to develop and validate a tool for assessing the quality of molecular diagnostic studies, more specifically those based in'-omics' technologies; and secondly, to critically evaluate the current state of research on the diagnostic application of new molecular technologies.

### 3.2 Specific objectives:

- **Development of a tool for assessing the quality of diagnostic accuracy studies that use '-omics' technologies.**

    1. To produce a guideline for evaluating the methodological quality of diagnostic accuracy studies that use '-omics' technologies.
    2. To validate the developed guideline, through an evaluation of its applicability and consistency.

- **Evaluation of the current state of research on the diagnostic application of new molecular technologies.**

    3. To describe the methodological quality of a sample of diagnostic accuracy studies that use '-omics' technologies.
    4. To evaluate if the authors of molecular diagnostic accuracy studies make appropriate conclusions with regard to the clinical application of their test, considering the design and patient population used in the studies.

UNIVERSITAS
Miguel
Hernández

# PART 4

## METHODOLOGY

A detailed report of the methods can be found in the research articles which form the results section of this doctoral thesis (Parts 5.1 – 5.3). Additionally, a summary of the methodology can be found in the following pages in Spanish. Part 4.1 describes the methodology used to develop and validate a guideline for assessing the quality of diagnostic accuracy studies that use '-omics' technologies. Part 4.2 describes the methodology used for the evaluation of current research on the diagnostic application of new molecular technologies.



## METODOLOGÍA

Se puede encontrar un informe detallado de la metodología en los artículos que forman la sección de resultados de esta tesis doctoral (los apartados 5.1 a 5.3). Asimismo, en las siguientes páginas se encuentra un resumen de la metodología en español. En el apartado 4.1, se describe la metodología utilizada para desarrollar y validar una guía para evaluar la calidad de estudios de exactitud diagnóstica que utilizan tecnologías '-ómicas'. En el apartado 4.2, se describe la metodología utilizada para evaluar el estado actual de la investigación sobre la aplicación diagnóstica de las nuevas tecnologías moleculares.

UNIVERSITAS
Miguel
Hernández

## 4.1Desarrollo de una herramienta para evaluar la calidad metodológica de estudios de exactitud diagnóstica que utilizan tecnologías '-ómicas'

### 4.1.1 Resumen de los métodos utilizados para alcanzar el objetivo específico 1: Generación de la guía QUADOMICS.

QUADOMICS, la nueva guía, es una adaptación de QUADAS para su aplicación en estudios diagnósticos que evalúan tecnologías '-ómicas'. Por tanto, los procedimientos para el desarrollo de esta herramienta se basaron en los utilizados para la consecución de la guía QUADAS (Whiting P et al, 2003). El desarrollo se llevó a cabo mediante las siguientes etapas: 1) decisiones preliminares, 2) definición de las fases, 3) generación de los primeros ítems, 4) evaluación de los ítems seleccionados 5) diseño final de la guía.

*1) Decisiones preliminares:*

Se adoptaron las siguientes decisiones preliminares acerca del objetivo de la nueva guía y las situaciones donde se pretende aplicar: El objetivo principal de la guía es la evaluación de la calidad metodológica de la investigación diagnóstica basada en tecnologías '-ómicas'. Al igual que la guía QUADAS, evalúa los estudios incluidos en una revisión sistemática o metaanálisis. La guía se aplica a estudios de exactitud diagnostica para uso clínico o en programas de cribado.

Así mismo, se decidió que la guía incorporara las distintas fases del desarrollo de una nueva prueba diagnóstica. Con esta incorporación a la herramienta podremos evaluar las necesidades metodológicas de cada tipo de diseño utilizado.

*2) Definición de las fases:*

Se definieron cuatro fases del proceso de validación clínica de una nueva prueba diagnóstica, basándose en las propuestas de Feinstein, Sackett, Haynes, y Pepe (apartado 1.1.5). Las fases están ordenadas de acuerdo a la secuencia de la investigación seguida y están relacionadas con la fuerza de la evidencia que cada fase proporciona a la utilidad clínica de la prueba. En la primera fase, se utiliza la prueba para distinguir entre casos de enfermedad manifiesta y controles sanos, mientras la segunda incluye un espectro de pacientes más amplio tanto de casos como de controles. Puede, por ejemplo,

incluir casos con distinto grado de la enfermedad a estudio y una amplia gama de controles con sintomatología parecida a los casos. Se define una tercera fase opcional en la cual se detecta la probabilidad de resultados falsos positivos o falsos negativos, midiendo la exactitud diagnostica en ciertos subgrupos de pacientes relevantes. En la cuarta fase, la prueba se aplica a una serie de pacientes con las mismas características a aquella población donde se va a aplicar la prueba en la práctica real. Es decir, por ejemplo, para estudiar una nueva herramienta diagnostica clínica se estudia pacientes reclutados de manera consecutiva por sospecha clínica de la enfermedad en cuestión.

El paso previo a la aplicación de la nueva guía a un artículo es su asignación a una de estas cuatro fases.

*3) Generación de los primeros ítems:*

Se elaboró una lista de ítems potenciales. Dicha lista incluyó todos los ítems de QUADAS y varios nuevos referidos a las fuentes de error más frecuentes en el campo de las tecnologías '-ómicas, previamente identificados con la realización de una revisión sistemática (Lumbreras B et al, 2009). Se evaluó la aplicabilidad de cada ítem a la investigación '-ómica' y a las distintas fases de desarrollo de una nueva prueba diagnóstica. Algunos ítems fueron descartados, otros aplicables únicamente a estudios de ciertas fases, y para algunos ítems se modificó la descripción y explicación en QUADAS para mejorar su relevancia para investigación '-ómica'.

*4) Evaluación de los ítems seleccionados:*

Todos los miembros del equipo investigador aplicaron la guía a tres estudios de distintas fases como un breve piloto para resolver dificultades de su aplicación.

*5) Diseño final de la guía:*

Se elaboró una lista final de los ítems a incluir en QUADOMICS. Se modificó la descripción de los ítems cuando fue necesario y se especificó a qué fases de investigación se debía aplicar cada ítem.

### 4.1.2 Resumen de los métodos utilizados para alcanzar el objetivo especifico 2: validación de QUADOMICS, mediante una evaluación de su aplicabilidad y su consistencia.

Se evaluó la aplicabilidad y la consistencia de la nueva guía QUADOMICS, mediante su aplicación a una muestra de estudios primarios de investigación diagnóstica que utilizaban tecnologías '-ómicas'.

*Búsqueda bibliográfica y selección de estudios:*
Se identificaron artículos originales mediante una búsqueda sistemática en Medline con los términos MeSH "Genomics", "Sensitivity and specificity" y "Diagnosis". La búsqueda se limitó a los artículos publicados desde el 1 de enero de 2006 hasta el 17 de junio 2009 (la fecha de la búsqueda). Los títulos y resúmenes de todos los posibles artículos fueron revisados. Se seleccionaron los artículos en función de los siguientes criterios: artículos de investigación originales cuyo objetivo principal fue evaluar la exactitud diagnóstica de una prueba basada en tecnologías '-ómicas' para su uso en la práctica clínica o en un programa de cribado. Se seleccionaron únicamente estudios en lengua inglesa que presentaban una medida de exactitud diagnóstica (por ejemplo, sensibilidad y especificidad, el área bajo la curva ROC, el odds ratio diagnóstico, razones de verosimilitud) o que proporcionaban los datos suficientes para su cálculo.

*Síntesis de datos:*
Tres investigadores de manera independiente evaluaron la calidad metodológica de todos los artículos seleccionados a través de la aplicación de la guía QUADOMICS. Como herramienta de referencia, además de un ejemplar de QUADOMICS, a cada revisor se le proporcionó una copia de la publicación QUADOMICS (Lumbreras B et al, 2008), el desarrollo de la guía QUADAS (Whiting P et al, 2003) y el artículo de la evaluación de QUADAS en la cual se detallan algunas modificaciones a los ítems iníciales (Whiting PF et al, 2006). Los tres investigadores se reunieron para comparar sus observaciones y generar la clasificación de consenso. Se resolvieron los desacuerdos por discusión. Durante este proceso, se debatieron las dificultades con la aplicación de ciertos ítems, y se exploraron modos de mejorar la descripción de dichos ítems para facilitar la aplicación de la guía. Para evaluar la consistencia de QUADOMICS, se calculó el porcentaje de acuerdo entre la evaluación original de cada revisor y la

calificación de consenso, tanto en general como para cada ítem por separado. No se calculó el estadístico kappa de Cohen para el acuerdo entre evaluadores, ya que está fuertemente influenciado por la prevalencia de las características evaluadas y puede no reflejar la realidad (Lantz CA et al, 1996). La consistencia se consideró 'baja' cuando la concordancia con el consenso fue inferior al 60% para al menos un revisor, o si dos o más de los revisores tenía menos de un 80% de acuerdo con el consenso. Se evaluaron las razones para la baja consistencia y se trató de reformular el ítem cuando fue necesario.

## 4.2 Evaluación del estado actual de la investigación de nuevas tecnologías de diagnóstico molecular.

### 4.2.1 Resumen de los métodos utilizados para alcanzar el objetivo específico 3: Describir la calidad metodológica de una muestra de estudios de exactitud diagnóstica que utilizan tecnologías '-ómicas'.

Se describió la calidad metodológica de los estudios identificados para la evaluación de la aplicabilidad y consistencia de QUADOMICS (apartado 4.1.2). Para la descripción de cada ítem de QUADOMICS se utilizó la variable de consenso creada durante el proceso de validación.

*Estrategia de búsqueda, selección de estudios y síntesis de datos:*

Descrito en el apartado 4.1.2.

*Análisis de datos:*

Teniendo en cuenta que había estudios de distintos fases de validación, y que en la guía QUADOMICS hay ítems que solo se aplican a estudios de fase IV, se describió la calidad de cada artículo mediante el cálculo del porcentaje de cumplimiento de los ítems aplicados. Para identificar los déficits metodológicos más frecuentes, se calculó la proporción de estudios que cumplió cada criterio de calidad por separado. Todos los cálculos estadísticos se llevaron a cabo con Stata/SE 8.0 (StataCorp, College Station, TX, USA).

**4.2.2 Resumen de los métodos utilizados para alcanzar el objetivo específico 4: Evaluar si los autores de estudios de diagnóstico que aplican métodos moleculares hacen conclusiones pertinentes con respecto a la aplicación clínica de su prueba, teniendo en cuenta el diseño y la población de pacientes utilizados en los estudios**.

*Búsqueda bibliográfica y selección de estudios:*

Se identificaron estudios de diagnóstico basados en métodos moleculares mediante una búsqueda sistemática de Medline con los siguientes términos MeSH: "Diagnosis", "Genomics", "Microarray analysis", "Molecular diagnostic techniques", o "Sensitivity and Specificity"; y las siguientes palabras claves: "diagnos*", "genomics", "proteomics", "molecular", o "genetic", "diagnostic test". Se seleccionaron artículos originales publicados en el año 2006 que incluían humanos, en los cuales el objetivo principal fue evaluar el valor diagnóstico de una determinada prueba diagnóstica basada en técnicas moleculares. Un investigador llevó a cabo la selección de los estudios y para determinar la fiabilidad del proceso de selección, una muestra aleatoria de 200 resúmenes fue evaluada de forma independiente por dos investigadores más. El acuerdo con el revisor inicial fue de 94% y 83%, respectivamente.

*Extracción y síntesis de datos:*

Dos investigadores extrajeron de manera independiente los siguientes datos de cada artículo: factor de impacto de la revista, categoría de la revista; si los autores procedieron del ámbito de laboratorio o la clínica; enfermedad de estudio, metodología molecular utilizada y tamaño muestral. Se elaboró la variable resultado (sobreinterpretación de la aplicabilidad clínica) utilizando reglas pre-definidas que consideraban las conclusiones de los autores con respecto al uso clínico de la prueba, el diseño del estudio y la exactitud diagnóstica alcanzada. Cada estudio se asignó a uno de los siguientes tres posibles diseños:

1) Grupo 1: Estudio con controles sanos o con controles que tienen un diagnóstico alternativo.
2) Grupo 2: Estudio con una serie consecutiva de pacientes o con sujetos clínicamente relevantes.
3) Grupo 3: Otros estudios.

Se anotaron las declaraciones de los autores que aparecían en los artículos referentes a la aplicabilidad clínica de la prueba y si estimaba la necesidad de una evaluación clínica adicional. Se clasificó cada afirmación según el siguiente esquema:

- Con respecto a las declaraciones sobre aplicabilidad clínica se clasificó cada estudio como "definitivamente favorable", "prometedor" o "no favorable".
- Con respecto a las declaraciones sobre evaluación clínica adicional se clasificó cada estudio como "menciona la necesidad de evaluación clínica adicional" o "no menciona la necesidad de evaluación clínica adicional".

En los estudios de los grupos 1 ó 3, (tipo casos y controles u otros estudios) se definió la sobreinterpretación si los autores fueron definitivamente favorables con respecto a la aplicación clínica de la prueba (con o sin mencionar evaluación clínica adicional), o si fueron moderadamente favorables (declaraciones prometedoras) y no mencionaron la necesidad de evaluación clínica adicional.

En los estudios del grupo 2, (con una población clínicamente relevante) se definió la sobreinterpretación si tuvieron conclusiones definitivamente favorables y la exactitud diagnóstico de la prueba fue insuficiente.

*Análisis estadístico:*

Se calcularon odds ratios y sus intervalos de confianza del 95% mediante regresión logística no condicional. Se desarrollaron los modelos multivariables considerando todas las variables con $p < 0{,}10$ en análisis univariado y utilizando una selección por pasos hacia delante. Se incluyeron el diseño del estudio y el índice de exactitud diagnóstica, como factores de ajuste en el análisis multivariable, porque fueron los criterios para juzgar el sobreinterpretación, y también podría estar relacionado con otras características del estudio, así actuando como clásicos factores de confusión. El tamaño del estudio y el factor de impacto bibliográfico se clasificaron en cuartiles. Los análisis se realizaron con Stata /SE 8.0 (StataCorp, College Station, TX, EE.UU.).

UNIVERSITAS
Miguel
Hernández

# PART 5

## RESULTS

The results of this doctoral thesis are presented in three published articles. In the following pages, each article can be found in pdf format, along with a brief description of the main findings in Spanish.

## RESULTADOS

Los resultados de esta tesis doctoral se presentan en tres artículos publicados. Cada artículo se puede encontrar en formato pdf en las páginas siguientes, junto con una breve descripción de los principales hallazgos en español.

**Article 1:**
Lumbreras B, Porta M, Márquez S, Pollán M, Parker LA, Hernandez-Aguado I. QUADOMICS: An adaptation of the Quality Assessment of Diagnostic Accuracy Assessment (QUADAS) for the evaluation of the methodological quality of studies on the diagnostic accuracy of '-omics' based technology. Clin Biochem 2008;41:1316-25.

**Article 2:**
Parker LA, Gomez Saez N, Lumbreras B, Porta M, Hernández-Aguado I. Methodological Deficits in Diagnostic Research Using '-Omics' Technologies: Evaluation of the QUADOMICS Tool and Quality of Recently Published Studies. PLoS One 2010;5:e11419.

**Article 3:**
Lumbreras B, Parker LA, Porta M, Pollán M, Ioannidis JPA, Hernández-Aguado I. Overinterpretation of clinical applicability in molecular diagnostic research. Clin Chem 2009;55:786-94.

UNIVERSITAS
Miguel
Hernández

## 5.1 Resumen de los hallazgos principales del primer artículo

(Referencia: Lumbreras B, Porta M, Márquez S, Pollán M, Parker LA, Hernandez-Aguado I. QUADOMICS: An adaptation of the Quality Assessment of Diagnostic Accuracy Assessment (QUADAS) for the evaluation of the methodological quality of studies on the diagnostic accuracy of '-omics' based technology. Clin Biochem 2008;41:1316-25.)

**QUADOMICS: Una adaptación de QUADAS para la evaluación de la calidad metodológica de estudios sobre la exactitud diagnóstica de tecnologías basadas en '-ómicas'.**

En la nueva guía adaptada, QUADOMICS, permanecieron 12 de los 14 criterios de QUADAS. El ítem de QUADAS referente a la independencia entre la nueva prueba y el estándar de referencia fue eliminado ya que las tecnologías '-ómicas' son aún novedosas y no forman parte de ningún estándar de referencia. Asimismo el ítem de QUADAS que se refiere a pérdidas de pacientes durante el estudio fue descartado ya que este tema ha sido incorporado en la nueva descripción del ítem 1 (*¿Se describieron claramente los criterios de selección?*). Con respecto a esto, dos ítems fueron reformulados para clarificar algunos aspectos relevantes para la investigación '-omics'. La descripción del ítem 1 fue ampliada prestando atención al problema de sesgos de selección en estudios '-ómicas' derivado de la dificultad de conseguir muestras. La nueva descripción es más estricta y exige un diagrama de flujo que describe la selección de pacientes en todo caso.

Por otra parte, se adaptó la descripción del ítem 10 (*¿Se describió la ejecución de la prueba de estudio con suficiente detalle para permitir su replicación?*) ya que debido a la complejidad de las nuevas tecnologías '-ómicas' hay numerosos aspectos analíticos adicionales que se deben especificar en los métodos. En la nueva descripción se remite a los autores a guías analíticas como MIAME y MIAPE (Brazma A et al, 2001; Taylor CF et al, 2007).

El primer paso de la evaluación de la calidad metodológica de estudios sobre la exactitud diagnóstica de tecnologías '-ómicas' con QUADOMICS, es la asignación del estudio a uno de las cuatro fases de validación. Se definieron las fases con respecto al

diseño del estudio y el espectro de los pacientes incluidos, como se menciona en métodos sección 4.1.1, punto 2. Se decidió que el ítem 2 (*¿El espectro de pacientes era representativo de los pacientes que recibirán la prueba en la práctica?*) y el 14 (*¿La información clínica de la que se disponía cuando se interpretaron los resultados de la prueba, estará presente cuando se aplique la prueba en la práctica?*) debían ser aplicados únicamente a estudios de la última fase (fase IV). Estudios en las fases I a III son preliminares, y por su propio diseño tiene una población de estudio artificial que no cumple estos criterios. Lo importante es que el lector sea consciente de que estos estudios no proporcionan pruebas definitivas de eficaz clínica.

Por último se introdujeron cuatro nuevos ítems que trataron fuentes de errores comunes de investigación diagnóstica basada en tecnologías '-ómicas'. El ítem 3 (*¿Se describió el tipo de muestra de manera completa?*), es necesario porque los biomarcadores estudiados en investigación '-ómica' pueden adoptar distintos perfiles dependiendo del tipo de muestra usada, por ejemplo puede haber una mayor concentración de mutaciones genéticas en tejido que en suero. El ítem 5 (*¿Se describieron los tratamientos y procedimientos pre-analíticos con suficiente detalle y fueron similares para todas las muestras? Y, si se mencionaron diferencias, ¿se evaluó su efecto en los resultados?)* trata la posible influencia de la variación en los procedimientos pre-analíticos de las muestras. Por otra parte, el ítem 4 (*¿Se describieron los procedimientos y los tiempos para la recogida de las muestras biológicas con respecto a los factores clínicos con suficiente detalle?*) considera los pacientes que han proporcionado las muestras y como las diferencias en factores clínicos pueden influir en el perfil '-ómico' estudiado. Para ayudar en su aplicación se dividió este criterio en dos partes de la siguiente manera: ítem 4.1 se refiere a los factores clínicos y fisiológicos de los pacientes y el ítem 4.2 se refiere a los procedimientos diagnósticos o tratamientos que han recibido antes de la recogida de la muestra. El cuarto ítem nuevo es el ítem 16 trata que evalúa si los investigadores han tomado medidas para evitar el overfitting (*¿Es probable que se evitara la presencia de overfitting?*).

Una traducción de la guía QUADOMICS, presentada en la Sociedad Española de Epidemiología, Zaragoza 2009, se encuentra en el anexo 1.

Available online at www.sciencedirect.com
ScienceDirect
Clinical Biochemistry 41 (2008) 1316–1325
CLINICAL BIOCHEMISTRY

Review

# QUADOMICS: An adaptation of the Quality Assessment of Diagnostic Accuracy Assessment (QUADAS) for the evaluation of the methodological quality of studies on the diagnostic accuracy of '-omics'-based technologies

Blanca Lumbreras [a,e,*], Miquel Porta [b,e], Soledad Márquez [c], Marina Pollán [d,e], Lucy A. Parker [a,e], Ildefonso Hernández-Aguado [a,e]

[a] *Public Health Department, Miguel Hernández University, Alicante, Spain*
[b] *Institut Municipal d'Investigacio Medica, Universitat Autonoma de Barcelona, Spain*
[c] *Andalusian Agency for Health Technology Assessment, Seville, Spain*
[d] *Cancer and Environmental Epidemiology Area, National Centre for Epidemiology, Instituto de Salud Carlos III, Madrid, Spain*
[e] *CIBER en Epidemiología y Salud Pública (CIBERESP), Spain*

Received 23 April 2008; received in revised form 23 June 2008; accepted 25 June 2008
Available online 9 July 2008

**Abstract**

**Objectives:** To adapt the Quality Assessment of Diagnostic Accuracy Assessment (QUADAS) to the particular methodological challenges posed by research on '-omics'-based diagnostic tests.

**Design and methods:** We generated new guidelines by appraising the suitability of each criterion from QUADAS to '-omics'-based diagnostic research, and by adding new items that addressed specific sources of error. In addition, we defined four phases in the evaluation of a diagnostic test.

**Results:** Twelve of the 14 criteria from QUADAS were retained in the new tool. The items relating to selection criteria and the description of the test were reformulated, and the criteria about external validation and the availability of clinical data were applied only in studies in the last research phase. Four new items were incorporated to QUADOMICS related to pre-analytical conditions and methods to avoid overfitting.

**Conclusions:** QUADOMICS is an adaptation of QUADAS to the special nature of '-omics'-based diagnostic research. The tool adds new items that assess quality issues specific to this research, and may enhance the application of '-omics'-based discoveries to clinical \and public health practice.


*Keywords:* QUADAS; Genomics; Proteomics; Arrays; Diagnosis; Guidelines; Error; Variability

**Contents**



* Corresponding author. Fax: 965919551.
*E-mail address:* blumbreras@umh.es (B. Lumbreras).

0009-9120/$ - see front matter © 2008 The Canadian Society of Clinical Chemists. Published by Elsevier Inc. All rights reserved.
doi:10.1016/j.clinbiochem.2008.06.018

## Introduction

New '-omics'-based diagnostic tests are continuously being developed and promoted for use in clinical practice [1], often without proper assessment [2,3]. There is, hence, a need for tools specifically tailored to assess the quality of research on such tests. Journals and research groups have made proposals to enhance the quality of traditional diagnostic research reports [4]. Proposals like STARD (Standards for Reporting of Diagnostic Accuracy) [5] and QUADAS (Quality Assessment of Diagnostic Accuracy Assessment) [6], provide methodologically sound criteria to guide decisions on the use of diagnostic tests in the management of patients and in interpretation of metaanalysis. However, neither STARD nor QUADAS are presently suited for '-omics'-based diagnostic research. The main sources of error described in this area are associated with chance (overfitting) and the analytical and pre-analytical characteristics of the test [7]. Analytical features are partially covered in the available guidelines but because of the complexities of new '-omics'-based methods, these points should be more strictly standardized. Moreover, other important aspects such as overfitting, the pre-analytical procedures or the biological variability of the samples, among others, which have become central to this field because of the higher biological instability of the biomarkers, do not appear in those recommendations.

Initiatives specifically aimed at improving the quality of '-omics'-based diagnostic research have limitations too. MIAME [8] or MIAPE [9] for instance, focus only on the analytical characteristics of the techniques, an early step in the clinical validation of a new diagnostic test. In '-omics'-based diagnostic research, we face problems that are common to traditional diagnostic research, and difficulties that are specific to and particularly important in these new technologies [7,10]. Moreover, quality recommendations should be adapted to the particular characteristics of each phase of test development.

Similarly to phases in drug development, Feinstein [11], Pepe et al. [12] and Sackett and Haynes [13] categorized diagnostic research in different phases, which guide the process of development that a diagnostic test needs to undergo before clinical application. This is a particularly important issue in systematic reviews and metaanalysis, where studies in different phases should also be assessed separately [14], because they answer different research questions and often have different quality requirements.

The aims of this project were, to adapt QUADAS to the particular methodological challenges posed by new molecular diagnostic tests, and to fit QUADAS to each study phase, in order to contribute to the development of specific recommendation on '-omics'-based diagnostic research.

## Materials and methods

Based on work by Whiting et al. in the development of QUADAS [6], our project proceeded through the following stages: 1) preliminary decisions, 2) definition of phases, 3) preliminary item generation, 4) evaluation of the guidelines, and 5) final generation of the guidelines (Fig. 1).

### *Preliminary decisions*

The Steering Committee (see author's affiliations) started with decisions about:

- *Technologies included in the '-omics' definition*: we included technologies that provide a comprehensive analysis of the complete, or near-complete, cellular specific constituents, such as RNAs, DNAs, proteins, and intermediary metabolites. We did not include techniques that only identify some proteins or a single mutation.
- *Fields where these recommendations may be applied*: studies of diagnostic accuracy for clinical practice and screening programs. Prognosis studies were excluded.
- *Aim of these recommendations*: to assess the quality of '-omics'-based diagnostic research for individual studies or when considering the potential inclusion of such studies in systematic reviews and metaanalysis.

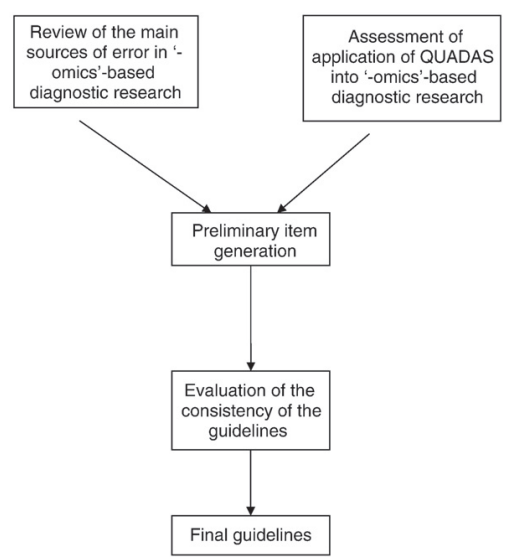



Fig. 1. Flow diagram 1: Overview of the stages of development of the recommendations.

*Definition of phases*

Based on previous work [11–13], Table 1 shows the 4 phases of the process of clinical validation of a new diagnostic test that we used: from the evaluation of the ability of the test to discriminate between healthy controls and known cases of disease, until the ultimate validation phase, which should be carried out in a population as close as possible to that in which the test would be applied in practice. Phases are ordered according to the research sequence usually followed; phases are also related to the strength of evidence –from weakest to strongest– that each phase provides in support of the clinical diagnostic utility of the test. Studies in preliminary phases (phases 1–3) are important in the development of a new diagnostic test. However, excellent results in these phases are not a proof of clinical utility. For instance, in a study [14] to evaluate the diagnostic potential of SELDI-TOF MS in malignant bile duct stricture, the authors collected samples from patients in different phases of cholangiocarcinoma and a group of healthy volunteers. This is a preliminary phase study (phase 1) and despite the authors' conclusions ('serum markers have important diagnostic implications for unknown bile duct stricture'), does not provide evidence that the test would be effective in a clinical situation where patients would be symptomatic and competing diagnoses would be present (phase 4). Defining the study phase as a first step in the quality assessment tool is key to establishing the clinical applicability of findings.

A final phase in diagnostic test development involves prospective observational and prospective randomized trials to measure the value of a new diagnostic test upon health outcomes, once the test has been accepted clinically and made commercially available. We have not covered this issue in this study because the quality requirements of this type of study are distinct from those validating the diagnostic utility of a test before clinical acceptation. However, we do agree that evaluating whether a test influences positively health outcomes is a key aspect.

Table 1
Description of the different phases involved in the clinical validation of a new diagnostic test

| Phase | Description |
|---|---|
| Phase 1 | The test is used to distinguish cases with overt disease from healthy controls. Likewise, some studies in this phase may compare pathological tissue with adjacent healthy tissue. |
| Phase 2 | The spectrum of disease under comparison is broadened. The test is now challenged with different types of diseased cases and a wider range of controls; thus, study patients may have diverse co-morbidities and disease severities, and controls a variety of illnesses and co-morbidities, sometimes with symptoms similar to the disease of interest. |
| Phase 3 | This is an optional phase that aims to detect particular sources of error in the test. The objective is usually to measure the presence of false positive or false negative results in specific groups of patients with particular characteristics that may influence the performance of the test (treatments, autoimmune disease, etc.). It is also time to detect changes in accuracy of the test according to technical modifications. |
| Phase 4 | The test is evaluated in a consecutive series of patients or of healthy people (screenees) that reflect with the maximum degree of fidelity the clinical or public health setting where the test would be used. |

Adapted from Feinstein [11], Pepe et al. [12] and Sackett et al. [13].

*Preliminary item generation*

The initial list of items to be incorporated in the guidelines included: a) all items from QUADAS [6], and b) additional items that specifically addressed main sources of error central in '-omics'-based diagnostic research: specimen collection and management, biological variation, reproducibility and reporting of the analytical conditions of the diagnostic test and overfitting. We also incorporated the definition of the study phase.

The application to genomics and proteomics of each item included in QUADAS was next assessed. To do so, we followed the definitions and applications of QUADAS when possible and, if necessary, we modified them to better address the specific concerns of '-omics'-based research. We assessed the suitability of each item of QUADAS to each study phase.

*Evaluation of the guidelines*

To assess the applicability and consistency of the preliminary guidelines, all researchers independently applied the items to three original articles in '-omics'-based diagnostic research [15–17]. We selected articles from phases 1, 2 and 4; we did not collect a study from phase 3 because this phase is optional to detect particular sources of error. The observer agreement for the application of QUADOMICS was high (kappa 0.89). The main problem arose from application of an initial item ('were the sources, collection and handling of the specimens clearly described? Were pre-analytical procedures similar for the whole sample? And, if differences in procedures were reported, were their effects on the results assessed?'), because it included various different aspects. We decided to divide this item into three different criteria as QUADOMICS finally shows. Then, the Steering Committee convened a second consensus meeting to evaluate the utility of the list of items proposed and to discuss again the explanation of each item. Some items were excluded from the list and others were modified. Further work to validate QUADOMICS in a larger sample of articles is in process.

*Final generation of the guidelines*

## Results

The list with the 16 items included the QUADOMICS tool is shown in Table 2. The two items eliminated from QUADAS were: a) the independence between the reference standard and the index test, because at present '-omics'-based diagnostic tests are not used either as a gold standard or as a part of a gold standard; and b) the description of the withdrawals from the study, because it is already included in the reformulated item 1.

Four new criteria were incorporated to QUADOMICS; criteria 3, 4, 5 and 16. Two more items have new specific descriptions in their definitions (criteria 1 and 10), and two

Table 2
Items included in QUADOMICS, the adaptation of QUADAS to studies on the diagnostic accuracy of '-omics'-based diagnostic research

| Item | Yes | No | Unclear | Not applied |
|---|---|---|---|---|
| 1. Were selection criteria clearly described? | | | | |
| 2. Was the spectrum of patients representative of patients who will receive the test in practice? | | | | |
| 3. Was the type of sample fully described? | | | | |
| 4. Were the procedures and timing of biological sample collection with respect to clinical factors described with enough detail? | | | | |
| 4.1. Clinical and physiological factors | | | | |
| 4.2. Diagnostic and treatment procedures. | | | | |
| 5. Were handling and pre-analytical procedures reported in sufficient detail and similar for the whole sample? And, if differences in procedures were reported, was their effect on the results assessed? | | | | |
| 6. Is the time period between the reference standard and the index test short enough to reasonably guarantee that the target condition did not change between the two tests? | | | | |
| 7. Is the reference standard likely to correctly classify the target condition? | | | | |
| 8. Did the whole sample or a random selection of the sample receive verification using a reference standard of diagnosis? | | | | |
| 9. Did patients receive the same reference standard regardless of the result of the index test? | | | | |
| 10. Was the execution of the index test described in sufficient detail to permit replication of the test? | | | | |
| 11. Was the execution of the reference standard described in sufficient detail to permit its replication? | | | | |
| 12. Were the index test results interpreted without knowledge of the results of the reference standard? | | | | |
| 13. Were the reference standard results interpreted without knowledge of the results of the index test? | | | | |
| 14. Were the same clinical data available when test results were interpreted as would be available when the test is used in practice? | | | | |
| 15. Were uninterpretable/intermediate test results reported? | | | | |
| 16. Is it likely that the presence of overfitting was avoided? | | | | |

items (criteria 2 and 14) are only applied to studies in the last phase of clinical validation (phase 4).

As previous studies did [18], we have added some concrete examples to illustrate the new or modified items included in QUADOMICS (annex 1).

*Notes:*

1. *For the explanations on 'what is meant by this item?' and 'how to score this item?', we refer readers to the QUADAS guidelines. The exceptions are explained in each item.*
2. *Unless otherwise indicated, the item is pertinent to all study phases.*

Determine the phase of diagnostic research according to the design of the study (Table 1):

1 ___ 2 ___ 3 ___ 4 ___.

1. Were selection criteria clearly described?

*a. What is meant by this item*

Specific problems in '-omics'-based research lead us to suggest a description of the criterion stricter than in QUADAS. In 'omics'-based disciplines, availability of samples is a key issue, and researchers often use biobanks with already collected samples. In such cases, '-omics'-based studies are prone to selection bias because sample availability (e.g. tumour tissue) may be associated to clinical and other variables that can influence the discrimination quality of the tests evaluated [19]. The study should hence thoroughly describe the flow of patients from the theoretical study population to the sample finally studied, and the sources of the subjects. Characteristics of patients excluded and included should be compared.

*b. How to score this item*

If detailed information on sources of samples, selection criteria and a flow diagram are included along with a comparison between included and excluded patients, the item should be scored as "yes". Otherwise this item should be scored as "no". If the paper does not provide enough information to answer clearly the above questions the item should be scored as "unclear". Lack of explicit information, including a flow diagram, will yield a "no".

2. Was the spectrum of patients representative of patients who will receive the test in practice?

Referred to QUADAS. Socio-demographic characteristics (such as sex or ethnicity [20]) and clinical factors (like disease stage [21]) can have even more influence on an '-omics'-based diagnostic test than on traditional laboratory tests [10].

*Situations in which this item does apply*

In contrast to QUADAS, this criterion will only be applied to studies in phase 4; phase 1–3 studies do not reproduce the real clinical setting where the test will be applied.

3. Was the type of sample used fully described?

*a. What is meant by this item*

Biomarkers in '-omics'-based diagnostic research can adopt different behaviour or characteristics according to the type of sample collected; for instance, potential marker candidates will be present at a higher concentration in the compartment in which the disease process actually takes place (tissue) than after dilution in peripheral blood [7]. A description of the samples and the processes in their retrieval is essential to reproduce the technique and to know the limitations and applications of the test.

*b. How to score this item*

To score positively in this item, the report should present a detailed description of the type of sample (serum, plasma, other body fluids, tissue, etc.). Moreover, the authors should spe-

cifically list the type of plasma specimen (e.g., EDTA, heparin, citrate), since they could give different results.

4. Were the procedures and timing of biological sample collection with respect to clinical factors described with enough detail?

4.1. Clinical and physiological factors

4.2. Diagnostic and treatment procedures

*a. What is meant by these items*

Observed proteomic patterns may reflect changes in blood concentrations of lipids or hormones, the presence of signs as jaundice and cachexia, the subject's menstrual cycle, ischemia [22], nutritional status, or the effect of diagnostic or treatment procedures [23], and not necessarily the presence of the disease of interest.

*b. How to score these items*

These items would be scored as "yes" if the study includes an analysis of potential factors affecting the protein/metabolite/peptide profile, and a procedure to control biases that they may induce (for instance, stratification). Otherwise, these criteria should be scored as "no".

5. Were handling of specimens and pre-analytical procedures reported in sufficient detail and similar for the whole sample? And if differences in procedures were reported, was their effect on the results assessed?

*a. What is meant by this item*

In '-omics'-based diagnostic research pre-analytical procedures are often more complex than in classic clinical research, and procedures are hence more likely to affect measures of the target marker (e.g. proteins and mRNA tend to have high biological instability) [24,25]. The differential handling of samples, for instance, may be related to different methods and time of preservation, and whole batches of samples should be run under the same conditions [26].

*b. How to score this item*

Any process related to the pre-analytical handling of the samples that could affect the results should be described, and a comparison of the results according to the different procedures be supplied (number of freezing cycles, type of anticoagulant, timing and storing of specimens, time from blood draw until centrifugation and storage, details on centrifugation conditions, etc.). Otherwise, authors should state that the whole set of samples has undergone the same pre-analytical process.

6. Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests?

Referred to QUADAS.

7. Is the reference standard likely to correctly classify the target condition?

Referred to QUADAS.

8. Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis?

Referred to QUADAS.

9. Did patients receive the same reference standard regardless of the index test result?

Referred to QUADAS.

10. Was the execution of the index test described in sufficient detail to permit replication of the test?

*a. What is meant by this item*

This criterion is similar to QUADAS. However, reporting of analytical procedures in '-omics'-based diagnostic research may be more complex than in traditional laboratory research. Hence, a simple citation to a technical article may not be enough. Authors should follow the recommendations for reporting each technique, such as MIAME (Minimum information about a microarray experiment) [8], MIAPE (Minimum Reporting Requirements for Proteomics) [9], Guidelines in Publication of Peptide and Protein Identification Data [27], International standards for reporting metabolomic experimental results [28] and recommendations for the description of sequence variants [29], among others. Studies published before the availability of these guidelines should cover basic aspects as:

- *Mass-spectrometry*: Description of the use of particular technologies: column chromatography, capillary electrophoresis, the use of software to analyze MS data and gel electrophoresis (and its processing and analysis). It should also cover molecular interaction experiments and statistical analysis of data.
- *Microarray data*: description of the set of hybridization experiments as a whole; definition of all arrays used in the experiment; laboratory conditions under which the hybridizations were carried out; measurements to get processed data (the original scan of the arrays, microarray quantification matrices based on image analysis and final gene expression matrix).
- *All*: analytical variability of the test described and controlled. The authors should explicitly describe the degree of instrument or observer variation and the methods used to control this variation (control procedures, reproducibility assessments, calibration, samples collected and run in a random order, etc.).

*b. How to score this item*

Studies that report having followed some of the guidelines above or studies previous to the publication of the recommendations that cover the aspects formerly mentioned are scored positively.

11. Was the execution of the reference standard described in sufficient detail to permit its replication?

Referred to QUADAS.

12. Were the index test results interpreted without knowledge of the results of the reference standard?

Referred to QUADAS.

13. Were the reference standard results interpreted without knowledge of the results of the index test?

Referred to QUADAS.

14. Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?

Referred to QUADAS.

*Situations in which this item does apply*

Only for studies in phase 4; phase 1–3 studies do not attempt to reproduce the real clinical setting where the test will be applied.

15. Were uninterpretable/intermediate test results reported?

Referred to QUADAS and to the later modification proposed in scoring this item [30].

16. Is it likely that overfitting was avoided?

*a. What is meant by this item*

Overfitting may occur in the analysis of large datasets when multivariate models show apparent discrimination that is actually caused by data over-interpretation, and hence give rise to results that are not reproducible [31,32]. The chance of overfitting, however, can be reduced by appropriate application of validatory estimation and assessment, such as through application of cross-validation. To develop and validate a method of classification, it is best to have a large collection of samples, which allow analyses of an independent training test and test set. In practice, usually, only a limited number of samples are available, and several methods are used to deal with overfitting, such as cross-validation (from simply splitting the sample in two parts to the most extreme version, "leave-one-out") and resampling methods (bootstrap, jackknife and permutation tests) [33,34].

*b. How to score this item*

This item will be scored as "yes" if the authors performed a validation test in an independent set of samples or used some approach to deal with overfitting. However, if the study used the same sample for the test and training set, it should be scored 'no'.

## Discussion

The recommendations included in QUADOMICS represent an adaptation of QUADAS that may be applied to the quality assessment of individual diagnostic accuracy studies on '-omics'-based research, and to candidates for inclusion in systematic reviews or metaanalysis. We found that as well as the modification of two original items, at least 4 new items were needed in order to address the specific design features and errors that are relevant in studies of "-omics" derived diagnostic tests.

QUADAS is a useful and reliable tool but is generic for all type of diagnostic research; its authors reported that work is being carried out to adapt the guidelines to different diagnostic topics and designs [6]. QUADOMICS is a tool adapted to assess the quality of diagnostic studies in a highly dynamic field which faces the challenge of sieving the huge amount of results recently produced and translating them into clinical and public health practice [35,36]. Systematic reviews will have a key role in this endeavour, hence the opportunity for and relevance of a suitable assessment tool.

At this stage, some features of our proposal are worth highlighting. We wanted to stress the relevance of reporting the diagnostic phase of every specific study. As previously mentioned, grouping diagnostic studies from different phases when performing a systematic review is not recommendable as they answer different research questions. We hold that the combination of heterogeneous studies should be completely avoided in '-omics' research. While in other diagnostic fields some studies comparing cases of diseased subjects with a spectrum of non-diseased controls could, under certain conditions, contribute to the estimations of accuracy indexes, in '-omics' this procedure is more likely to give flawed results. Our proposal prevents this mixture as it reports the study phase and recognizes the applicability of some items of the tool exclusively to studies in certain phases.

The main reason why "-omics" diagnostic studies in preliminary phases of research are more prone to give mistaken results is overfitting. Although the problem of overfitting has already been recognized in the traditional diagnostic area; it came to the forefront several years ago when a study reported that a blood test, based on a pattern-recognition proteomics analysis of serum, was nearly 100% sensitive and specific for ovarian cancer [37]. However, these data did not demonstrate reproducibility in independent subjects and the results were explained simply by chance and bias [31]. A relevant feature of QUADOMICS is the inclusion of a specific item to ascertain the presence of overfitting and the methods used to deal with it in the reviewed studies.

Another concern of discovery phases in "-omics" diagnostic research is the influence that the type of biological sample and its collection and handling procedures have on the test results. Our proposed tool adds in three new criteria in order to check these significant characteristics. We also wanted to stress the importance that studies report appropriately the analytical procedures and therefore suggested that authors follow the guidelines MIAME [8] and MIAPE [9] or other appropriate recommendations [27–29] when describing the execution of the tests. These recommendations are useful and opportune in a field where the standardization of techniques is particularly necessary.

QUADAS was a decisive step in contributing to an adequate process of systematic review of diagnostic studies and its evaluation proved that the tool was reproducible and needed merely minor changes [30]. This adaptation, QUADOMICS, has the advantage of building upon the previous original and high quality work of QUADAS contributors; however, the new tool may face challenges regarding the reproducibility of the added items. In order to avoid inconsistencies in the application of the tool we have assured that precision in the writing took priority over applicability, that is, we chose to be stricter in the scoring of items rather than to enable wide but imprecise application. As a result, the tool is very demanding but reproducible.

In spite of the high expectations, few of the many "-omics" tests proposed have moved on from the discovery phase to an appropriate validation phase. Furthermore, excellent results in preliminary phases are not a proof of clinical utility, as the few present clinical applications demonstrate [3]. The usual gap existing between basic research and clinical practice is even

greater in '-omics'-based diagnostic research. Most of the work is devoted to overcoming technological challenges. This is indeed essential but more attention should be paid to an efficient process in order to confirm discoveries through independent validation studies [36]. Availability of quality assessment tools that integrate basic requirements as well as clinical study design features and bias control could remind researchers of the need to translate basic results to practice through appropriate studies. The publication of STARD had a positive effect on the quality of diagnostic research [38,39]. Tools such as QUADAS primarily designed to be applied in systematic reviews have a prospective positive effect on researchers when designing their diagnostic studies. In addition to providing reviewers of "-omics" diagnostic studies with an adequate tool, QUADOMICS also contributes to the opportune design of validation studies. The next important step is the evaluation of QUADOMICS through its application to a sufficient sample of empirical studies.

## Acknowledgments

This work was supported by Spanish Agency for Health Technology Assessment, Exp PI06/90311, Instituto de Salud Carlos III.

We acknowledge partial funding and support to this research from the CIBER en Epidemiología y Salud Pública (CIBER-ESP), Spain.

The funding sources, Spanish Agency for Health Technology Assessment and CIBER, had no role in the design, conduct, or reporting of the study or in the decision to submit the manuscript for publication.

## Appendix A

*Annex 1: Examples of application of QUADOMICS in real studies*

References are listed at the end of this annex.

Item 1: Were selection criteria clearly described?

- Example 1 [1].

*Presentation*: The authors analyzed the clinical utility of a proteomic test in the diagnosis of recurrent bladder cancer and compared its usefulness with cytology.

*Extract*: Twenty-three clinical sites in 9 states, including academic, private practice, and veterans' facilities, prospectively enrolled 668 consecutive patients with a history of bladder cancer between September 2001 and February 2002 (figure of a flow diagram).

*Comment*: The study detailed how patients were selected for inclusion (consecutively), selection criteria (history of bladder cancer between September 2001 and February 2002), and it included a flow diagram of eligible patients and reasons for exclusion from the study. This item would be scored as yes.

Item 2: Was the spectrum of patients representative of patients who will receive the test in practice?

- Example 2 [2].

*Presentation*: In this study the authors developed a ProteinChip Array as a non-invasive method, in contrast to renal biopsy, for the detection of renal transplant rejection.

*Extract*: We conducted a retrospective study of midstream urine samples from 23 consecutive transplant patients that were subjected to SELDI time-of-flight mass spectrometry in an attempt to identify biomarkers for rejection. A total of 23 urine samples were collected from 13 patients showing biopsy-proven renal allograft rejection and from 10 patients without histological signs of rejection. All 23 patients had clinical symptoms and signs of acute allograft rejection and underwent renal biopsy.

*Comment*: The authors included a consecutive sample of patients with clinical symptoms of transplant rejection. This population represents the patients who would receive the test in practice based on the method of recruitment (consecutive patients) and in the symptoms and signs of the patients (patients with renal transplant where the available tests do not provide a diagnosis of rejection).

Item 3: Was the type of sample fully described?

- Example 3 [3].

*Presentation*: The objective of this study was to evaluate the simultaneous detection of expression levels of a multiple mRNA marker panel in the peripheral blood of colorectal cancer (CRC) patients for use in complementary CRC diagnosis. The authors collected twenty-seven tumour tissue specimens and 80 peripheral blood specimens from CRC patients.

*Extract*: Among 80 pairs of CRC tissue and adjacent normal colorectal tissue surgically removed from the patients, 27 were randomly selected for further analysis. Additionally, a 5-mL sample of peripheral blood was obtained from each of the 80 CRC patients at the time of surgical resection and from 98 healthy volunteers serving as normal controls. To prevent contamination of epithelial cells, peripheral blood samples were obtained through a catheter inserted into a peripheral vessel, and the first 5 mL of blood were discarded.

*Comment*: The authors specified the type of sample, and in the case of the blood sample, they described with detail the method of collection. This item should be scored as yes.

Item 4: Were the procedures and timing of biological sample collection with respect to clinical factors described with enough detail?

4.1. Clinical and physiological factors.

- Example 4 [4].

*Presentation*: The study analyses serum proteome to evaluate the role of some proteins as diagnostic biomarkers for

idiopathic osteonecrosis of the femoral head (IONFH). The authors selected 10 patients with IONFH and 10 normal subjects.

*Extract*: Serum samples: To minimize individual variation, genders and ages of patients were matched in both the normal and the IONFH groups in the proteomic study.

*Comment*: In this case, the authors considered that factors such as gender and age could affect the results. Therefore, they controlled those possible biases through matching the samples. This item should be scored as yes.

4.2. Diagnostic and treatment procedures.

Example 5 [5].

*Presentation*: In this study the authors searched for endometriosis-specific proteins to distinguish women with and without endometriosis.

*Extract*: All women had no other diseases on physical examination and biochemical tests. None of them had received any hormonal treatment in the 3 months before this study.

*Comment*: The author detailed the absence of potentially known factors, diagnosis of other diseases and treatment procedures (hormonal treatment), which could affect the protein profile in the diagnosis of endometriosis. This item should be scored as yes.

Item 5: Were handling and pre-analytical procedures reported in sufficient detail and similar for the whole sample? And, if differences in procedures were reported, was their effect on the results assessed?

- Example 6 [6].

*Presentation*: This study aimed to develop and test serum protein profiles as indicatives of the presence of breast cancer. The sample size included serum samples from 78 patients 1 day prior to surgery for breast cancer and 29 healthy female volunteers.

*Extract*: Serum samples: All samples were collected and processed following a standardized protocol: the samples were collected in a 10 $cm^3$ Serum Separator Vacutainer Tube (BD Diagnostics. Plymouth, UK), and centrifuged 30 min later at 3000 rpm for 10 min. The serum samples were distributed into 1-mL aliquots and stored at −70 °C. After thawing on ice, the serum samples were randomized over different 96-well microtitration racks (Matrix) and then stored at −70 °C until the experiment.

*Study design*: we used a randomized block design to avoid any potential batch effects. At the available 106 samples from both groups were randomly distributed across 3 plates in roughly equal proportions. For breast cancer, the distribution of stadia across plates was again in random fashion and in approximately equal proportions. The position on the plates of samples allocated to each plate was randomized as well. Each plate was then assigned to a distinct day. Analyses were carried out on 3 consecutive days, Tuesday to Thursday, processing a single plate each day.

*Comment*: In this case, the authors thoroughly described the conditions of the samples before the analysis. In order to avoid the different handling of samples and its adverse consequences, they also used a randomized block design. This item should be scored as yes.

Item 10: Was the execution of the index test described in sufficient detail to permit replication of the test?

- Example 7 [7].

*Presentation*: This study evaluated proteomic approaches to identify new biomarkers for detection and monitoring of ovarian cancer through the analysis of three sets: 1) 21 ovarian cancers, 18 benign diseases, and 20 normal patients; 2) 32 ovarian cancers, 30 benign ovarian diseases, and 30 age-matched healthy controls; and, 3) samples collected before and after chemotherapy from 18 ovarian cancer patients.

*Extract*: To assess inter- and intra-assay reproducibility, a pooled serum sample (from 5 normal sera) was processed multiple times during experiments on the second and third sample sets. The order in which samples were processed and the spotting allocation of samples in chips and bioprocessors were randomized using an in-home experiment design software.

*Comment*: Besides of basic features covering aspects of protein chip array analysis and bioinformatics and statistics procedures, it is essential the description of the measure of inter- and intra-assay reproducibility. This item should be scored as yes.

Item 16: Is it likely that the presence of overfitting was avoided?

- Example 8 [8].

*Presentation*: The authors evaluated autoantibody signatures on a panel of 22 peptides for the early detection of prostate cancer. The study included a sample of 139 different types of cases and 149 controls.

*Extract*: These samples were randomly separated into a training set (129 samples, including 59 cancers and 70 controls) and a validation set (128 samples, including 60 cancers and 68 controls). The training samples were used to identify phage-peptides with high specificity and sensitivity for the detection of prostate cancer. A total of 22 phage clones were selected, with 97.1% specificity and 88.1% sensitivity for detection of prostate cancer in this group of 129 serum samples. These results were then tested against the second independent validation set of 60 patients with prostate cancer and 68 control subjects. Within this validation cohort, the 22 selected phage-peptide clones had a specificity of 88.2% and a sensitivity of 81.6% for the detection of prostate cancer.

*Comment*: To avoid overfitting, the authors split the initial sample in two independent groups: the training set (with 129 samples), where the authors identified the peptides associated with prostate cancer, and the validation set (with 128 samples), where the previous results were independently tested. This is the most suitable approach to validate a proteomic diagnostic test; hence, we should score this item as yes.

- Example 9 [9].

*Presentation*: This study aimed to discover potential biomarkers in serum proteomics for the detection and monitoring of adjuvant chemotherapy for ovarian cancer. The sample included untreated ovarian cancer patients (64) and non-cancer population (31 patients with benign ovarian diseases and 30 healthy female volunteers). An additional 16 postoperative patients with epithelial ovarian cancer were recruited for identifying potential biomarkers related to adjuvant chemotherapy.

*Extract*: From SELDI spectra of training set, we identified a total of 156 raw peaks in the m/z region of 1000–20,000. Using Biomarker Patterns Software, we compared the spectrum generated from control group with the spectrum generated from untreated cancer group. This comparison yielded a model consisting of 4 peaks that discriminated between non-cancer sera and cancer serum from patients with ovarian cancer. These 4 peaks corresponded to m/z ratios of 6195, 6311, 6366, and 11,498 (Fig. 1). The m/z 6195, 6311, and 6366 peaks were down-regulated in the cancer group, and the m/z 11,498 peak was up-regulated in the cancer group. The accuracy of this model was shown in Table 2. A blind test set consisted of 23 cancer cases and 20 controls were used for evaluation of this multivariate model to distinguish ovarian cancer from non-cancer cohort. In our study, 19 out of 20 of the true non-cancer cases were correctly classified, and 20 of 23 cancer samples, including all 4 stage I cancers, were correctly classified as malignant. This result yielded a sensitivity of 87.0%, and a specificity of 95.0%.

*Comment*: The authors carried out an initial analysis in a training set to identify the potential biomarkers. Then, they validated this pattern in an independent sample. Therefore, overfitting could have not been avoided: we should score this item as no.

## References


[1] Thomas DC. High-volume "-omics" technologies and the future of molecular epidemiology. Epidemiology 2006;17:490–1.
[2] Check E. Proteomics and cancer: running before we can walk? Nature 2004;429:496–7.
[3] Marshall E. Getting the noise out of gene arrays. Science 2004;306:630–1.
[4] Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med 2004;140:189–202.
[5] Bossuyt PM, Reitsma JB, Bruns DE, et al. Standards for Reporting of Diagnostic Accuracy. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Clin Chem 2003;49:7–18.
[6] Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol 2003;3:25–37.
[7] Zolg W. The proteomic search for diagnostic biomarkers: lost in translation? Mol Cell Proteomics 2006;5:1720–6.
[8] Brazma A, Hingamp P, Quackenbush J, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 2001;29:365–71.
[9] Taylor CF. Minimum reporting requirements for proteomics: a MIAPE primer. Proteomics 2006;6(Suppl 2):39–44.
[10] Lay JO, Borgmann S, Liyanage R, Wilkins CL. Problems with the omics. Trends Analyt Chem 2006;25:1046–56.
[11] Feinstein AR. Clinical epidemiology: the architecture of clinical research. Philadelphia: WB Saunders; 1985.
[12] Pepe MS, Etzioni R, Feng Z, et al. Phases of biomarker development for early detection of cancer. J Natl Cancer Inst 2001;93:1054–61.
[13] Sackett DL, Haynes RB. Evidence base of clinical diagnosis: the architecture of diagnostic research. BMJ 2002;324:539–41.
[14] Scarlett CJ, Saxby AJ, Neilson AQ, Bell C, Samra JS, Hugh T, et al. Proteomic profiling of cholangiocarcinoma: diagnostic potential of SELDI-TOF MS in malignant bile stricture. Hepatology 2006;44:658–66.
[15] Inoue M, Sakaguchi J, Sasagawa T, Tango M. The evaluation of human papillomavirus DNA testing in primary screening for cervical lesions in a large Japanese population. Int J Gynecol Cancer 2006;16:1007–13.
[16] Ahn BY, Song ES, Cho YJ, Kwon OW, Kim JK, Lee NG. Identification of an anti-aldolase autoantibody as a diagnostic marker for diabetic retinopathy by immunoproteomic analysis. Proteomics 2006;6:1200–9.
[17] Huang LJ, Chen SX, Huang Y, et al. Proteomics-based identification of secreted protein dihydrodiol dehydrogenase as a novel serum markers of non-small cell lung cancer. Lung Cancer 2006;54:87–94.
[18] Dupuy A, Simon RM. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. J Natl Cancer Inst 2007;99:147–57.
[19] Porta M, Hernández-Aguado I, Lumbreras B, Crous-Bou M. 'Omics' research, monetization of intellectual property and fragmentation of knowledge: can clinical epidemiology strengthen integrative research? J Clin Epidemiol 2007;60:1220–5.
[20] Villanueva J, Martorella AJ, Lawlor K, et al. Serum peptidome patterns that distinguish metastatic thyroid carcinoma from cancer-free controls are unbiased by gender and age. Mol Cell Proteomics 2006;5:1840–52.
[21] Cowen EW, Liu CW, Steinberg SM. Differentiation of tumour-stage mycosis fungoides, psoriasis vulgaris and normal controls in a pilot study using serum proteomic analysis. Br J Dermatol 2007;157:946–53.
[22] Dash A, Maine IP, Varambally S, Shen R, Chinnaiyan AM, Rubin MA. Changes in differential gene expression because of warm ischemia time of radical prostatectomy specimens. Am J Pathol 2002;161:1743–8.
[23] Lin DW, Coleman IM, Hawley S, et al. Influence of surgical manipulation on prostate gene expression: implications for molecular correlates of treatment effects and disease prognosis. J Clin Oncol 2006;24:3763–70.
[24] Hsieh S, Chen RK, Pan YH, Lee HL. Systematic evaluation of the effects of sample collection procedures on low-molecular-weight serum/plasma proteome profiling. Proteomics 2006;6:3189–98.
[25] Baumann S, Ceglarek U, Fiedler GM, Lembcke J, Leichtle A, Thiery J. Standardized approach to proteome profiling of human serum based on magnetic bead separation and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Clin Chem 2005;51:6973–80.
[26] Hu J, Coombes KR, Morris JS, Baggerly KA. The importance of experimental design in proteomic mass spectrometry experiments: some cautionary tales. Brief Funct Genomic Proteomic 2005;3:322–31.
[27] Bradshaw RA. Revised draft guidelines for proteomic data publication. Mol Cell Proteomics 2005;4:1223–5.
[28] Castle AL, Fiehn O, Kaddurah-Daouk R, Lindon JC. Metabolomics Standards Workshop and the development of international standards for reporting metabolomics experimental results. Brief Bioinform 2006;7:159–65.
[29] den Dunnen JT, Antonarakis SE. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. Hum Mutat 2000;15:7–12.
[30] Whiting PF, Weswood ME, Rutjes AW, Reitsma JB, Bossuyt PN, Kleijnen J. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. BMC Med Res Methodol 2006;6:9–16.
[31] Ransohoff DF. Rules of evidence for cancer molecular-marker discovery and validation. Nat Rev Cancer 2004;4:309–14.
[32] Diamandis EP. Analysis of serum proteomic patterns for early cancer diagnosis: drawing attention to potential problems. J Natl Cancer Inst 2004;96:353–6.
[33] Mertens BJ, De Noo ME, Tollenaar RA, Deelder AM. Mass spectrometry proteomic diagnosis: enacting the double cross-validatory paradigm. J Comput Biol 2006;13:1591–605.
[34] Baggerly KA, Morris JS, Edmonson SR, Coombes KR. Signal in noise: evaluating reported reproducibility of serum proteomic tests for ovarian cancer. J Natl Cancer Inst 2005;97:307–9.
[35] Chin L, Gray JW. Translating insights from the cancer genome into clinical practice. Nature 2008;452:553–63.

[36] Hanash SM, Pitteri SJ, Faca VM. Mining the plasma proteome for cancer biomarkers. Nature 2008;452:571–9.
[37] Petricoin EF, Ardekani AM, Hitt BA, et al. Use of proteomic patterns in serum to identify ovarian cancer. Lancet 2002;359:572–7.
[38] Lumbreras-Lacarra B, Ramos-Rincón JM, Hernández-Aguado I. Methodology in diagnostic laboratory test research in clinical chemistry and clinical chemistry and laboratory medicine. Clin Chem 2004;50: 530–6.
[39] Smidt N, Rutjes AW, van der Windt DA, et al. The quality of diagnostic accuracy studies since the STARD statement: has it improved? Neurology 2006;67:792–7.

## Further reading

[1] Grossman HB, Soloway M, Messing E, et al. Surveillance for recurrent bladder cancer using a point-of-care proteomic assay. JAMA 2006;295:299–305.
[2] Reichelt O, Múller J, von Eggeling F, et al. Prediction of renal allograft rejection by urinary protein analysis using ProteinChip Arrays (surface-enhanced laser desorption/ionization time-of-∗flight mass spectrometry). Urology 2006;67:472–5.
[3] Yeh CS, Wang JY, Wu CH, et al. Molecular detection of circulating cancer cells in the peripheral blood of patients with colorectal cancer by using membrane array with a multiple mRNA marker panel. Int J Oncol 2006;28:411–20.
[4] Tan X, Cai D, Wu Y, et al. Comparative analysis of serum proteomes: discovery of proteins associated with osteonecrosis of the femoral head. Transl Res 2006;148:114–9.
[5] Zhang H, Niu Y, Feng J, Guo H, Ye H, Cui H. Use of proteomic analysis of endometriosis to identify different protein expression in patients with endometriosis versus normal controls. Fertil Steril 2006;86:274–82.
[6] de Noo ME, Deelder A, van der Werff M, Ózalp A, Mertens B, Tollenaar R. MALDI-TOF serum protein profiling for the detection of breast cancer. Onkologie 2006;29:501–6.
[7] Kong F, White CN, Xiao X, et al. Using proteomic approaches to identify new biomarkers for detection and monitoring of ovarian cancer. Gynecol Oncol 2006;100:247–53.
[8] Bradford TJ, Wang X, Chinnaiyan AM. Cancer immunomics: using autoantibody signatures in the early detection of prostate cancer. Urol Oncol 2006;24:237–42.
[9] Zhang H, Kong B, Qu X, Jia L, Deng B, Yang Q. Biomarker discovery for ovarian cancer using SELDI-TOF-MS. Gynecol Oncol 2006;102:61–6.

## 5.2 Resumen de los hallazgos principales del artículo 2

**(Referencia:** Parker LA, Gomez Saez N**,** Lumbreras B, Porta M, Hernández-Aguado I. Methodological Deficits in Diagnostic Research Using '-Omics' Technologies: Evaluation of the QUADOMICS Tool and Quality of Recently Published Studies. PLoS One 2010;5:e11419.)

**Los déficits metodológicos en la investigación diagnóstica que utiliza tecnologías '-ómicas': Aplicabilidad y consistencia de QUADOMICS y evaluación de la calidad de estudios recientes.**

Este artículo trata de alcanzar los objetivos específicos 2 y 3 de esta tesis doctoral, es decir la validación de la nueva guía QUADOMICS mediante el análisis de su aplicabilidad y consistencia, y una descripción de la calidad metodológica de estudios diagnósticos basados en tecnologías '-ómicas'.

La estrategia de búsqueda identificó 164 artículos potenciales, de los cuales se seleccionaron 59 para revisión del texto completo. Finalmente se seleccionaron 45 artículos para su inclusión en el estudio. Los mismos 45 artículo se utilizaron para la validación de QUADOMICS (apartado 5.2.1) y la evaluación de la calidad metodológica de una muestra de estudios diagnósticos que utilizan tecnologías '-ómicas' (apartado 5.2.2).

1) Evaluación de la aplicabilidad y consistencia de QUADOMICS:

De manera independiente, tres investigadores aplicaron QUADOMICS a 45 estudios diagnósticos que usaron tecnologías '-ómicas'. Las observaciones de cada investigador se compararon con el consenso establecido y el porcentaje de acuerdo entre cada investigador y el consenso establecido fue 83%, 90% y 82% respectivamente.

Cuatro de los ítems no se aplicaron a todos los estudios. Ítems 2 y 14 se debe aplicar únicamente a estudios de fase IV. Asimismo, los ítems 9 (¿Los pacientes recibieron el mismo estándar de referencia a pesar del resultado de la prueba de estudio?) y 13 (¿Se interpretaron los resultados del estándar de referencia sin conocimiento de los resultados obtenidos con la prueba de estudio?) se puntuaron como 'no aplicable' en varios

estudios. Los motivos fueron los siguientes: 1) la prueba '-ómica' se llevó a cabo después del estándar de referencia y 2) varios estudios no incluyeron una prueba independiente como estándar de referencia ya que utilizaban el diagnóstico anterior de la enfermedad en cuestión como referencia para el cálculo de la exactitud diagnóstica. La falta de un estándar de referencia independiente y claramente definido contribuyó a varios problemas con la aplicación de los ítems de QUADOMICS que referían al estándar de referencia.

A continuación se describen los cuatro ítems que presentaron las mayores dificultades en su aplicación y cuya consistencia se considera 'baja' según nuestra definición anteriormente escrita (concordancia con el consenso fue inferior al 60% para al menos un revisor, o si dos o más de los revisores alcanzaron menos del 80% de acuerdo con el consenso).

*Ítem 4.1: ¿Se describieron los procedimientos y los tiempos para la recogida de las muestras biológicas con respecto a los factores clínicos con suficiente detalle? - ¿Factores clínicos y fisiológicos?*
El desacuerdo se centró en la definición de lo que se debe considerar 'con suficiente detalle'. Para el presente estudios se decidió que un estudio cumpliría este criterio si los autores proporcionaban alguna información clínica adicional como el estadío de la enfermedad, además de la edad y el sexo de los pacientes. Para ayudar en la aplicación de este criterio se aconseja que los investigadores que pretendan usar QUADOMICS como herramienta para evaluar la calidad de estudios incluidos en una revisión sistemática, se deciden de antemano qué aspectos clínicos pueden influir en el perfil '-ómico' estudiado y cuáles se deben indicar como mínimo para cumplir con este criterio.

*Ítem 6: ¿El periodo de tiempo entre la aplicación del estándar de referencia y la prueba de estudio fue suficientemente corto para garantizar que la condición no hubiera cambiado?*
Este criterio es especialmente relevante cuando consideramos pruebas proteómicas porque el perfil proteico puede variar substancialmente en diferentes estadios de la enfermedad. La dificultad en la aplicación de este criterio se debió a que varios de los estudios evaluados carecían de un estándar de referencia independiente y bien descrito, donde habían seleccionado pacientes diagnosticados de la enfermedad bajo estudio y

una serie de controles. Se decidió que este criterio debía aplicarse considerando el momento del diagnóstico como el estándar de referencia. Los estudios cumplen este criterio si el diagnóstico está confirmado en el momento de recoger la muestra para la prueba '-ómica', o si se describe el tiempo que ha trascurrido desde el diagnóstico hasta la recogida de muestra y se considera suficientemente corto para garantizar que la condición no haya cambiado. Por otra parte aquellos estudios que no mencionan cuándo se ha diagnosticado a los pacientes, se debe marcar como 'no se aclara'.

*Ítem 11: ¿Se describió la ejecución del estándar de referencia con suficiente detalle para permitir su replicación?*

De manera parecida, la aplicación de este ítem fue difícil por la ausencia de un estándar de referencia independiente y claramente descrito en muchos de los estudios evaluados. Se evaluó si los criterios usados para diagnosticar los casos de enfermedad, o para establecer la ausencia de enfermedad en los controles, fueron descritos con suficiente detalle. Había estudios que describieron adecuadamente el diagnóstico en casos pero no fue así en los controles Se recomienda que aquellos investigadores que pretendan usar QUADOMICS como herramienta de evaluar la calidad de estudios incluidos en una revisión sistemática, deciden previamente: 1) si quieren incluir artículos que utilizan el diagnóstico ya establecido como un estándar de referencia y 2) si deciden incluirlos, qué información mínima deben presentar los autores para asegurar la ausencia de enfermedad en los controles.

*Ítem 15: ¿Se informó sobre los resultados no interpretables o intermedios?*

Muchos estudios no mencionan de manera explícita la presencia o ausencia de resultados no interpretables (sobre todo la ausencia). En este caso, la modificación de QUADAS dice que se debe evaluar si el estudio describe los resultados para todos los pacientes seleccionados en el estudio inicialmente. Obtuvimos baja consistencia al aplicar este criterio porque a veces fue difícil establecer si todos los pacientes que habían entrado el estudio proporcionaron resultados que se utilizaron para el cálculo del índice de exactitud diagnóstica. De hecho, algunos estudios presentaban el cálculo de la sensibilidad para varios perfiles de biomarcadores sin clarificar las muestras incluidas para su cálculo. Se decidió que estos casos se debían considerar como 'no está claro'.

2) Evaluación de la calidad metodológica de una muestra de estudios diagnósticos que utilizan tecnologías '-ómicas':

De los 45 estudios evaluados, 35 (78%) eran de fase I, es decir utilizaron un diseño de casos y controles, y 6 (13%) se llevaron a cabo en una muestra consecutiva de pacientes similares a aquellos que recibirán la prueba en la práctica.

Había mucha variación en la calidad de los artículos estudiados: uno cumplió solo 2 de los 13 criterios aplicados (15%) y otro cumplió 12 de los 13 ítems aplicados (92%). En general, la calidad metodológica de los estudios evaluados fue pobre: media de cumplimiento 55% ± 18%. A continuación se describe algunos de los fallos más comunes en la muestra de estudios evaluados.

*Aspectos relacionados con las pacientes y con las muestras*

Ningún estudio describió la selección de pacientes con suficiente detalle (ítem 1) y se consideraron los factores clínicos o fisiológicos del paciente que se somete a la prueba en menos de la mitad de los artículos evaluados (ítem 4.1, 20 estudios, 45%). Por otra parte, se consideraron los procedimientos diagnósticos o tratamientos recibidos por las pacientes antes de la recogida de la muestra en la mitad de los estudios (ítem 4.2, 22 estudios, 49%)

*Aspectos relacionados con las prueba de estudio*

19 (42%) estudios no describieron la prueba '-ómica' en suficiente detalle para permitir su replicación (ítem 10). Solo 20 (44%) de los estudios mencionó que la interpretación de la nueva prueba '-ómica' se realizó de manera independiente a los resultados del estándar de referencia (ítem 13). Esta omisión sugiere que el sesgo de revisión podría estar presente y podría conllevar a la sobreestimación de la exactitud diagnóstica.

*Aspectos relacionados con el estándar de referencia*

La mayoría de los estudios no incluyó una prueba independiente como referencia sino que utilizaron el diagnostico establecido de la enfermedad bajo estudio. Solo 21 (47%) de los estudios describieron este proceso con suficiente detalle (ítem 11) y en 24 (53%) de los estudios no se podía evaluar el sesgo de progresión de la enfermedad porque no

se mencionó el periodo de tiempo transcurrido entre el diagnóstico inicial de la enfermedad y la prueba de estudio.

*Overfitting*

Asimismo, en 22 (49%) de los estudios no se indicaron medidas para evitar el overfitting y en 3 estudios (7%) no fue claro si la validación de la prueba fue llevada a cabo en la misma muestra de pacientes utilizadas para construir el modelo o en una población independiente. Estudios vulnerables al overfitting podrían presentar resultados optimistas que no son reproducibles en otras muestras de pacientes.

Los datos suplementarios para el artículo se encuentran en el anexo 2.

UNIVERSITAS
Miguel
Hernández

OPEN ACCESS Freely available online

PLoS one

# Methodological Deficits in Diagnostic Research Using '-Omics' Technologies: Evaluation of the QUADOMICS Tool and Quality of Recently Published Studies

**Lucy A. Parker[1,2]*, Noemí Gómez Saez[1], Blanca Lumbreras[1,2], Miquel Porta[2,3], Ildefonso Hernández-Aguado[1,2]**

**1** Departamento de Salud Pública, Universidad Miguel Hernández, Alicante, Spain, **2** Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública, (CIBERESP), Barcelona, Spain, **3** Institut Municipal d'Investigació Mèdica, Facultat de Medicina, Universitat Autònoma de Barcelona, Barcelona, Spain

## Abstract

***Background:*** QUADOMICS is an adaptation of QUADAS (a quality assessment tool for use in systematic reviews of diagnostic accuracy studies), which takes into account the particular challenges presented by '-omics' based technologies. Our primary objective was to evaluate the applicability and consistency of QUADOMICS. Subsequently we evaluated and describe the methodological quality of a sample of recently published studies using the tool.

***Methodology/Principal Findings:*** 45'-omics'- based diagnostic studies were identified by systematic search of Pubmed using suitable MeSH terms (''Genomics'', ''Sensitivity and specificity'', ''Diagnosis''). Three investigators independently assessed the quality of the articles using QUADOMICS and met to compare observations and generate a consensus. Consistency and applicability was assessed by comparing each reviewer's original rating with the consensus. Methodological quality was described using the consensus rating. Agreement was above 80% for all three reviewers. Four items presented difficulties with application, mostly due to the lack of a clearly defined gold standard. Methodological quality of our sample was poor; studies met roughly half of the applied criteria (mean $\pm$ sd, 54.7$\pm$18.4%). Few studies were carried out in a population that mirrored the clinical situation in which the test would be used in practice, (6, 13.3%); none described patient recruitment sufficiently; and less than half described clinical and physiological factors that might influence the biomarker profile (20, 44.4%).

***Conclusions:*** The QUADOMICS tool can consistently be applied to diagnostic '-omics' studies presently published in biomedical journals. A substantial proportion of reports in this research field fail to address design issues that are fundamental to make inferences relevant for patient care.

**Citation:** Parker LA, Gómez Saez N, Lumbreras B, Porta M, Hernández-Aguado I (2010) Methodological Deficits in Diagnostic Research Using '-Omics' Technologies: Evaluation of the QUADOMICS Tool and Quality of Recently Published Studies. PLoS ONE 5(7): e11419. doi:10.1371/journal.pone.0011419

**Editor:** Antje Timmer, HelmholtzZentrum München, Germany

**Received** February 8, 2010; **Accepted** June 9, 2010; **Published** July 2, 2010



**Funding:** This work was supported by the Spanish Agency for Health Technology Assessment, Exp PI06/90311, Instituto de Salud Carlos III and CIBER en Epidemiologi-a y Salud Pública (CIBERESP) in Spain. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

* E-mail: lparker@umh.es

## Introduction

Technological advances in the past 20 years have permitted large-scale parallel measurements of biochemical and cellular constituents for study as a unified whole, spurring what may be referred to as the '-omics' revolution. [1–3] By adding the suffix '-omics', we can refer to the comprehensive study of almost any cellular constituent. For instance, transcriptomics refers to analysis of total mRNA expression and proteomics refers to the analysis of the proteome, the total protein content. The coupling of these high throughput technologies with computer-assisted discrimination systems may substantially influence the future of clinical diagnosis, leading to diagnostic tests based on multi-marker patterns, biomarker profiles or signatures, rather than on a single alteration [1,4].

Despite rigorous and vigorous promotion of '-omics' based technologies for diagnosis of human diseases, few of the many tests proposed have been introduced into clinical practice with clearly documented clinical benefits. [5–7] Analysis and interpretation of the diagnostic capacity of '-omics' based technologies has presented unique challenges, [8] and reproducing the initial claims of diagnostic accuracy in independent populations has often proved complex. [9,10] The apparent -but in fact artifactual- power to discriminate between diagnostic groups using '-omics' technologies may actually be due to methodological features of the studies; most notably, differences in the pre-analytical procedures, [11] in the clinical or pathophysiological characteristics of the patients who provided the biological samples, [12–14] or simply chance. [15,16] Consequently, in '-omics' studies investigators must consider the potential genetic variation between different individuals, or how certain physiological characteristics (disease pathophysiology, stress, menstruation) could influence the serum protein profile of study participants. When designing and analysing their experiments, investigators must also consider the

relative lack of stability of some of the cellular constituents detected by '-omics' techniques, such as RNA degradation and repetitive freezing cycles. Furthermore, the tendency to develop or 'discover' the biomarker patterns using the available data, [17] rather than having a predefined hypothesis as to which biomarkers are likely to be involved, make these studies susceptible to overfitting [15,16] (i.e., the apparent discrimination is due to chance and results cannot be reproduced in other populations). Additionally, '-omics' technologies may be subject to limitations common to all diagnostic research. For example, one common problem in study design is the tendency to collect two groups of patients for discrimination separately (in what can be considered a diagnostic case-control study), instead of prospectively recruiting a group of patients with clinical suspicion of the disease under question, and then using the '-omics' technology to discriminate between patients who are finally diagnosed with the disease and those who are not. [18,19]

Achievement of all legitimate clinical and commercial interests requires that the provision of '-omics'-based diagnostic services be evidence based. [20] Tools for evaluating the quality of diagnostic research reports included in a systematic review, such as QUADAS, [21] have made a considerable impact in promoting evidence based diagnosis. Nevertheless, there is some concern that quality appraisal tools generic to all diagnostic tests may not be sufficiently adequate for this complex field, as such tools do not address the issues specific to the '-omics' field previously mentioned. Consequently, we proposed an adaptation to the QUADAS guideline to take into account the particular challenges presented by '-omics' based technologies. QUADOMICS [22] incorporates four new items addressing the type of sample used, differences in pre-analytical conditions, the clinical and physiological characteristics of the patients providing biological samples, and overfitting. Furthermore, it calls for users to classify each study into one of four phases of biomarker validation, according to the population in which the study is carried out. [23–25] In the first three phases a case control design may be used, and the objective could be to show discrimination between patients with overt disease and healthy individuals, to challenge the test with competing diagnoses, diverse co-morbidities or varying levels of disease severity, or to evaluate changes in diagnostic accuracy according to particular patient characteristics. However, in the fourth phase of evaluation, the test should be evaluated in a prospective series of individuals that reflect, with the maximum degree of fidelity, the clinical or public health setting where the test would be used.The evaluation of study phase was incorporated into QUADOMICS to increase recognition of issues related to the spectrum of patients studied, [26] and the requirements for synthesising results from studies in different phases when performing a meta-analysis. [27,28]

As with any quality appraisal tool, it is essential that QUADOMICS be easy to apply and consistent, i.e., that independent users make analogous observations and judgements when appraising the same study. Accordingly, the primary objective of this study was to evaluate the applicability and consistency of the QUADOMICS tool by applying it to a broad selection of studies in triplicate. An associated secondary objective was the assessment of the methodological quality of the selection of recently published '-omics' diagnostic studies, using this tool.

## Methods

The study consisted of two parts: 1) the evaluation of the applicability and consistency of the QUADOMICS tool, and 2) the evaluation of the methodological quality of a selection of recent published studies. The same selection of studies was used for both parts.

### Search Strategy

We identified original research articles by a systematic search of the Pubmed database combining the medical subject headings (MeSH) "Genomics", "Sensitivity and specificity" and "Diagnosis". The search was limited to articles published from 1st January 2006 through June 17 2009 (the date of the search). The titles and abstracts of all potential articles were reviewed and articles were selected based on the following criteria: original research articles in which the key objective was to evaluate the diagnostic accuracy of an '-omics' based test for use in clinical practice or a screening programme (we used the definition of '-omics' applied in the development of QUADOMICS). [22] Studies which used '-omics' techniques for the discovery of a biomarker pattern but then used standard laboratory techniques such as immunohistochemistry, ELISA or PCR to identify the biomarkers and validate the pattern were not selected. Furthermore, we only selected studies which presented a diagnostic accuracy measurement (e.g., sensitivity and specificity, area under ROC curve, diagnostic odds ratio, likelihood ratios) or that provided enough information for their calculation. Studies in which the main aim was to validate biomarkers for prognostic use or to predict the response to treatment were also excluded, as were articles published in languages other than English.

### Evaluation of the applicability and consistency of the QUADOMICS tool

Three investigators (LP, NG, BL) independently assessed the quality of all selected articles using the QUADOMICS tool. For reference, each reviewer was provided with a copy of the QUADOMICS publication, [22] the development of QUADAS publication [21] and the article evaluating QUADAS and providing some modifications to the items. [29] All three researchers met to compare their observations and generate the consensus rating after 8 articles had been reviewed, after 21, and finally after all 45; any disagreements were solved by discussion. During this process the authors explored the potential motives for the lack of agreement and discussed methods to improve the description of the item in the QUADOMICS guideline in order to avoid future discrepancies.

To evaluate the consistency of the QUADOMICS tool, we calculated the percentage agreement between each reviewer's original assessment and the consensus rating, both overall and for each item separately. We chose not to report Cohen's kappa statistic for inter-rater agreement because it is strongly influenced by the prevalence of each rating and can be misleading. [30] We regarded the consistency as "low" if agreement with the consensus was less than 60% for at least one reviewer, or if two or more reviewers had less than 80% agreement with the consensus. The reasons for limited consistency were evaluated and the item was reworked if necessary.

### Evaluation of the methodological quality of the selected articles

We used the consensus variables created during the evaluation of applicability and consistency of QUADOMICS to describe the methodological quality of the articles. As not all of the items were applied to every article (for instance, some criteria are only applied to articles in phase 4), we summarised the overall quality of each article by calculating the percentage of applied articles which scored positively. Finally, to identify if certain methodological short-comings

were more common than others, we calculated the proportion of articles which met or failed to meet each item separately.

### Data analysis

Univariate descriptive statistics and 95% confidence intervals were computed as customary. [31,32] All computations were carried out using STATA/SE 8.0 (StataCorp, College Station, TX, USA).

## Results and Discussion

The search strategy provided 164 potential articles, of which 59 were selected for full text revision and 45 were finally selected (Figure S1). The references of the 45 selected articles can be found in Annex S1 and a list of the study phase, study size, index test and reference standard of each study is found in Table S1.

### Applicability and consistency of QUADOMICS

Overall, the percentage agreement with the consensus rating was above 80% for all three reviewers (table 1). Of the 17 quality items, up to 4 were not applied to some of the articles. These included items 2 and 14, which should only be applied to studies in phase IV, as directed in the QUADOMICS background document. [22] Additionally, items 9 and 13 were only applied to some articles due to one or both of the following reasons: 1) the index test was almost exclusively performed after the reference diagnosis, and 2) many studies did not have an independent reference standard but, rather, the index test was tested against the diagnosis itself (which was also the criteria used by the authors to select the patients). For example, some studies selected a group of patients with the disease in question and a group of controls, either healthy individuals or with an alternative diagnosis. The lack of an independent reference test is a common problem in studies that seek to validate the diagnostic application of new '-omics' based technologies and it contributed to difficulties in the application of the QUADOMICS items that refer to the reference standard. When possible, we applied these quality items by considering how and when the initial diagnosis was made, or how the diagnosis was ruled out in the controls. We decided that it would be unfair to score studies negatively for all items that

**Table 1.** Consistency in the application of the QUADOMICS tool to 45 diagnostic '-omics' studies: % agreement with the consensus[1].

| | Reviewer 1 | | Reviewer 2 | | Reviewer 3 | |
|---|---|---|---|---|---|---|
| | % | (95%CI) | % | (95%CI) | % | (95%CI) |
| Study Phase | 91.1 | (78.8–97.5) | 97.8 | (88.2–99.9) | 73.3 | (62.9–88.8) |
| 1. Were selection criteria clearly described? | 100 | | 100 | | 100 | |
| 2. Was the spectrum of patients representative of patients who will receive the test in practice? | 95.2 | (84.2–99.4) | 100 | | 97.7 | (87.7–99.9) |
| 3. Was the type of sample fully described? | 86.7 | (73.2–94.5) | 91.1 | (78.8–97.5) | 77.8 | (87.7–99.9) |
| 4. Were the procedures and timing of biological sample collection with respect to clinical factors described with enough detail? | | | | | | |
| 4.1. Clinical and physiological factors | 86.7 | (73.2–94.5) | 68.9 | (53.2–81.4) | 73.3 | (58.1–85.4) |
| 4.2. Diagnostic and treatment procedures. | 88.9 | (75.2–95.8) | 86.7 | (73.2–94.5) | 80.0 | (65.4–90.4) |
| 5. Were handling and pre-analytical procedures reported in sufficient detail and similar for the whole sample? and, if differences in procedures were reported, was their effect on the results assessed? | 64.4 | (48.8–78.1) | 93.3 | (81.7–98.6) | 88.9 | (75.2–95.8) |
| 6. Is the time period between the reference standard and the index test short enough to reasonably guarantee that the target condition did not change between the two tests? | 68.9 | (53.2–81.4) | 84.4 | (70.5–93.5) | 53.3 | (37.9–68.3) |
| 7. Is the reference standard likely to correctly classify the target condition? | 80.0 | (65.4–90.4) | 88.9 | (75.2–95.8) | 64.4 | (48.8–78.3) |
| 8. Did the whole sample or a random selection of the sample receive verification using a reference standard of diagnosis? | 80.0 | (65.4–90.4) | 93.3 | (81.7–98.6) | 73.3 | (58.1–85.4) |
| 9. Did patients receive the same reference standard regardless of the result of the index test? | 80.0 | (65.4–90.4) | 82.2 | (67.9–92.0) | 97.8 | (88.2–99.9) |
| 10. Was the execution of the index test described in sufficient detail to permit replication of the test? | 84.4 | (70.5–93.5) | 77.8 | (87.7–99.9) | 88.9 | (75.2–95.8) |
| 11. Was the execution of the reference standard described in sufficient detail to permit its replication? | 77.8 | (87.7–99.9) | 80.0 | (65.4–90.4) | 62.2 | (46.5–76.2) |
| 12. Were the index test results interpreted without knowledge of the results of the reference standard? | 88.9 | (75.2–95.8) | 91.1 | (78.8–97.5) | 91.1 | (78.8–97.5) |
| 13. Were the reference standard results interpreted without knowledge of the results of the index test? | 88.9 | (75.2–95.8) | 97.8 | (88.2–99.9) | 100 | |
| 14. Were the same clinical data available when test results were i nterpreted as would be available when the test is used in practice? | 97.6 | (87.4–99.9) | 100 | | 100 | |
| 15. Were uninterpretable/intermediate test results reported? | 57.8 | (42.2–72.0) | 93.3 | (81.7–98.6) | 73.3 | (58.1–85.4) |
| 16. Is it likely that the presence of overfitting was avoided? | 73.3 | (58.1–85.4) | 93.3 | (81.7–98.6) | 84.4 | (70.5–93.5) |
| **Overall** | **83.0** | **(80.2–85.5)** | **89.9** | **(87.5–91.9)** | **82.3** | **(79.5–84.9)** |

[1]A consensus rating was achieved by discussion between the three reviewers for every item of each study separately.

doi:10.1371/journal.pone.0011419.t001

mentioned the reference standard as they will not always be subject to the biases addressed by every quality item.

When each item was analysed individually, four items -4.1, 6, 11 and 15- showed a low consistency according to our definition (one reviewer with less than 60% agreement with consensus, or 2+ reviewers with less than 80%). The motives for limited agreement are next discussed individually for each item.

**Item 4.1: Were the procedures and timing of biological sample collection with respect to clinical factors described with enough detail? -Clinical and pathophysiological factors.** There was some disagreement as to what constituted 'enough detail'. Reporting sex and age of the patients in a descriptive table should not be considered sufficient to score positively. Ideally, authors should perform an analysis of the influence of procedures and timing of biological sample collection on the results of the test (example excerpt below). Nevertheless, in this review it was decided that studies scored positively as long as they provided some additional clinical information (apart from sex and age), such as cancer stage. It is advised that, before carrying out a systematic review, the authors discuss what is considered to be 'enough detail'.

Example. Score positively:

*"… was employed to determine whether potentially confounding clinical variables such as patient age, sex, time from transplantation, HCV status, immunosuppressive therapy (…), and peripheral blood monocyte, lymphocyte, and neutrophil counts could be influencing gene-expression levels."* [No. 17 in Annex S1]

**Item 6: Is the time period between the reference standard and the index test short enough to reasonably guarantee that the target condition did not change between the two tests?** As discussed, most studies in '-omics' technologies selected patients with established diagnosis and a control group, and used this classification as the reference standard. Thus, to evaluate disease progression bias [33] one should consider the time period between the initial diagnosis of the established condition and performance of the index test. This item is especially relevant for proteomics-based tests when the biomarker profile may be considerably different at different stages in disease. To score positively the diagnosis should be confirmed at the time of sample collection, and the disease stage should be noted or the time since diagnosis should be stated, so that disease progression bias can be evaluated (example excerpts below). If the authors fail to mention time since diagnosis this item should be marked unclear. If the authors mention time since diagnosis but the reviewer considers it to be too long (refer to QUADAS), [21] this item should be scored as no. If the test is based on a DNA microarray it is unlikely to be affected by the time since diagnosis and so this item will be scored as yes.

Example. Score positively:

*"At the time the sample was taken, all patients were classified by the clinician, according to standard criteria, as having active or inactive renal or systemic lupus."* [No. 22 in Annex S1] *or "The clinical stage distribution of the 132 patients was as follows: stage I (n = 16); stage II (n = 56); stage III (n = 44); and stage IV (n = 16)."* [No. 43 in Annex S1]

Example. Score unclear:

*"Sera from pathologically confirmed lung cancer and benign tobacco-induced or tobacco-associated chronic lung disease patients were collected…"* [No. 12 in Annex S1]

**Item 11: Was the execution of the reference standard described in sufficient detail to permit its replication?** The application of this item was made more complicated by the absence of an independent reference test in many of the studies. We evaluated whether the diagnostic criteria which gave rise to patient selection were described in enough detail. On several occasions, the diagnostic process for the cases with the disease of interest was described in sufficient detail; yet, there was relatively little information relating to how the authors established the absence of disease in the comparison group. Consistency was limited for this item because the reviewers dealt with this situation differently. We recommend that before carrying out a review, the authors discuss firstly whether they want to include studies that use prior diagnoses as the reference diagnosis, and secondly, if they choose to include them, what information should be given as a minimum to rule out the disease in the comparison group.

**Item 15: Were uninterpretable/intermediate test results reported?** We experienced difficulties in evaluating this item as few studies mentioned uninterpretable results. We sought to apply the modification to this item made in the evaluation of QUADAS. *"If the authors do not report any uninterpretable/indeterminate/intermediate results, and if results are reported for all patients who were described as having been entered into the study then this item should also be scored as "yes".* [21] Nevertheless, problems arose because it was difficult to judge if all patients described as having entered into the study contributed to the results presented, as often authors reported the diagnostic accuracy for different biomarker patterns (e.g., different protein peaks), without actually providing the crude patient numbers (example excerpt below). It was agreed that in this case we would mark the item "unclear".

Example. Score positively:

*"…the test group had 52 patients and 33 controls."* → *"Analyses of the spectra from the 85 testing samples showed that the classification algorithm correctly predicted 94% (80 of 85) of all of the samples, with 94% (49 of 52) of DLBCL samples and 94% (31 of 33) of the control samples. The specificity was 94% and the sensitivity was 94%."* [No. 43 in Annex S1]

Example. Score unclear:

*"Cancers (62 samples) and controls (31 samples) were collected into identical tubes and processed in an identical manner."* → *"Varying numbers of the most significant peaks were then used to develop ANNs to discriminate between cancer and non-cancer with 10-fold cross-validation. The ANNs developed using the seven most significant peaks performed best giving a sensitivity of 94% and specificity of 96%."* [No. 37 in Annex S1]

## Quality of selected articles

Out of 45 included articles, 35 were considered to be in phase 1 (78%). Only 6 articles (13.3%) reflected the clinical situation in which the test would be used in practice, phase 4. This finding has important implications given that the case-control design used in phases I-III can lead to an overestimation of diagnostic accuracy. [34,35]

There were 15 (33.3%) studies published in 2008, 13 (28.9%) each in 2006 and 2007, and 4 (8.9%) in 2009.

It is worth mentioning that the main goal in developing QUADOMICS, like QUADAS, was not for assessing the absolute quality in a cross-sectional sample of studies examining different

technologies at different stages in development but, rather, for use in systematic reviews to identify differences in design and conduct that could lead to bias or variation in accuracy within a set of studies examining the same index test. Nevertheless, we have outlined how QUADOMICS can be tailored to suit the different phases of development and in such, any methodological shortcoming highlighted in our analysis was relevant considering the stage of development. Accordingly, up to four items were not applied to some of the selected articles and we evaluated the absolute quality of the studies by calculating the proportion of applied criteria that scored positively.

There was substantial variation in the number of quality criteria met by the selected articles, with one article meeting only 2 of 13 applied criteria (15.4%), [36] and another meeting 12 of 13 applied criteria (92.3%). [37]On average, the selected studies scored positively in just over half of the applied criteria (mean ± standard deviation, 54.7±18.4%). We have reported the percentage of applied criteria which scored positively to summarise the quality of the studies only. We do not believe that a critical threshold should be used when judging study quality [38]. We provide QUADOMICS as a tool that allows systematic reviewers and other readers to identify potential methodological weaknesses in a study, which could have biased the diagnostic accuracy, and therefore judge themselves whether study results are valid. The use of a critical threshold would not appropriately distinguish between a study with a single methodological shortcoming that completely invalidates the results, and a study that does not properly address a number of less influential items.

That being said, the methodological quality of the articles was generally poor, with numerous studies failing to address critical details. This in itself is a relevant finding because high quality studies are imperative if we are to ensure that the application of '-omics' based diagnostic tests to clinical practice is evidence based. To identify the most common methodological short-comings, we explored the proportion of articles that met or failed to meet each item separately (Table 2). The most relevant findings are discussed in more detail below.

**Aspects relating to the patient population and samples studied (Items 1–5).** In general, the description of the sample population was poor and none of the articles scored positively for item 1 due to the absence of a flow diagram describing the flow of patients in the selection process. The limited description of the patient population observed in these studies was disconcerting as this information is essential in order to assess external validity. Interestingly, even one of the phase 4 studies, scored negatively for the item on patient spectrum (item 2, example excerpt below). This study sought to validate a proteomics based urine test for the diagnosis of ovarian cancer. [39] Although it was considered to be phase 4 due to the inclusion of a consecutive series of patients, it is likely that by selecting women undergoing surgery the study selected a more severely diseased patient population than would normally receive the urine based test:

**Table 2.** Evaluation of the methodological quality of 45 diagnostic '-omics' studies using the QUADOMICS tool.

| Item | Yes | (%) | No | (%) | Unclear | (%) | N/A | (%) |
|---|---|---|---|---|---|---|---|---|
| 1. Were selection criteria clearly described? | 0 | — | 45 | (100) | 0 | — | 0 | — |
| 2. Was the spectrum of patients representative of patients who will receive the test in practice? | 4 | (8.9) | 1 | (2.2) | 1 | (2.2) | 39 | (86.7) |
| 3. Was the type of sample fully described? | 40 | (88.9) | 4 | (8.9) | 1 | (2.2) | 0 | — |
| 4. Were the procedures and timing of biological sample collection with respect to clinical factors described with enough detail? | | | | | | | | |
| 4.1. Clinical and physiological factors | 20 | (44.4) | 25 | (55.6) | 0 | — | 0 | — |
| 4.2. Diagnostic and treatment procedures. | 22 | (48.9) | 22 | (48.9) | 1 | (2.2) | 0 | — |
| 5. Were handling and pre-analytical procedures reported in sufficient detail and similar for the whole sample? and, if differences in procedures were reported, was their effect on the results assessed? | 38 | (84.4) | 7 | (15.6) | 0 | — | 0 | — |
| 6. Is the time period between the reference standard and the index test short enough to reasonably guarantee that the target condition did not change between the two tests? | 20 | (44.4) | 1 | (2.2) | 24 | (53.3) | 0 | — |
| 7. Is the reference standard likely to correctly classify the target condition? | 33 | (73.3) | 6 | (13.3) | 6 | (13.3) | 0 | — |
| 8. Did the whole sample or a random selection of the sample receive verification using a reference standard of diagnosis? | 24 | (53.3) | 14 | (31.1) | 7 | (15.6) | 0 | (0.0) |
| 9. Did patients receive the same reference standard regardless of the result of the index test? | 1 | (2.2) | 0 | — | 0 | — | 44 | (97.8) |
| 11. Was the execution of the reference standard described in sufficient detail to permit its replication? | 21 | (46.7) | 24 | (53.3) | 0 | — | 0 | — |
| 12. Were the index test results interpreted without knowledge of the results of the reference standard? | 20 | (44.4) | 25 | (55.6) | 0 | — | 0 | — |
| 13. Were the reference standard results interpreted without knowledge of the results of the index test? | 6 | (13.3) | 0 | — | 0 | — | 39 | (86.7) |
| 14. Were the same clinical data available when test results were interpreted as would be available when the test is used in practice? | 5 | (11.1) | 1 | (2.2) | 0 | — | 39 | (86.7) |
| 15. Were uninterpretable/intermediate test results reported? | 33 | (73.3) | 2 | (4.4) | 10 | (22.2) | 0 | — |
| 16. Is it likely that the presence of overfitting was avoided? | 20 | (44.4) | 22 | (48.9) | 3 | (6.7) | 0 | — |

doi:10.1371/journal.pone.0011419.t002

Example:

*"Urine samples and paired blood samples were prospectively collected from 209 consecutive women admitted for an exploratory laparotomy for an ovarian neoplasm at the Gynaecological Department at Rigshospitalet, Copenhagen between June 2006 and August 2007."* [No. 25 in Annex S1]

Only half of the studies considered the diagnostic or treatment procedures undergone by the patient before the sample was taken (Item 4.2: 22, 49.9%), and even fewer described the clinical and pathophysiological factors that might influence the biomarker profile [13,14] (Item 4.1: 20 studies, 44.4%). Most articles clearly described the type of sample used and the pre-analytical procedures in sample preparation (Item 3: 40, 88.9%, Item 5: 38, 84.4%).

**Aspects relating to the test being evaluated (Items 10, 13, 14).** 19 (42.2%) studies did not describe the index test in enough detail (Item 10). Less than half of the studies (Item 13: 20, 44.4%) mentioned whether the index test result was interpreted without knowledge of the reference standard; such omission suggests that review bias was possible. [19,26] On the other hand, one of the phase 4 studies was subject to a kind of over blinding, and scored negatively in item 14 (example excerpt below). This study evaluated a gene expression profile for the identification of the tissue of origin in the case of metastatic, poorly differentiated specimens. [40] Although blinding of the reference diagnosis is necessary to avoid review bias, in clinical practice the clinician interpreting the test would have access to details such as patient sex and tumour pathology.

Example. *"… investigators who interpreted the Pathwork Tissue of Origin Test results for making a tissue determination were blinded to patient sex, histology, or morphology information, and reference diagnosis"* [No. 21 in Annex S1]

**Aspects relating to the reference test (Items 6, 11).** Over half of the articles did not describe the reference test in enough detail (Item 11: 21, 46.7%). As mentioned earlier many of the articles did not actually include an independent reference test. In this case we evaluated the diagnosis of the target condition or selection criteria for the comparison group. Furthermore, over half of the articles failed to mention any time period with regard to diagnosis, making it difficult to judge whether the target condition could have changed (item 6: 24, 53.3% unclear).

**Overfitting (Item 16).** 22 (48.9%) studies did not effectively control for overfitting, and in 3 studies (6.7%) it was not clear if validation was carried out in samples from the same patients in which the model was built. Only studies that validated their biomarker signature in an independent set of patient samples scored positively for this item; i.e., studies that performed internal validation using cross validation alone did not score positively. We deem this an important finding because it is likely that the results presented in these studies are overly optimistic [41] and may not be reproducible in other patient populations. [42]

Finally, there was no apparent change in the proportion of studies meeting each item separately over the 4 years studies (data not shown), but numbers were small.

## Conclusions

In this study we showed that three reviewers could apply the QUADOMICS tool to a broad sample of diagnostic '-omics' studies with reasonable consistency. A small number of items were difficult to apply to studies that did not use an independent test for determining the reference diagnosis. This problem with item applicability arose in studies which used a healthy or alternative diagnosis comparison group and, thus, it was closely linked to the study phase of the articles (phases I–III). On one hand, the importance of this problem is limited because systematic reviews and meta-analyses carried out to inform decision makers of the evidence supporting the use of a test in clinical practice should focus on studies with more clinically relevant populations (phase IV). On the other hand, it is highly important that the quality of early phase studies is adequately assessed in order to weigh up the evidence and decide if it is a sensible use of resources to proceed to studies in more clinically relevant populations. Here, we have outlined how the QUADOMICS criteria can be applied to these studies.

In practice the QUADOMICS guideline will be used to evaluate studies included in a systematic review and, therefore, studies should all be addressing the same diagnostic question, and be in the same phase. Similar to QUADAS, [21] reviewers should tailor the guideline to suit their specific review question. For example, if they want to assess the utility of the test for use in clinical practice, they should only include phase IV studies, and make some decisions before evaluating the studies (e.g., what should be the appropriate reference standard, how much information is considered to be 'sufficient detail' or how long is too long for the time period between reference and index test). On the other hand, a review carried out to assess the preliminary evidence in favour of a new '-omics' test in order to judge whether it would be sensible or appropriate to carry out a large scale prospective evaluation may include studies from earlier phases which use the case-control type design. While it would extremely important to consider differences between the two diagnostic groups with regard to pre-analytical conditions (item 5), or the clinical characteristics of the patients providing samples (item 4), it would be inappropriate to score a study negatively because it does not meet item 2 ('Was the spectrum of patients representative of patients who will receive the test in practice?'). In this case the tailoring of the guideline would involve eliminating the items that are not applicable as well as making decisions as how specific items should be scored. By applying QUADOMICS to a broad range of articles from different subjects, we have shown that it is flexible, and we believe that the ability to be tailored to the different study phases is one of its key strengths.

The methodological quality of our selection of 45 '-omics' based diagnostic studies was poor. It is alarming, for example, that none of the studies included a flow diagram describing the patient recruitment process; such diagrams are also strongly recommended in the Standards for Reporting of Diagnostic Accuracy (STARD) publication. [43] This deficiency is not specific to the '-omics' field; for instance, a recent review of commercial tests for HIV, TB or malaria showed that only 13% of studies reviewed met the STARD criterion which recommends the flow diagram. [44] This issue is in fact a reporting item and therefore only indirectly linked to quality. Studies that meet this criterion do not automatically have clinically relevant populations, yet in studies that do not clearly describe patient recruitment it is impossible to evaluate whether the results are applicable to our context. It is arguable that reporting items have no place in instruments measuring methodological quality however, despite increased sensitisation to issues related to the quality of reporting, diagnostic research remains poorly reported [45] and evaluating methodological quality relies on transparent and good quality reporting. In such we feel that such items do help draw attention of the readers to potential methodological limitations, and thus reduce assumptions that the methodology was sound.

There were other threats to the validity of the studies. For instance, it is now recognised that patient treatment regimes or other clinical and pathophysiological characteristics may influence the parameters studied, such as proteins, and thus bias '-omics' studies. [13,14,46,47] Nevertheless, few of the studies we assessed actually reported these details, let alone analysed their potential effect. Furthermore, in nearly half of the articles the diagnostic model was not validated in an independent set of patients; such shortfall may lead to overfitting and the production of results that are not reproducible. Coupled with the fact that very few of the studies were actually carried out in a consecutive set of patients with clinical suspicion of the disease in question, the problem illustrates the relative lack of attention paid in '-omics' research to design issues that are fundamental when we aim at making inferences relevant for patient care.

One limitation of this study is the external validity of our assessment of the quality of recent articles published in this field, our secondary objective. We do not presume to have included all diagnostic '-omics' studies published in 2006 through 2009. While our sample was not restricted to any particular field or technique, it is clear that it was limited to reports indexed by Medline, and adequately tagged with the selected MeSH terms. Nevertheless for our primary objective, we feel that the selected sample was sufficiently diverse to adequately assess the applicability and consistency of the QUADOMICS tool.

Another issue is related to the three reviewers used to evaluate the consistency and applicability of QUADOMICS. While the three reviewers had different backgrounds and varying levels of research experience, in principle it would have been beneficial to include a larger number of reviewers with a wider knowledge of the diseases of interest. Furthermore, two of the three observers were involved in the development of the tool, and hence may have found the tool easier to apply. However, in practice QUADOMICS will be used to evaluate the quality of studies addressing the same diagnostic question and reviewers will decide a priori how each item should be scored. In such situations it is likely that application would be more straightforward and that reviewer observations would be more consistent. Here we provide an evaluation of the tool in general, rather than for every subject separately, because at this stage in the development of QUADOMICS, we felt it was important to ensure the tool was applicable to a broad range of real studies.

For ethical, clinical and economic reasons, the application of '-omics' based tests in clinical practice requires valid and reliable research that can be reproduced in clinically relevant patient populations. [23–25] While some of the methodological deficiencies we described were linked to the specific peculiarities of '-omics' based research, other important aspects -which have long been considered fundamental in traditional diagnostic research, such as the description of the index test and test reproducibility- are being overlooked in '-omics' research. The QUADOMICS tool was proposed for the assessment of the methodological quality of diagnostic research using '-omics' based technology. [22] We show that the tool can consistently be applied to a broad range of these studies. Furthermore, we hope that it will help sensitize researchers, clinicians and other decision makers to the serious threats to the validity inherent to this type of research, and ensure that the provision of '-omics' tests to the clinic is evidence based.

## Supporting Information

**Figure S1** Flow diagram of search and selection process.
Found at: doi:10.1371/journal.pone.0011419.s001 (0.26 MB TIF)

**Annex S1** References of the 45 articles evaluated.
Found at: doi:10.1371/journal.pone.0011419.s002 (0.04 MB DOC)

**Table S1** Characteristics of 45 studies evaluating the diagnostic use of an '-omics' based test.
Found at: doi:10.1371/journal.pone.0011419.s003 (0.10 MB DOC)

## Author Contributions

Conceived and designed the experiments: BL IHA. Analyzed the data: LAP. Wrote the paper: LAP. Data acquisition: LAP NGS BL. Data interpretation: LAP NGS BL MP IHA. Critical review of manuscript: NGS BL MP IHA. Final approval of manuscript: NGS BL MP IHA.

## References


1. Ghosh D, Poisson LM (2009) "Omics" data and levels of evidence for biomarker discovery. Genomics 93: 13–16.
2. Thomas DC (2006) High-volume "-omics" technologies and the future of molecular epidemiology. Epidemiol 17: 490–491.
3. Finn WG (2007) Diagnostic pathology and laboratory medicine in the age of "omics": a paper from the 2006 William Beaumont Hospital Symposium on Molecular Pathology. J Mol Diagn 9: 431–436.
4. Negm RS, Verma M, Srivastava S (2002) The promise of biomarkers in cancer screening and detection. Trends Mol Med 8: 288–293.
5. Check E (2004) Proteomics and cancer: running before we can walk? Nature 429: 496–497.
6. Diamandis EP (2007) Oncopeptidomics: A useful approach for cancer diagnostics? Clin Chem 53: 1004–1006.
7. Ioannidis JP (2007) Is molecular profiling ready for use in clinical decision making? Oncologist 12: 301–311.
8. Lumbreras B, Porta M, Marquez S, Pollán M, Parker LA, et al. (2009) Sources of error and its control in studies on the diagnostic accuracy of "-omics" technologies. Proteomics Clin Appl 3: 173–184.
9. Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, et al. (2002) Use of proteomic patterns in serum to identify ovarian cancer. Lancet 359: 572–577.
10. Wagner L (2004) A test before its time? FDA stalls distribution process of proteomic test. J Natl Cancer Inst 96: 500–501.
11. Baggerly KA, Morris JS, Coombes KR (2004) Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. Bioinformatics 20: 777–785.
12. Ransohoff DF (2005) Bias as a threat to the validity of cancer molecular-marker research. Nat Rev Cancer 5: 142–149.
13. Porta M, Pumarega J, Ferrer-Armengou O, López T, Alguacil J, et al. (2007) Timing of blood extraction in epidemiologic and proteomic studies: Results and proposals from the PANKRAS II Study. Eur J Epidemiol 22: 577–588.
14. Porta M, Pumarega J, López T, Jariod M, Marco E, et al. (2009) Influence of tumor stage, symptoms and time of blood draw on serum concentrations of organochlorine compounds in exocrine pancreatic cancer. Cancer Causes Control 20: 1893–1906.
15. Baggerly KA, Morris JS, Edmonson SR, Coombes KR (2005) Signal in noise: evaluating reported reproducibility of serum proteomic tests for ovarian cancer. J Natl Cancer Inst 97: 307–309.
16. Simon R, Radmacher MD, Dobbin K, McShane LM (2003) Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst 95: 14–8.
17. Carrol S, Goodstein D (2009) Defining the scientific method. Nat Methods 6: 237.
18. Lumbreras B, Parker LA, Porta M, Pollán M, Ioannidis JPA, et al. (2009) Overinterpretation of clinical applicability in molecular diagnostic research. Clin Chem 55: 786–794.
19. Haynes RB, Sackett DL, Guyatt GH, Tugwell P (2006) Clinical epidemiology.How to do clinical practice research. 3rd. ed. Philadelphia: Lippincott, Williams & Wilkins.
20. Ransohoff DF (2007) How to improve reliability and efficiency of research about molecular markers: roles of phases, guidelines, and study design. J Clin Epidemiol 60: 1205–1219.
21. Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PMM, Kleijnen J (2003) The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol 3: 25–37.

22. Lumbreras B, Porta M, Márquez S, Pollán M, Parker LA, et al. (2008) QUADOMICS: An adaptation of the Quality Assessment of Diagnostic Accuracy Assessment (QUADAS) for the evaluation of the methodological quality of studies on the diagnostic accuracy of '-omics' based technology. Clin Biochem 41: 1316–1325.
23. Feinstein AR (1985) Clinical epidemiology: the architecture of clinical research, Philadelphia: WB Saunders.
24. Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson ML, et al. (2001) Phases of biomarker development for early detection of cancer. J Natl Cancer 93: 1054–1061.
25. Sackett DL, Haynes RB (2002) Evidence base of clinical diagnosis: The architecture of diagnostic research. BMJ 324: 539–541.
26. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, et al. (2004) Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med 140: 189–202.
27. Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, et al. (1994) Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med 120: 667–676.
28. Tatsioni A, Zarin DA, Aronson N, Samson DJ, Flamm CR, et al. (2002) Challenges in systematic reviews of diagnostic technologies. Ann Intern Med 142: 1048–1055.
29. Whiting PF, Weswood ME, Rutjes AW, Reitsma JB, Bossuyt PN, et al. (2006) Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. BMC Med Res Methodol 6: 9–16.
30. Lantz CA, Nebenzahl E (1996) Behavior and interpretation of the κ statistic: Resolution of the two paradoxes. J Clin Epidemiol 49: 431–434.
31. Armitage P, Berry G, Matthews JNS (2002) Statistical methods in medical research. 4th edition. Oxford: Blackwell.
32. Kleinbaum DG, Kupper LL, Muller KE (1998) Applied regression analysis and other multivariable methods. 3rd edition. Pacific Grove, CA: Duxbury.
33. Porta M, ed (2008) A dictionary of epidemiology. 5th edition. New York: Oxford University Press 69,226.
34. Lijmer JG, Mol BW, Heistekamp S, Bonsel GJ, Prins MH, et al. (1999) Emprirical evidence of design-related bias in studies of diagnostic tests. JAMA 282: 1061–1066.
35. Rutjes AWS, Reitsma JB, DiNisio M, Smidt N, van Rijn JC, et al. Evidence of bias and variation in diagnostic accuracy studies. CMAJ 176: 469–476.
36. Pasinetti GM, Unger LH, Lange DJ, Yemul S, Deng H, et al. (2006) Identification of potential CSF biomarkers in ALS. Neurology 66: 1218–1222.
37. Belluco C, Petricoin EF, Mammano E, Facchiano F, Ross-Rucker S, et al. (2007) Serum proteomic analysis identifies a highly sensitive and specific discriminatory pattern in stage 1 breast cancer. Ann Surg Oncol 14: 2470–2476.
38. Whiting P, Harbord R, Kleijnen J (2005) No role for quality scores in systematic reviews of diagnostic accuracy studies. BMC Med Res Methodol 5: 19.
39. Petri AL, Simonsen AH, Yip TT, Hogdall E, Fung ET, et al. (2009) Three new potential ovarian cancer biomarkers detected in human urine with equalizer bead technology. Acta Obstet Gynecol Scand 88: 18–26.
40. Monzon FA, Lyons-Weiler M, Buturovic LJ, Rigl CT, Henner WD, et al. (2009) Multicenter validation of a 1,550-gene expression profile for identification of tumor tissue of origin. J Clin Oncol 27: 2503–2508.
41. Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KGM (2003) Internal and external validation of predictive models: A simulation study of bias and precision in small samples. J Clin Epidemiol 56: 441–447.
42. Taylor JM, Ankerst DP, Andridge RR (2008) Validation of biomarker-based risk prediction models. Clin Cancer Res 14: 5977–5983.
43. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, et al. (2003) Standards for Reporting of Diagnostic Accuracy. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Clin Chem 49: 7–18.
44. Fontela PS, Pant Pai N, Schiller I, Dendukuri N, Ramsay A, et al. (2009) Quality and reporting of diagnostic accuracy studies in TB, HIV and malaria: evaluation using QUADAS and STARD standards. PLoS One 4: e7753.
45. Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Bossuyt PM, et al. (2006) The quality of diagnostic accuracy studies since the STARD statement: has it improved? Neurology 67: 740–741.
46. Porta M, Ferrer-Armengou O, Pumarega J, López T, Crous-Bou M, et al. (2008) Exocrine pancreatic cancer clinical factors were related to timing of blood extraction and influenced serum concentrations of lipids. J Clin Epidemiol 61: 695–704.
47. Hoppin JA, Tolbert PE, Taylor JA, Schroeder JC, Holly EA (2002) Potential for selection bias with tumor tissue retrieval in molecular epidemiology studies. Ann Epidemiol 12: 1–6.

The supporting information from this article can be found in Annex 2 of the thesis.

UNIVERSITAS
Miguel
Hernández

## 5.3 Resumen de los hallazgos principales del artículo 3

(Referencia: Lumbreras B, Parker LA, Porta M, Pollán M, Ioannidis JPA, Hernández-Aguado I. Overinterpretation of clinical applicability in molecular diagnostic research. Clin Chem 2009;55:786-794.)

**Sobre-interpretación de la aplicabilidad clínica de la investigación diagnóstica con métodos moleculares.**

Se identificaron 108 estudios diagnósticos de pruebas basadas en métodos moleculares. 82 (76%) de ellos utilizaron un diseño preliminar con el reclutamiento de casos de la enfermedad en cuestión y su comparación con controles sanos o controles con un diagnóstico alternativo. Solo 15 (11%) estudios utilizaron una población considerada clínicamente relevante, es decir que reclutaron una serie de pacientes con las mismas características de aquellas poblaciones que recibirán la prueba en la práctica.

Con respecto a las conclusiones de los autores, 104 estudios (96%) fueron definitivamente favorables o moderadamente favorables a la aplicabilidad clínica de la prueba estudiada. Se consideraba que en 61 (56%) de los estudios estas conclusiones representaban sobreinterpretación debido al uso de un diseño preliminar, la falta de reconocimiento de la necesidad de estudios adicionales y, en algunos casos, la pobre exactitud diagnóstica.

Había varias variables que se asociaron con sobreinterpretación. Por ejemplo, el factor de impacto de la revista de publicación; los estudios publicados en revistas con mayor impacto mostraron más tendencia a sobreinterpretar sus resultados con respecto a artículos publicados en revistas de menor impacto. En un análisis multivariable hecho por cuartiles del factor de impacto, se encontró que cada incremento en cuartil aumentó la probabilidad de sobreinterpretación en 1,7 (1,1-2-7).

Asimismo, la sobreinterpretación de la aplicabilidad clínica fue más común en artículos escritos por autores procedentes del ámbito de laboratorio comparados con aquellos procedentes del ámbito clínico (OR ajustado 18,7 IC95% 1,4-249,3).

Los datos suplementarios para el artículo se encuentran en el anexo 3, y en el anexo 4 se encuentra un extracto del periódico *El Mundo*, del 9 abril 2009 que refiere al artículo.

Clinical Chemistry 55:4
786–794 (2009)



# Overinterpretation of Clinical Applicability in Molecular Diagnostic Research

Blanca Lumbreras,[1] Lucy A. Parker,[1*] Miquel Porta,[2] Marina Pollán,[3] John P.A. Ioannidis,[4] and Ildefonso Hernández-Aguado[1]

**BACKGROUND:** We evaluated whether articles on molecular diagnostic tests interpret appropriately the clinical applicability of their results.

**METHODS:** We selected original-research articles published in 2006 that addressed the diagnostic value of a molecular test. We defined overinterpretation of clinical applicability by means of prespecified rules that evaluated study design, conclusions regarding applicability, presence of statements suggesting the need for further clinical evaluation of the test, and diagnostic accuracy. Two reviewers independently evaluated the articles; consensus was reached after discussion and arbitration by a third reviewer.

**RESULTS:** Of 108 articles included in the study, 82 (76%) used a design that used healthy controls or alternative-diagnosis controls, only 15 (11%) addressed a clinically relevant population similar to that in which the test might be applied in practice, 104 articles (96%) made definitely favorable or promising statements regarding clinical applicability, and 61 (56%) of the articles apparently overinterpreted the clinical applicability of their findings. Articles published in journals with higher impact factors were more likely to overinterpret their results than those with lower impact factors (adjusted odds ratio, 1.71 per impact factor quartile; 95% CI, 1.09–2.69; $P = 0.020$). Overinterpretation was more common when authors were based in laboratories than in clinical settings (adjusted odds ratio, 18.7; 95% CI, 1.41–249; $P = 0.036$).

**CONCLUSIONS:** Although expectations are high for new diagnostic tests based on molecular techniques, the majority of published research has involved preclinical phases of research. Overinterpretation of the clinical applicability of findings for new molecular diagnostic tests is common.



With the remarkable advances in genomic and proteomic technologies, a large number of studies on new molecular diagnostic tests are being published. Expectations are high for the development of noninvasive molecular diagnostic tests, yet analysis and interpretation of the data have presented unique challenges *(1)*. Few of the many proposed tests have been introduced into clinical practice with clearly documented benefits *(2–4)*. Today, more than ever, intense promotion of molecular-diagnostic techniques strengthens the need to ensure that the provision of diagnostic tests in clinical settings is evidence-based; however, offering guidance for the introduction of a new diagnostic test into clinical practice remains a challenge *(5)*. Besides the increased sensitivity to issues of reporting *(6)* and quality assessment *(7)*, several authors *(8–10)* have proposed a formal structure to guide the process of diagnostic-test development.

In the path toward a successful clinical application, a diagnostic test should be evaluated in distinct populations that are similar to those in which the test is intended for eventual use (in clinical practice or in public health). Although preliminary studies may evaluate the ability of the test to distinguish between known disease cases and control individuals who are either healthy or have a specific, different diagnosis, excellent results in the preliminary, preclinical phases do not prove clinical utility. Application of a test in the real world usually involves a different spectrum of disease than preliminary studies, because real-life diagnostic investigations tend to address primarily patients suspected of the target condition and not patients with

[1] Public Health Department, Miguel Hernández University, Alicante, Spain [CIBER en Epidemiología y Salud Pública (CIBERESP)]; [2] Institut Municipal d'Investigació Mèdica, Facultat de Medicina, Universitat Autònoma de Barcelona, Spain [CIBER en Epidemiología y Salud Pública (CIBERESP)]; [3] Cancer and Environmental Epidemiology Area, National Centre for Epidemiology, Instituto de Salud Carlos III, Madrid, Spain [CIBER en Epidemiología y Salud Pública (CIBERESP)]; [4] Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece.
* Address correspondence to this author at: Public Health Department, Miguel Hernández University, E-03550 Alicante, Spain. Fax +34 965 919551; e-mail lparker@umh.es.
Received November 26, 2008; accepted January 30, 2009.
Previously published online at DOI: 10.1373/clinchem.2008.121517

severe clear-cut disease or obviously healthy people. Moreover, other, competing diagnoses are prevalent in real life, whereas most healthy control– or alternative diagnosis–control studies typically exclude patients with diagnoses that compete in the differential diagnosis. Analytical issues (e.g., reproducibility) *(11, 12)* and potential biases *(13)* may also complicate the transition from discovery to clinical translation *(1)*. Although these conceptual and methodologic requirements have long been established, it is unknown whether the new generations of studies on molecular diagnostic tests recognize and integrate the extra requirements for clinical translation or, by contrast, whether they tend to overinterpret or exaggerate preliminary results as providing conclusive evidence for clinical applicability.

Our aim was to analyze a large sample of recent articles on molecular-diagnostic tests to determine whether the authors' assessment of the clinical applicability of their results was coherent with their study design and findings or whether they overinterpreted the clinical significance of the available information.

## Materials and Methods

### DATA SOURCES AND SEARCHING

We identified diagnostic-accuracy studies on molecular research through a computerized search of MEDLINE that used the medical subject headings (MeSH): "Diagnosis" and "Genomics" or "Microarray analysis"; "Molecular diagnostic techniques" (MeSH) and "Sensitivity and Specificity" (MeSH); "diagnos*" and "genomics" or "proteomics"; and finally, "molecular" or "genetic" and "diagnostic test." The searches were carried out on May 11, 2007. The full search strategy is documented in Fig. 1 in the Data Supplement that accompanies the online version of this article at http://www.clinchem.org/content/vol55/issue4.

### STUDY SELECTION

We selected original research articles that used human participants in studies in which the main objective was to address the diagnostic value of a given test whose methodology was based on molecular techniques. The term "molecular techniques" included technologies that provide a comprehensive analysis of cellular-specific constituents, such as RNA, DNA, proteins, and intermediary metabolites, as well as techniques such as in situ hybridization of chromosomes for cytogenetic analysis, identification of pathogenic organisms via analysis of species-specific DNA sequences, and detection of mutations with the PCR. To maintain a focus on recent research, we limited our sample to articles published in 2006.

A single investigator screened the titles and abstracts according to specific criteria. Reviews, editorials, letters, and case reports were excluded. We also excluded preevaluation studies that focused on the analytical aspects of a diagnostic test (technical aspects on how a method is applied or how measurements are made) and studies that aimed to monitor disease prognosis or treatment effects.

To assess the reliability of the selection process, 2 investigators independently assessed a random sample of 200 abstracts; they agreed with the initial reviewer 94% and 83% of the time.

### DATA EXTRACTION AND DEFINITIONS

Two investigators independently extracted data from each article. The data extractors assigned each study to one of 3 following study designs according to previous definitions *(14)*: (*a*) healthy-control or alternative diagnosis–control study; (*b*) consecutive series or series of clinically relevant patients in which the spectrum of patients/samples reflects, as closely as possible, populations in which the test may be used in practice; and (*c*) studies that could not be assigned with confidence to either of the 2 other groups. Table 1 details the operational definitions for each type of design. Furthermore, all statements in the articles referring to clinical applicability and potential need for further clinical evaluation were recorded, as follows:

- *Statements regarding clinical applicability of the test*. Statements on clinical applicability were graded as definitely favorable, as promising, or as unfavorable. Conditional language such as "may" was considered as promising; however, if the authors affirmed that a study reflected the clinical evaluation of the test under question or that the test could be considered an option for diagnosis, it was marked as definitely favorable. The final weight of the decision regarding overinterpretation was based in the abstract.
- *Statements regarding further clinical evaluation of the test*. The presence or absence of statements regarding the need for further clinical evaluation was recorded for each study. A distinction was made between studies that mentioned further clinical evaluation as a desirable possibility and those that stated clinical evaluation was necessary. Only the latter were considered to "mention need of further clinical evaluation."

We defined overinterpretation of clinical applicability with the following rules, which were agreed upon up front and evaluated in a pilot study of 10 articles to ensure that they were operational (Table 2). In brief, overinterpretation was defined in studies with healthy or alternative-diagnosis controls when authors gave a conclusion that was definitely favorable for the appli-

**Table 1. Rules for classification of study designs of molecular-diagnostic studies.**

| Study design | Description |
|---|---|
| Consecutive series or patient series based on a clinically relevant population | Consecutively enrolled patients with clinical suspicion of disease |
| | Individuals presenting at a specific center or group of centers who have symptoms indicative of the disease in question |
| | Consecutive samples sent to diagnostic lab for analysis and possible diagnosis of the disease in question |
| | In screening, when participants share the same characteristics as target population (e.g., asymptomatic "at risk" individuals) |
| Healthy control or alternative-diagnosis control | Clear selection of disease-positive cases and healthy controls |
| | Diseased tissue and healthy adjacent tissue from same patient |
| | The same patient is tested before and after treatment/surgery is performed |
| | Analysis of amplified spectrum of cases and controls (e.g., severe disease, mild disease, benign disease, healthy controls) |
| | Selection of large variety of controls that might pose a diagnostic challenge (but still compared with definitely disease-positive cases) |
| | Studies stating "consecutive series or patient series," yet results clearly indicating that investigators used a healthy-control or alternative diagnosis–control study (e.g., include a healthy control group) |
| Other | Studies that do not follow a healthy-control or alternative diagnosis–control design, but it is not clearly evident that investigators use consecutive series or patient series based on a clinically relevant population. |

cation of the test to the clinic (with or without mentioning the requirement of further clinical evaluation), or if authors stated that the assessed test was promising but did not mention the need for further clinical evaluation. In studies including patient series, any statement in a study that concluded that the test had clinical applications was classified as overinterpretation if the study had unacceptable diagnostic accuracy, as follows: Both sensitivity and specificity were <60% in the main analysis; either sensitivity or specificity was <50% in the main analysis without justification of the merits of the test as an exclusion/inclusion test; the lower limits of the CIs of both sensitivity and specificity were <50%; the area under the ROC curve was <0.55 or had CIs that reached to <0.50; or, an accuracy index was absent, along with insufficient information provided to calculate sensitivity or specificity.

Transcriptions of a selection of the articles examined and their classifications are provided in Annex 1 of the online Data Supplement for illustrative purposes, and some detailed examples are described in the Results. The degree of observer agreement regarding the presence or

**Table 2. Rules for the assessment of overinterpretation.**

| Study design | Overinterpretation | Not overinterpretation |
|---|---|---|
| Consecutive series or patient series based on a clinically relevant population | Definitely favorable comments regarding clinical application of a test with unacceptable diagnostic accuracy | Definitely favorable, promising, or unfavorable comments regarding the clinical applicability of a test evaluated with acceptable diagnostic accuracy |
| | Promising statements regarding clinical application of a test with unacceptable diagnostic accuracy, but *without* mentioning the need for further clinical evaluation | Promising statements regarding clinical application of a test with unacceptable diagnostic accuracy, but *with* statement mentioning the need for further clinical evaluation |
| Healthy control or alternative-diagnosis control | Definitely favorable comments regarding clinical application of the test under study | Unfavorable comments regarding clinical application |
| Other | Promising statements regarding clinical application, but *without* mentioning the need for further clinical evaluation | Promising statements regarding clinical application, but *with* statement mentioning the need for further clinical evaluation |

absence of overinterpretation was 79% at this stage. Discrepancies were resolved by consensus and by independent review by a third investigator. The reviewers were aware of the journal source and authorship.

From each study we also recorded the following variables: Thomson Reuters' bibliographic impact factor; journal categories selected by Thomson Reuters' Web of Science (Journal Citation Reports 2006); whether the authors were based in a laboratory, in a clinical setting, or both; the disease studied; the molecular methodology used, categorized as gene-targeting techniques (PCR-based and microarray), protein-targeting techniques (mass spectrometry or 2-dimensional gel electrophoresis, antibody array or protein microarray), and other; mention of previous studies on the same test and how the results were reported; and description of other diagnostic tests for the same diagnostic problem. We also recorded the sample size; in proteomic or genomic studies in which a pattern-recognition model is developed in a training set and then applied in an independent "validation" set *(13)*, we recorded only the number of patients/samples included in the validation set.

#### STATISTICAL ANALYSIS

To assess the association between the outcome variable (overinterpretation) and the variables listed in the previous paragraph, we computed odds ratios and their 95% CIs by means of unconditional logistic regression. Multivariable models considered all variables with $P$ values $<0.10$ in univariate analyses and used stepwise forward selection. We always included study design and accuracy index as adjusting factors in the multivariable analysis, because they were included in the criteria for judging overinterpretation (as discussed above) and because they could be related with other study characteristics, thus acting as classic confounders. Study size and bibliographic impact-factor data were categorized in quartiles. Analyses were carried out with STATA/SE 8.0 (StataCorp).

## Results

#### EVALUATED ARTICLES

After screening the titles and abstracts of 1614 articles retrieved in the electronic searches, we considered 147 articles potentially eligible for the study after reviewing the abstracts. After examination of the full texts, we ultimately included 108 articles (see Annex 2 and Flowchart in the online Data Supplement).

Table 3 lists the characteristics of the sample of 108 reports. Most of the included reports (83%) used a healthy-control or alternative diagnosis–control design to assess diagnostic accuracy. Regarding the measurement of diagnostic accuracy, more than half (n = 58) of the studies reported classic diagnostic indexes (sensitivity and specificity, or area under the ROC curve). We presented sensitivity and specificity in the same category as area under the ROC curve because 9 of the 12 studies that reported area under the ROC curve presented it along with sensitivity and specificity values; however, when we separately analyzed the 3 studies that reported only area under the ROC curve, we obtained similar results. The sample size ranged from 4 to 8156, with a median of 68.

Thirty-one reports (29%) mentioned previous studies on the same tests; of these 31 reports, 15 quantitatively described the results of the previous studies. More than two thirds (n = 75) of the studies mentioned the existence of other diagnostic tests for the same diagnostic problem. Approximately half (n = 53, 49%) of the reports stated the need for studies other than diagnostic evaluations, such as identification of biomarkers or assessment of prognostic value.

#### OVERALL STANCE AND INTERPRETATION OF THE RESULTS

Half (n = 54, 50%) of the articles studied made definitely favorable statements with regard to clinical application, whereas 50 studies (46%) made statements that were classified as promising. Only 4 studies made unfavorable statements regarding the evaluated diagnostic test. About half (n = 57, 53%) of the articles mentioned the need to evaluate the test's diagnostic performance in further studies.

Fifty-seven (59%) of the 97 studies that did not use a clinically relevant population overinterpreted the clinical applicability. Of the 15 studies carried out with a clinically relevant population, 4 studies (3%) were also deemed to have overinterpreted their results because of insufficient diagnostic accuracy. In combination, overinterpretation of the clinical applicability of the test under study was apparent in more than half (n = 61, 56%) of the examined articles.

Authors solely based in clinical settings were much less likely to overinterpret results, and articles published in journals focusing on medical specialties were also less likely to do so. Furthermore, a higher impact factor for a journal was associated with a higher chance of overinterpretation (Table 4). Multivariable analyses indicated that laboratory-based authors were more likely than clinic-based authors to overinterpret the clinical implications of their results (odds ratio adjusted for study design, type of diagnostic accuracy index, and impact factor, 18.7; 95% CI, 1.41–249.26; $P = 0.026$). Articles from journals with impact factors in the upper quartile were more likely to overinterpret than those from the lowest quartile (odds ratio adjusted for study design, type of diagnostic accuracy index, and authorship, 4.33; 95% CI, 1.03–18.23; $P =$

**Table 3. Main characteristics of the 108 articles on molecular-diagnostic tests: overall results and results according to whether they overinterpreted clinical applicability.**

| Variables | No. of studies | Overinterpretation of studies? Yes, n (%) | Overinterpretation of studies? No, n (%) | *P*[a] |
|---|---|---|---|---|
| Study design | | | | 0.042 |
| Consecutive series or series of clinically relevant patients | 15 | 4 (27) | 11 (73) | |
| Healthy control or alternative-diagnosis control | 82 | 50 (61) | 32 (39) | |
| Other | 11 | 7 (64) | 4 (36) | |
| Accuracy index | | | | 0.014 |
| Sensitivity and specificity, or area under the ROC curve | 57 | 25 (44) | 32 (56) | |
| Predictive values or accuracy | 14 | 10 (71) | 4 (29) | |
| Diagnostic index not calculated | 36 | 26 (72) | 10 (28) | |
| Sample size by quartile, n | | | | 0.44[b] |
| Q1 (4–37) | 27 | 15 (56) | 12 (44) | |
| Q2 (38–68) | 26 | 17 (65) | 9 (34) | |
| Q3 (69–107) | 27 | 16 (59) | 11 (41) | |
| Q4 (108–8156) | 26 | 12 (46) | 14 (54) | |
| Journal category | | | | 0.025 |
| Medicine | 36 | 15 (42) | 21 (58) | |
| Oncology | 32 | 16 (50) | 16 (50) | |
| Biomedical or general science | 19 | 14 (74) | 5 (26) | |
| Laboratory and methodology | 21 | 16 (76) | 5 (24) | |
| Impact factor by quartile | | | | 0.050[b] |
| Q1 (<2.15) | 25 | 11 (44) | 14 (56) | |
| Q2 (2.16–3.87) | 25 | 12 (48) | 13 (52) | |
| Q3 (3.88–5.74) | 28 | 18 (64) | 10 (36) | |
| Q4 (5.75–51.30) | 21 | 14 (67) | 7 (33) | |
| Not classified[c] | 9 | 6 (67) | 3 (33) | |
| Authorship | | | | 0.005 |
| Clinic-based | 11 | 2 (18) | 9 (82) | |
| Both clinic- and laboratory-based | 79 | 34 (43) | 45 (57) | |
| Laboratory-based | 26 | 15 (58) | 11 (42) | |
| Technique used | | | | 0.30[d] |
| Gene-targeting techniques | | | | |
| PCR-based | 34 | 20 (59) | 14 (41) | |
| Microarray | 20 | 14 (70) | 6 (30) | |
| Protein-targeting techniques | | | | |
| Mass spectrometry or 2D gel electrophoresis | 44 | 20 (46) | 24 (55) | |
| Antibody array or protein microarray | 9 | 6 (67) | 3 (33) | |
| Other | | | | |
| Lipidomics | 1 | 0 (0) | 1 (100) | |

*Continued on page 791*

0.045). The association between overinterpretation and impact factor was linear (odds ratio, 1.71 per quartile; 95% CI, 1.09–2.69; $P = 0.020$). We calculated cross-tabulations to see the differences between journals with high vs low impact factors. The only difference observed was in journal category. The higher-

**Table 3. Main characteristics of the 108 articles on molecular-diagnostic tests: overall results and results according to whether they overinterpreted clinical applicability. (*Continued from page 790*)**

| Variables | No. of studies | Overinterpretation of studies? Yes, n (%) | No, n (%) | *P*[a] |
|---|---|---|---|---|
| Disease type | | | | 0.57[d] |
| Cancer | 61 | 31 (51) | 26 (49) | |
| Infectious disease | 19 | 14 (74) | 5 (26) | |
| Congenital disorders | 10 | 6 (60) | 4 (40) | |
| Autoimmune disease and transplant rejection | 8 | 4 (50) | 4 (50) | |
| Neurologic disease | 6 | 3 (50) | 3 (50) | |
| Other[e] | 4 | 3 (75) | 1 (25) | |
| Total | 108 | 61 (57) | 47 (44) | |

[a] *P* values from $\chi^2$ univariate test of homogeneity unless otherwise stated.
[b] $\chi^2$ test of tendency.
[c] Articles that did not enter the Thomson Reuters' ISI Web of Knowledge Journal Citation Report, edition 2006. Excluded from the statistical analysis were articles that were published in *BMC Medical Genetics*, *World Journal of Gastroenterology*, *Taiwan Journal of Obstetrics & Gynecology*, *Molecular Diagnosis & Therapy*, *Molecular Cancer*, *Translational Research*, *Journal of Zhejiang University. Science. B*, and *Journal of Thoracic Oncology*.
[d] Fisher exact test (2-tailed).
[e] Adenomyosis, endometriosis, osteonecrosis of the femoral head, and idiopathic pulmonary fibrosis.

impact journals included a higher proportion of those categorized as "laboratory and methodology," whereas the lower-impact journals included more "biomedical or general science" journals ($P = 0.010$).

### EXAMPLES IN THE ASSESSMENT OF OVERINTERPRETATION

*Example 1 (reference 25 in Annex 1 in the online Data Supplement)*. This study used an alternative diagnosis–

**Table 4. Multivariable analyses: variables significantly associated with overinterpretation of results.**

| Variables | n (%) | Adjusted odds ratio[a] | 95% CI | *P* |
|---|---|---|---|---|
| Study design | | | | |
| Consecutive series or series of clinically relevant patients | 15 (13.9) | 1.00 | | |
| Healthy control or alternative-diagnosis control | 82 (75.9) | 4.54 | 1.13–18.15 | 0.032 |
| Other | 11 (10.2) | 5.67 | 0.88–36.80 | 0.069 |
| Accuracy index | | | | |
| Sensitivity and specificity, or area under the ROC curve | 57 (52.8) | 1.00 | | |
| Predictive values or overall accuracy | 14 (12.9) | 1.85 | 0.42–8.13 | 0.417 |
| Diagnostic index not reported | 36 (33.3) | 2.87 | 1.03–7.96 | 0.043 |
| Authorship | | | | |
| Clinic-based | 11 (9.5) | 1.00 | | |
| Both clinic- and laboratory-based | 79 (68.1) | 4.50 | 0.44–46.14 | 0.206 |
| Laboratory-based | 26 (22.4) | 18.73 | 1.41–249.26 | 0.026 |
| Impact factor (by quartiles) | | | | |
| Linear relationship[b] | | 1.71 | 1.09–2.69 | 0.020 |

[a] Logistic regression model controlling for the effects of study design, type of accuracy index, authorship, and bibliographic-impact factor.
[b] Reference category is the previous quartile.

control design, and the statements regarding clinical applicability were considered definitely favorable: "This rapid MS-MA is a good primary screening method that can be implemented in a diagnostic laboratory to determine the methylation patterns of patients with suspected PWS or A." The authors confirm that the diagnostic test is a good primary-screening method, despite the limited conclusiveness of the study design; therefore, the study was considered as overinterpretation.

*Example 2 (reference 40 in Annex 1 in the online Data Supplement)*. This study used a healthy-control design, and we did not consider it to have overinterpreted its results. The statements regarding clinical applicability were judged as simply promising ("This study shows that free-circulating DNA can be detected in cancer patients compared with disease-free individuals, and suggests a new, non invasive approach for early detection of cancer."). The authors additionally specify the need of further studies to evaluate the test ("Further studies are needed to understand the correlation of these new molecular markers with cancer diagnosis, outcome of disease, and eventually treatment response.").

*Example 3 (reference 87 in Annex 1 in the online Data Supplement)*. This study used a clinically relevant population, and we considered the statements regarding clinical utility as definitely favorable ("Component-based testing and the whole-allergen CAP are equally relevant in the diagnosis of grass-, birch- and cat-allergic patients."). The authors specify the need for further clinical evaluation ("The clinical relevance of each allergen needs to be validated separately before the implementation of multiallergen panels into routine diagnostic settings."). This study had acceptable diagnostic accuracy (sensitivity, 72%; specificity, 92%) and therefore was not considered to have overinterpreted the clinical applicability of its results.

*Example 4 (reference 54 in Annex 1 in the online Data Supplement)*. This study also used a clinically relevant population; we considered the statements regarding clinical utility as definitely favorable ("This PCR assay detects a variety of strains exhibiting characteristics of the EAEC group, making it a useful tool for identifying both typical and atypical EAEC."); however, the authors did not report any measure of diagnostic accuracy. The study was therefore considered overinterpretation.

## DISCUSSION

Although clinical evaluation is necessary before introducing a test into clinical practice, few recent diagnostic studies on molecular research have been carried out in a clinically relevant population. The authors almost always interpreted their findings as either definitely favorable or at least promising for the evaluated technology. More than half of the articles apparently overinterpreted the clinical applicability of their findings, and such interpretation was more likely for articles in which all of the authors were laboratory-based and in articles published in journals with higher impact factors. Most of the reviewed studies used healthy- or alternative diagnosis–control designs. These studies are not all equal *(14)*: Some may be affected by biases, whereas others may be unbiased. Such nonequivalence is one more reason why evaluations with study designs that come closer to the real-life clinical settings are warranted.

Some authors have stressed the need to measure the value of a diagnostic test on health outcomes as a final phase in the evaluation of its clinical utility, once the test has been accepted clinically and made commercially available *(6, 9)*. We have not covered this issue in this study; however, we do agree that evaluating whether a test positively influences health outcomes is a key aspect. We chose not to cover this aspect because few molecular-diagnostic tests have been incorporated into practice and because trials evaluating the clinical utility of such tests are still scarce. For example, no randomized trials have conclusively assessed the clinical utility of tests involving gene expression profiling, despite several thousand published articles on the subject (2 trials are ongoing) *(15)*.

Other empirical investigations of the methodologic aspects of diagnostic research have reported serious methodologic limitations *(16–19)*. In the present study, however, we examined the applicability of diagnostic-test results to practice on the basis of the study design and independently of other methodologic aspects. We documented that considerable distance often exists between study design and the clinical applicability of the molecular-diagnostic tests, even if the design and the data are methodologically sound.

With the continuing development of new diagnostic tests, comprehensive clinical evaluations are needed if clinical harm and unnecessary spending are to be avoided. As our results show, studies that make claims about the clinical applicability of molecular-diagnostic tests often have not evaluated populations of clinically relevant patients and therefore lack evidence on which to base their claims. Enticing promises exist across the field of molecular medicine *(20)*. The exaggeration of the clinical implications of preliminary investigations that we observed in our study may be due to different processes *(4, 21, 22)*, including commercial influences *(4)* and insufficient awareness by researchers of their own "interpretive biases" *(23, 24)*.

Overinterpretation can certainly arise when a strong result is obtained from a very small study. Indeed, the lack of reproducibility in analyses of

proteomic and genomic data is often ascribed to small samples: The main difficulty in conducting a satisfactory early assessment is obtaining sufficient numbers of individuals for both training and validation; thus, the results may be overinterpreted. Large sample sizes and replication in multiple independent data sets are necessary but not sufficient for reliable results, however.

Comprehensive clinical evaluation of a single diagnostic test is expensive in terms of both money and time *(25)*. Reliable consecutive series of samples that are representative of the real clinical settings of interest may be difficult to obtain in molecular-based research. Unless a well-thought-out research study is designed in collaboration with a clinical center, few groups are likely to hand over their "precious" clinical samples and their clinical and demographic data to a laboratory *(26)*. Clinicians may be more sensitive to the difficulties and implications of moving these tests to the bedside and thus may be more cautious in their interpretation. Such reticence would be consistent with our observation that articles by exclusively laboratory-based authors were more likely to overinterpret the clinical applicability of their results. Finally, the observed relationship between journal-impact factor and overinterpretation could be a form of bias: Studies with the more spectacular conclusions appear in journals with higher impact factors, many of which are also more biologically and industry oriented than clinically based.

Some caveats about our methods require some discussion. First, we used an operational search strategy and definition to identify a sufficiently large number of molecular-diagnostic studies, but there is no established and widely agreed strategy for identifying such studies in the literature. To evaluate the consistency of the selection process, 2 investigators assessed a random sample of the abstracts and achieved an adequate degree of agreement with the initial reviewer. Therefore, only one reviewer carried out the complete search of the potential reports through MEDLINE. We cannot totally exclude the potential for selective inclusion, but our hope is that it is not large. Furthermore, the internal validity of the type of study we conducted does not require the same completeness of the sample that systematic reviews and metaanalyses of research findings require.

More importantly, passing judgment on whether overinterpretation exists is not always straightforward, and there is a risk that our own assessments overinterpret the language of an article. To establish an adequate definition of overinterpretation, we took into account several aspects in each scientific report; however, we acknowledge that this scheme is not a perfectly objective rule. The agreement between the independent data extractors was less than perfect. Although such deficiencies may affect the exact extent of estimated overinterpretation, it does not affect our main conclusion that inferences on clinical applicability are exaggerated in this literature.

The requirements for the introduction of diagnostic tests into clinical practice are less strict than for the introduction of new treatments. Hence, flawed or exaggerated claims for diagnostic-research results could lead to the premature adoption of defective tests, which could translate into erroneous decisions with adverse consequences for health. All in all, our results emphasize the necessity for caution when interpreting the results of diagnostic-accuracy studies in molecular research.

**Author Contributions:** *All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.*

**Authors' Disclosures of Potential Conflicts of Interest:** *Upon manuscript submission, all authors completed the Disclosures of Potential Conflict of Interest form. Potential conflicts of interest:*

**Employment or Leadership:** None declared.
**Consultant or Advisory Role:** None declared.
**Stock Ownership:** None declared.
**Honoraria:** None declared.
**Research Funding:** Spanish Agency for Health Technology Assessment (Exp PI06/90311) and CIBER en Epidemiología y Salud Pública (CIBERESP), Instituto de Salud Carlos III, Government of Spain.
**Expert Testimony:** None declared.

**Role of Sponsor:** The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, or preparation or approval of manuscript.

**Acknowledgments:** This manuscript was presented in poster format at the Fifth Annual Meeting of Health Technology Assessment International (HTAi), Montréal, Canada, July 6–9, 2008. We thank Jonathan Whitehead for editorial help in preparing an early version of the manuscript.

## References

1. Zolg W. The proteomic search for diagnostic biomarkers: lost in translation? Mol Cell Proteomics 2006;5:1720–6.
2. Ioannidis JP. Molecular bias. Eur J Epidemiol 2005;20:739–45.
3. Check E. Proteomics and cancer: running before we can walk? Nature 2004;429:496–7.
4. Porta M, Hernández-Aguado I, Lumbreras B, Crous-Bou M. 'Omics' research, monetization of intellectual property and fragmentation of knowledge: can clinical epidemiology strengthen integrative research? J Clin Epidemiol 2007;60:1220–5.
5. Hernández-Aguado I. The winding road towards evidence based diagnoses. J Epidemiol Community Health 2002;56:323–5.

6. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Clin Chem 2003;49: 7–18.
7. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol 2003;3:25.
8. Feinstein AR. Clinical epidemiology: the architecture of clinical research. Philadelphia: WB Saunders; 1985. 812 p.
9. Sackett DL, Haynes RB. Evidence base of clinical diagnosis: the architecture of diagnostic research. BMJ 2002;324:539–41.
10. Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, et al. Phases of biomarker development for early detection of cancer. J Natl Cancer Inst 2001;93:1054–61.
11. Dobbin KK, Beer DG, Meyerson M, Yeatman TJ, Gerald WL, Jacobson JW, et al. Interlaboratory comparability study of cancer gene expression analysis using oligonucleotide microarrays. Clin Cancer Res 2005;11:565–72.
12. Storms V, Baele M, Coopman R, Willems A, de Baere T, Haesebrouck F, et al. Study of the intra- and interlaboratory reproducibility of partial single base C-sequencing of the 16S rRNA gene and its applicability for the identification of members of the genus *Streptococcus*. Syst Appl Microbiol 2002;25:52–9.
13. Ransohoff DF. Rules of evidence for cancer molecular-marker discovery and validation. Nat Rev Cancer 2004;4:309–14.
14. Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM. Case-control and two-gate designs in diagnostic accuracy studies. Clin Chem 2005;51:1335–41.
15. Ioannidis JP. Is molecular profiling ready for use in clinical decision making? Oncologist 2007;12: 301–11.
16. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. JAMA 1995;274: 645–51.
17. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282:1061–6.
18. Lumbreras-Lacarra B, Ramos-Rincon JM, Hernández-Aguado I. Methodology in diagnostic laboratory test research in clinical chemistry and clinical chemistry and laboratory medicine. Clin Chem 2004;50:530–6.
19. Yesupriya A, Evangelou E, Kavvoura FK, Patsopoulos NA, Clyne M, Walsh MC, et al. Reporting of human genome epidemiology (HuGE) association studies: an empirical assessment. BMC Med Res Methodol 2008;8:31.
20. Kyzas PA, Denaxa-Kyza D, Ioannidis JP. Almost all articles on cancer prognostic markers report statistically significant results. Eur J Cancer 2007;43: 2559–79.
21. van't Veer LJ, Bernards R. Enabling personalized cancer medicine through analysis of gene-expression patterns. Nature 2008;452:564–70.
22. Van den Bruel A, Aertgeerts B, Buntinx F. Results of diagnostic accuracy studies are not always validated. J Clin Epidemiol 2006;59:559–66.
23. Kaptchuk TJ. Effect of interpretive bias on research evidence. BMJ 2003;326:1453–5.
24. Porta M, ed. A dictionary of epidemiology. 5th ed. New York: Oxford University Press; 2008. Interpretive bias; p. 133.
25. Veenstra TD. Global and targeted quantitative proteomics for biomarker discovery. J Chromatogr B Analyt Technol Biomed Life Sci 2007; 847:3–11.
26. Lumbreras B, Porta M, Hernández-Aguado I. Assessing the social meaning, value and implications of research in genomics. J Epidemiol Community Health 2007;61:755–6.

The supplementary data from this article can be found in annex 3 of the thesis. Annex 4 shows a Spanish newspaper clipping that refers to the article.

# PART 6

## GLOBAL DISCUSSION OF FINDINGS

### 6.1 Overview of main findings

QUADOMICS is an adaptation of QUADAS. It can be used for assessing the quality of individual diagnostic accuracy studies that use '-omics'-based research technology. When performing a systematic review or meta-analysis in this field, QUADOMICS can be used to evaluate the methodological quality of the primary studies and can help determine whether certain methodological characteristics are the sources of heterogeneity in the meta-analysis. The development of QUADOMICS involved the following adaptations to QUADAS: an additional step of assigning the primary studies to one of four phases of diagnostic research; the elimination of 2 of the original QUADAS items; the modification of the description of two of the original QUADAS items; the application of 2 items only to phase IV studies (studies carried out in a population as close as possible to that in which the test would be applied in practice); and finally, the incorporation of 4 new items dealing with the type of sample, pre-analytical handling conditions, clinical and physiological characteristics, and overfitting. The newly developed tool was applied in triplicate to a sample of 45 diagnostic '-omics' studies to evaluate its consistency and applicability. The studies were systematically identified and were heterogeneous with regard to the disease of interest and the type of '-omics' technology used. The percentage agreement with the consensus scoring was high for all 3 reviewers, (82.3, 83.0 and 89.9). Some items were more difficult to apply because the studies had not included a clearly defined reference

standard. An explanation of how this situation can be dealt with, along with clear examples, was provided.

The methodological quality of 45 systematically identified diagnostic studies that used '-omics' based technology was evaluated with the newly developed QUADOMICS tool. Of the 45 studies evaluated, the majority were considered to be in phase I, that is that they used a case-control type design. The studies were subject to numerous methodological limitations which may have influenced their estimates of sensitivity and specificity. Notably, none of the studies included a flow diagram describing the patient recruitment process; and in less than half of the studies it was clear that the index test had been interpreted without knowledge of the reference standard. In a separate analysis, we explored how the authors of molecular diagnostic studies interpreted the clinical applicability of their findings. In a sample of 106 molecular diagnostic studies published in 2006, the authors almost always interpreted their findings as either definitely favourable or at least promising for the evaluated technology, even though most of them were considered to be in phase I. More than half of the articles apparently overinterpreted the clinical applicability of their findings, and this was more likely in articles where all authors were laboratory-based and in articles published in journals with higher impact factors.

Overall, the studies examined here were of poor quality and overinterpretation of preliminary research findings was common. The concept that diagnostic research is of poor quality is not new. Numerous investigators have described how diagnostic research pales in comparison to therapeutic research when it comes to both standards and methodological quality (Hernandez-Aguado I, 2002). Molecular diagnostic research has been shown to be especially susceptible to methodological deficits. One study based on 44 studies of genetic, molecular and proteomic tests showed that studies met an average of 9.8 (95% CI 8.8-10.6) of the 24 STARD criteria (Lumbreras B et al, 2006). Although STARD is a reporting guideline, reporting criteria are indirectly linked to methodological quality and so these results are comparable to those found here. Furthermore, we have described that some of the QUADOMICS items were difficult to apply due to the poor reporting of the primary research studies. We showed that none of the 45 '-omics' studies included a flow diagram describing the selection and recruitment of study participants. Such diagrams are also strongly recommended in the STARD

publication (Bossuyt PM et al, 2003). Similarly, a recent review of commercial tests for HIV, tuberculosis or malaria showed that only 13% of studies reviewed met the STARD criterion which recommends the flow diagram (Fontela PS et al, 2009).

We have shown that some modern '-omics' studies do not adequately address overfitting and hence may present results that are not reproducible. Overfitting can be avoided if results are validated in an independent patient population, and hence, we scored studies positively for QUADOMICS item 16 (Is it likely that the presence of overfitting was avoided?) only when independent validation had been performed. In our evaluation, just under half met this criterion. Although the proportion is fairly high, it does seem that there has been some improvement related to the problems of reproducibility and the importance of external validation. A research report from 2003 showed that only 26% of 84 studies using microarray for cancer diagnosis had attempted to carry out any kind of validation, either independent validation or cross validation [Ntzani EE et al, 2003]. Some high profile examples of the difficulties with reproducing '-omics' results can be seen in proteomics experiments (Baggerly KA, 2005; Ransohoff DF, 2005). Evaluation of the proteome is much more complex than genomics due to post translational modification, and potential for change depending on the experimental conditions. Researchers realized that proteomic signatures developed for very different diseases, under very diverse conditions were all actually identifying the same proteins, suggesting that the changes represent common cellular stress responses rather than meaningful tools to aid diagnosis (Petrak J et al, 2008).

The exaggeration of the clinical implications of preliminary investigations is difficult to rationalize, especially if we take into account different writing cultures. It could be linked to commercial influences, insufficient awareness by researchers of the limitations of their study designs or their own "interpretive biases" (Porta M et al, 2007; Ransohoff DF, 2010; Kaptchuk TJ et al, 2003). Furthermore, the tendency to err on the side of optimism is not surprising given the need of most researchers to provide results to secure future grant funding. Although we are unaware of other studies that attempt to describe analytically overinterpretation in this way, the gap between what studies claim and what actually impacts on clinical practice has frequently been described (Vitzthum F et al, 2005; Frangioni JV, 2006; Ioannidis JP, 2010). There are only a few examples of '-omics' tests that currently influence clinical decision making. The most progress

has been made in breast cancer prognosis, with both the MammaPrint assay (Agendia BV, The Netherlands) and the Oncotype DX (Genomic Health) now commercially available for developing individualized treatment plans. The MammaPrint assay is a 70-gene expression profile and was cleared by the U.S. Food and Drug Administration in 2007 (FDA, US Food and Drug Administration, 2007; Slodkowska EA et al, 2009). Oncotype DX is a 21-gene RT-PCR assay and has been validated widely in distinct patient populations (Habel LA et al, 2006; Toi M et al, 2010; Kelly CM et al, 2010).

### 6.2 Clinical implication of the development of a tool for assessing the quality of diagnostic accuracy studies that use 'omics' technologies

Molecular diagnostics is a highly dynamic field in which a great deal of research is currently being carried out (Ghosh D et al, 2009). New tests based in '-omics' technologies are continually proposed to improve diagnosis, prognosis and to predict the responsiveness to therapy in individualized medicine. Filtering the huge amount of information produced and translating research results into clinical and public health practice is a major challenge. Systematic reviews play a key role in this endeavour, and a suitable tool for assessing the methodological quality of the primary studies that are included in systematic reviews is necessary. QUADOMICS is a tool adapted to assess the quality of diagnostic studies using '-omics' based technologies which can consistently be applied to a broad range of disease conditions and technologies. QUADOMICS can help researchers that are carrying out systematic reviews, identify methodological weaknesses of the studies included in their review. Illustrating when the evidence base for a new test is limited to poor quality or potentially biased studies is important because it encourages decision makers to exercise caution.

The widespread application of QUADOMICS can therefore enhance evidence based diagnosis by aiding the assessment of research findings. The consequences and implications are the following. A systematic review can be performed concluding that there is sufficient evidence in favour of a new test, and using QUADOMICS can show that the evidence comes from good quality unbiased studies. In this scenario, reviewers, health technology assessment agencies and other decision makers should recommend the timely introduction of new tests into clinical practice. This in turn can potentially lead to the earlier diagnosis of disease, timely treatment and ultimately improve patient

health and save lives. On the other hand, a systematic review may conclude that there is sufficient evidence against a new test, for example, by showing that all studies with positive finding are of poor quality or seriously biased and methodologically sound studies do not show positive results. In this scenario, decision makers will not recommend introduction of the new test. In addition to avoiding potential adverse events in patients due to ineffective diagnostic procedures and misinformed clinical decisions, arriving at a clear and confident recommendation will help minimize the resources dedicated unnecessarily to the clinical validation of ineffective tests. Given the numerous commercial interests in the development of new molecular diagnostic tests, this latter point is important.

### 6.3 Clinical implication of the findings regarding current research on the diagnostic application of new molecular technologies

We have shown here that the research reports published in the '-omics' field tend to be of poor methodological quality and are potentially susceptible to biases that could influence the estimations of diagnostic accuracy reported. Poor quality research will hinder the evidence based transition of the new tests into clinical practice as described above. Furthermore, it may have the following consequences. Firstly, attempting to replicate spurious research findings from biased investigation would represent an inefficient use of time, money and other resources. More importantly, decisions based on spurious results from biased studies could lead to the adoption of ineffective tests in clinical and public health practice, which in turn may cause incorrect clinical decisions and ultimately cause harm to patients.

While some of the methodological deficiencies described were linked to the specific peculiarities of '-omics' based research, other important aspects –which have long been considered fundamental in traditional diagnostic research, such as the description of the index test and test reproducibility– are being overlooked in '-omics' research. I will now discuss some of methodological deficits that we observed in the sample of studies evaluated and how they may impact on clinical or public health practice.

*Study design:*

In both of our samples, a high proportion of the studies used a case control design (77.8% of the sample of '-omics' studies, and 75.9% of the sample of molecular diagnostic studies). This observation in itself has important clinical implications given that the case control design has been shown to lead to an over-estimation of diagnostic accuracy (Whiting P et al, 2004; Lijmer JG et al, 1999; Rutjes AW et al, 2005). Inflated estimates of either sensitivity or specificity may lead to incorrect clinical decisions involving treatment and intervention which in some cases may be risky or have secondary effects that are harmful to the patient. For example, a physician using a test presumed to be highly sensitive may be overly confident about ruling out a disease when a test result is negative and in this way the patient might not receive available treatment. Valid and reliable estimates of the sensitivity and specificity of a new test are required to ensure that correct clinical decisions are made. Furthermore these estimations must come from clinically relevant patient populations. The fact that so few of these molecular diagnostic studies were carried out in a clinically relevant population shows a lack of understanding of study design required for adequate clinical validation (Ransohoff DF, 2009).

*Description of patient population and external validity:*

It is necessary that estimates of the diagnostic validity of a new test come from well described patient populations. Here, it was shown that none of the '-omics' studies clearly described the selection criteria leading to patient selection, and more than half failed to report the clinical and physiological characteristics of the study population. Alarmingly some reports failed to present basic information such as the age and sex of the patients. Studies which fail to report the characteristics of the patient who have taken part in the study limit the reader's ability to judge the external validity of the study and may lead to incorrect clinical decisions. For example, health care workers are only able to judge whether the results of a study are applicable to their patient when the patient characteristics are reported. If a diagnostic test demonstrates high diagnostic accuracy during validation experiments that were carried out exclusively in young male patients, it is impossible to judge if the diagnostic accuracy will be the same in a 70 year only female patient. Consequently, the healthcare worker may use the test in the elderly woman presuming the diagnostic accuracy to be the same, unaware of any limitation in external validity, simply because the demographic details of the study population was not mentioned. It should be noted that some diagnostic procedures have shown to be

less accurate in the elderly population (Kurosaki M et al, 2008) and that socio-demographic characteristics such as sex do influence diagnostic accuracy (Whiting P et al, 2004; Roger VL et al, 1997). Furthermore, the potential clinical utility of a new test may not be equal in all patients due to differences in test performance in practice. For example, it is possible that the procedures involved in many '-omics' technologies such as DNA purification and amplification may be less successful in elderly patients due to degeneration of genetic material.

*Potential for bias due to variation in test procedures:*

In the previous paragraph, I have focused more on the need to report patient characteristics for evaluating the external validity of the findings. Variation in the clinical and other factors related to the patients who have provided the samples can also introduce bias into the studies, and failure to report patient characteristics makes it impossible to evaluate such biases. Bias may also be introduced by variation in the experimental conditions and so it is important that new '-omics' tests are validated in studies where all pre-analytical and analytical conditions are uniform. While there are standardized protocols for blood extraction in most institutions, these are generally not suitable for large scale validation experiments due to slight variations in the timing from blood draw to aliquoting and storage, and ensuing differences in the proteolysis of serum proteins (Latterich M et al, 2008). We have shown here that this limitation of '-omics' base research appears to be widely acknowledged by the research community. In our sample only 7, 16% of the studies failed to report pre-analytical in sufficient details, or reported that they were not equal for all subjects without analyzing their influence. These studies may have been subject to variation in the pre-analytical test conditions, thus producing biased estimates. Furthermore, potential variation was detected on a much larger scale regarding the clinical and physiological characteristics of the patients who provided the biological samples, as well as the diagnostic and treatment procedures they had undergone. Consideration of patient characteristics and using uniform experimental procedures are important in all diagnostic research, but is particularly relevant in '-omics' research given that certain biomarker profiles are especially susceptible to variation caused by these aspects. As previously mentioned, biased studies hinder the evidence based transition of new tests into clinical or public health practice.

*Overfitting and other challenges to reproducibility:*

Before assessing the utility of new genomic or proteomic signatures for clinical diagnosis it is essential to understand the complexity of the data analyses used to derive them. '-Omics' data typically involves the analysis of thousands of individual parameters and so one would expect to find many spuriously significant associations purely by chance (5%). Despite the body of literature addressing and highlighting the serious issue of overfitting in this type of research (Simon R et al, 2003), less than half of the studies in our sample adequately controlled for overfitting by validating their findings in a completely independent set of patient samples. It is important to recognise that attempting to replicate biomarker signatures that were obtained due to chance is an inefficient use of time, money and other resources. Most major journals now require that high throughput data be made publicly available upon publication of the article. Therefore it should become increasingly feasible to use publically available '-omics' data, generated by other investigators studying a related problem as a method for validating biomarker profiles. Results that are reproducible across multiple studies show strong evidence of a true association and thus are more likely to be clinically useful.

*Overinterpretation of preliminary research findings:*

It is possible that authors tend to be overly positive about their results in order to increase the chances of publication. The association between positive findings and publication has been demonstrated (Dickersin K et al, 1992). The overinterpretation described here is important, especially as it is apparent in high profile journals. One study has shown that claims from highly cited observational studies continue to be supported in the medical literature even when there is strong contradictory evidence from randomized trials (Tatsioni AT et al, 2007). The tendency to exaggerate the clinical relevance of preliminary research findings makes the evidence based provision of new diagnostic tests particularly challenging. Given the lack of knowledge regarding potential limitations or bias, one should be cautious. Flawed or exaggerated claims on diagnostic research could lead to the premature adoption of defective tests, which could translate into erroneous decisions with adverse consequences for health.

### 6.4 Limitations

One limitation of this work is that the '-omics' field is highly dynamic. While we attempted to identify and include *all* the pertinent threats to validity involved in diagnostic accuracy studies which use these new technologies –technologies are continually evolving and it is possible that new threats will be uncovered which are not included in the QUADOMICS tool. With regard to validation, the tool was applied in triplicate to 45 studies. Two of the three reviewers were involved in the development of the tool, and therefore may have found it easier to apply. For this reason it was necessary to include one researcher who had not been involved with the development of the tool. Furthermore, the tool was tested in a slightly artificial situation: In practice the tool would be used to evaluate the quality of studies included in a systematic review and all studies would be addressing the same question, diagnosing the same disease with the same '-omics' technology. In the validation procedure described here, the 45 studies addressed different diseases and used different technologies, which made application more challenging. It is likely that the consistency between reviewers would actually be higher if all we had been evaluating 45 studies on the same subject. Nevertheless, with this sample we have illustrated that the tool is applicable to a broad range of studies.

The diagnostic '-omics' studies included in our sample were subject to numerous methodological biases. Valid estimations of study quality rely on comprehensive and transparent reporting of the methodology. A number of the items in QUADAS and QUADOMICS are in essence reporting items – e.g. whether the selection criteria are clearly described, or whether the index and reference tests are described in sufficient detail to permit their replication. It follows that some studies are deemed to be of poor quality because they are poorly reported. It is not always clear if a study is subject to the relevant bias simply because the relevant points are not addressed in the report, e.g. if the study does not mention whether the test results were interpreted with or without knowledge of the reference standard. In these cases, an indirect link between study reporting and study quality is assumed– i.e. that studies poorly reported are probably poorly done – but this may not always be the case. However, it is likely that researchers who are aware of, and control for, potential biases are also aware of the need to report

such details. Fortunately, ventures like STARD [Bossuyt PM et al, 2003] are now available to guide researchers in reporting all of important details.

Additionally, we attempted to evaluate how the authors of molecular diagnostic studies interpreted the clinical applicability of their research findings. Passing judgment on whether or not there is overinterpretation is not straightforward, especially given different writing cultures, and the need to be optimistic with regard to the potential impact of the findings in order to convince editors to publish them. We developed and piloted a strategy to determine overinterpretation in order to make our evaluation as objective as possible. The agreement between the independent data extractors was less than perfect and while this may affect the exact extent of estimated overinterpretation, it does not affect the main conclusion that inferences on clinical applicability are exaggerated in this literature.

Finally, it is necessary to consider the external validity of our observations. We included 45 '-omics' diagnostic studies published in 2006 through 2009 for evaluating the quality of '-omics' based research, and 106 molecular diagnostic studies from 2006 studies for evaluating how the authors interpreted the clinical applicability of their findings. Neither sample was restricted to any particular field or technique, but clearly they were limited to published reports and to those indexed by Pubmed. It is arguable that studies indexed by Pubmed would actually be of better methodological vigour and subject to more rigorous peer review, and so our estimations of quality and overinterpretation would actually be underestimations of what is occurring in practice. It is not clear how only including studies published in the English language would influence quality assessment. It is possible that studies published in international journals would be of better quality but even if the language restriction does not influence our estimation of overall study quality it could be related to overinterpretation. That authors tend to send positive research findings to international journals and negative or less novel research findings to local journals has been demonstrated;  it follows that preliminary research findings which the authors deem to be definitive evidence of clinical utility of a new test would also be sent to an international journal rather than a local one.

# PART 7

## CONCLUSIONS

The introduction of a new diagnostic test into clinical practice does not follow the same rigorous structure as the introduction of a new treatment or pharmaceutical. Nevertheless, inappropriate or premature application of diagnostic procedures may lead to incorrect clinical decisions, unnecessary patient discomfort, and adverse patient outcomes. We have shown that '-omics' based diagnostic research is of poor quality and that authors of molecular based diagnostic studies show a tendency to overinterpret their results. The development of the QUADOMICS tool is therefore an important way to combat this problem. In addition to providing reviewers of '-omics' diagnostic studies with an adequate tool, it is hoped that QUADOMICS will help sensitize researchers, clinicians and other decision makers to the serious threats to the validity inherent in this type of research, therefore assuming a key role to ensure that the provision of '-omics' tests to the clinic is based in the best available evidence. A brief conclusion with regard to each of the four specific objectives can be found on the following page.

UNIVERSITAS
Miguel
Hernández

## 7.1 Conclusions with regard to the specific objectives

1. The QUADAS guide was adapted to incorporate the specific sources of error relevant to '-omics' technologies. The new tool was named QUADOMICS.

2. QUADOMICS proved to be applicable and consistent. Independent users made analogous observations and judgements when appraising the same study.

3. The methodological quality of a sample of diagnostic accuracy studies that use '-omics' technologies was poor. Studies were subject to bias caused by the complexities of the new technologies but also lacked methodological vigour long since established for diagnostic research.

4. Overinterpreting the clinical applicability of molecular diagnostic studies is common. Authors frequently interpret studies carried out in preliminary patient populations as providing definitive evidence of clinical applicability. Studies with solely laboratory based authors and those published in high impact scientific journals are especially prone to overinterpretation.

UNIVERSITAS
Miguel
Hernández

# PART 8

## REFERENCES


- Attia J, Ioannidis JPA, Thakkinstian A, McEvoy M, Scott RJ, Minelli C, et al. How to use an article about genetic association A: Background concepts. JAMA 2009;301:74-81.
- Attia J, Ioannidis JPA, Thakkinstian A, McEvoy M, Scott RJ, Minelli C, et al. How to use an article about genetic association B: Are the results of the study valid? JAMA 2009;301:191-7.
- Attia J, Ioannidis JPA, Thakkinstian A, McEvoy M, Scott RJ, Minelli C, et al. How to use an article about genetic association C: What are the results and will they help me in caring for my patient? JAMA 2009;301:304-8.
- Bachmann L, Puhan MA, ter Reit G, Bossuyt PM. Sample sizes of studies in diagnostic accuracy: literature survey. BMJ 2006; 332:1127-9.
- Baggerly KA, Morris JS, Edmonson SR, Coombes KR. Signal in noise: evaluating reported reproducibility of serum proteomic tests for ovarian cancer. J Natl Cancer Inst 2005;97:307-9.
- Bossuyt PM, Reitsma JB, Bruns DE Gatsonis CA, Glasziou PP, Irwig LM, et al. Toward complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Ann Intern Med 2003:138:40-4.
- Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med. 2003;138:W1-12.
- Botling J, Edlund K, Segersten U, Tahmasebpoor S, Engström M, Sunderström M et al. Impact of thawing on RNA integrity and gene expression analysis in fresh frozen tissue. Diagn Mol Pathol 2009;18:44-52.

- Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 2001;29:365-71.

- Check E. Proteomics and cancer: running before we can walk? Nature 2004;429:496-7.

- Correlogic Systems, Inc. Ovarian Cancer. Germantown, MD:Correlogic Systems, Inc: 2010 [Website last accessed 12 January 2011] Available from: http://www.correlogic.com/research-areas/ovarian-cancer.php

- Deutsch EW, Ball CA, Berman JJ, Bova GS, Brazma A, Bumgarner RE, et al. Minimum information specification for in situ hybridization and immunohistochemistry experiments (MISFISHIE). Nat Biotechnol 2008; 26:305-12.

- Diamandis EP. Oncopeptidomics: A useful approach for cancer diagnostics? Clin Chem 2007; 53:1004-6.

- FDA, US Food and Drug Administration. FDA Approves First Molecular-Based Lab Test to Detect Metastatic Breast Cancer. FDA News release July 16 2007. Washington, D.C: FDA; 2007

- Feinstein AR. Clinical epidemiology: the architecture of clinical research. Philadelphia: WB Saunders; 1985.

- Florkowski CM. Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests. Clin Biochem Rev 2008;29:S83-7.

- Fontela PS, Pant Pai N, Schiller I, Dendukuri N, Ramsay A, Pai M. Quality and reporting of diagnostic accuracy studies in TB, HIV and malaria: evaluation using QUADAS and STARD standards. PLoS One 2009;4:e7753.

- Frangioni JV. Translating in vivo diagnostics into clinical reality. Nat Biotechnol 2006;24:909-13.

- Garde AH, Hansen AM, Skovgaard LT, Christensen JM. Seasonal and biological variation of blood concentrations of total cholesterol, dehydroepiandrosterone sulphate, haemoglobin A1c, IgA, Prolactin, and free testosterone in healthy women. Clin Chem 2000;46;551-9.

- Ghosh D, Poisson LM. "Omics" data and levels of evidence for biomarker discovery. Genomics 2009;93:13-6.

- Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. A. Are the results of the study valid? Evidence-Based Medicine Working Group. JAMA. 1993;270:2598-601.

- Habel LA, Shak S, Jacobs MK, Capra A, Alexander C, Pho M, et al. A population-based study of tumor gene expression and risk of breast cancer death among lymph node-negative patients. Breast Cancer Res. 2006;8(3):R25.

- Hanash SM, Bobek MP, Rickman DS, Williams T, Rouillard JM, Kuick R et al. Integrating cancer genomics and proteomics in the post-genome era. Proteomics 2002;2:69-75.

- Hernandez-Aguado I. The winding road towards evidence based diagnoses. J Epidemiol Community Health 2002;56:323-5.

- Ioannidis JP. Expectations, validity, and reality in omics. J Clin Epidemiol 2010;63:945-9.

- Kaptchuk TJ. Effect of interpretive bias on research evidence. BMJ 2003; 326: 1453-55.

- Kelly CM, Krishnamurthy S, Bianchini G, Litton JK, Gonzalez-Angulo AM, Hortobagyi GN, Pusztai L. Utility of oncotype DX risk estimates in clinically intermediate risk hormone receptor-positive, HER2-normal, grade II, lymph node-negative breast cancers. Cancer. 2010;116:5161-7.

- Knottnerus JA, Buntinx F. The evidence base of clinical diagnosis: Theory and methods of diagnostic research, 2nd edition. Oxford: Blackwell Publishing Ltd. 2009

- Kurosaki M, Izumi N. External Validation of FIB-4: Diagnostic Accuracy Is Limited in Elderly Populations. Hepatology 2008 ;47:352.

- Lantz CA, Nebenzahl E. Behavior and interpretation of the κ statistic: Resolution of the two paradoxes. J Clin Epidemiol 1996;49:431-4.

- Latterich M, Abramovitz M, Leyland-Jones B. Proteomics: New Technologies and clinical applications. Eur J Cancer 2008;44:2737-41.

- Leeflang MMG, Bossuyt PMM, Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. J Clin Epidemiol 2009;62:5-12.

- Lijmer JG, Leeflang M, Bossuyt PM. Proposals for a phased evaluation of medical tests. Med Decis Making. 2009;29:E13-21.

- Lijmer JG, Mol BW, Heistekamp S, Bonsel GJ, Prins MH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282:1061-6.

- Lin DS, Coleman IM, Hawley S, Huang CY, Dumpit R, Gifford D et al. Influence of surgical manipulation on prostate gene expression: Implications for molecular correlates of treatment effects and disease prognosis. J Clin Oncol 2006;24:3763-70.

- Little J, Higgins JPT, Ioannidis JPA, Moher D, Gagnon F, von Elm E, et al. Strengthening the Reporting of Genetic Association studies (STREGA) – an extension of the STROBE statement. Eur J Clin Invest 2009;39:247-66.

- Lumbreras B, Jarrín I, Hernández Aguado I. Evaluation of the research methodology in genetic, molecular and proteomic tests. Gac Sanit 2006;20:368-73.

- Lumbreras B, Porta M, Márquez S, Pollán M, Parker LA, Hernandez-Aguado I. QUADOMICS: An adaptation of the Quality Assessment of Diagnostic Accuracy Assessment (QUADAS) for the evaluation of the methodological quality of studies on the diagnostic accuracy of '-omics' based technology. Clin Biochem 2008;41:1316-25.

- Lumbreras-Lacarra B, Ramos-Rincon JM, Hernandez-Aguado I. Methodology in diagnostic laboratory test research in clinical chemistry and clinical chemistry and laboratory medicine. Clin Chem 2004;50:530-6.

- Lumbreras B, Porta M, Marquez S, Pollán M, Parker LA, Hernández-Aguado I. Sources of error and its control in studies on the diagnostic accuracy of "-omics" technologies. Proteomics Clin Appl 2009;3:173–84.

- Marshall E. Getting the noise out of gene arrays, Science 2004;306;630-1.

- McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM, et al. REporting recommendations for tumour MARKer prognostic studies (REMARK). Eur J Cancer 2005;41;1690–6.

- Mulherin SA, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. Ann Intern Med 2002;137:598-602.

- Nature briefing. Proteomics, transcriptomics: what's in a name? Nature. 1999;402:715.

- Ntzani EE, Ioannidis JP. Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. Lancet. 2003;362:1439-44.

- Petrak J, Ivanek R, Toman O, Cmejla R, Cmejlova J, Vyoral D, et al. Déjà vu in proteomics. A hit parade of repeatedly identified differentially expressed proteins. Proteomics 2008;8:1744-9.

- Petricoin EF, Ardekani AM, Hitt BA, et al. Use of proteomic patterns in serum to identify ovarian cancer. Lancet 2002;359:572–7.

- Porta M, Hernández-Aguado I, Lumbreras B, Crous-Bou M. 'Omics' research, monetization of intellectual property and fragmentation of knowledge: can clinical epidemiology strengthen integrative research? J Clin Epidemiol 2007;60:1220-5.

- Porta M, Hernández-Aguado I, Lumbreras B, Crous-Bou M. 'Omics' research, monetization of intellectual property and fragmentation of knowledge: can clinical epidemiology strengthen integrative research? J Clin Epidemiol 2007;60:1220-5.

- Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 1978;299:926-30.

- Ransohoff DF, Gourlay ML. Sources of bias in specimens for research about molecular markers for cancer. J Clin Oncol 2010;28:698-704.

- Ransohoff DF. How to improve reliability and efficiency of research about molecular markers: roles of phases, guidelines, and study design. J Clin Epidemiol 2007;60:1205-19.

- Ransohoff DF. Lessons from controversy: Ovarian cancer screening and serum proteomics. J Natl Cancer Inst 2005;97:315-9.

- Ransohoff DF. Promises and limitations of biomarkers. In: Senn HJ, Kapp U, Otto F (Editors) Cancer Prevention II Recent Results in Cancer Research, Vol 181, II, pages 55-59, Springer-Berlag, Berlin 2009.

- Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. JAMA 1995;274:645-51.

- Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM. Case-control and two-gate designs in diagnostic accuracy studies. Clin Chem 2005;51:1335-41.

- Rutjes AWS, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PMM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ 2006;176:469-76.

- Schulz KF, Altman DG, Moher D; CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. PLoS Med. 2010;7:e1000251.

- Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst. 2003;95:14-8.

- Slodkowska EA, Ross JS. MammaPrint 70-gene signature: another milestone in personalized medical care for breast cancer patients. Expert Rev Mol Diagn 2009;9:417-22.

- Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Reitsma JB, Bossuyt PM, et al. Quality of reporting of diagnostic accuracy studies. Radiology 2005;235:347-53.

- Tatsioni AT, Bonitits NG, Ioannidis JPA. Persistence of contradicted claims in the literature. JAMA 2007;298:2517-26.

- Taylor CF, Paton NW, Lilley KS, Binz PA, Julian RK, Jones AR et al. The minimum information about a proteomics experiment (MIAPE). Nat Biotechnol 2007; 25:887-93.

- Timms JF, Arslan-Low E, Gentry-Maharaj A, Luo Z, T'Jampens D, Podust VN et al. Preanalytic influence of sample handling on SELDI-TOF Serum protein profiles. Clin Chem 2007;53:645-55.

- Toi M, Iwata H, Yamanaka T, Masuda N, Ohno S, Nakamura S, et al. Clinical significance of the 21-gene signature (Oncotype DX) in hormone receptor-positive early stage primary breast cancer in the Japanese population. Cancer. 2010;116:3112-8.

- Vitzthum F, Behrens F, Anderson NL, Shaw JH. Proteomics: from basic research to diagnostic application. A review of requirements & needs. J Proteome Res 2005;4:1086-97.

- von Elm E, Altman DG, Egger M, Pocock SJ, Gotzche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS Med 2007;4:1623-7.

- Wagner L. A test before its time? FDA stalls distribution process of proteomic test. J Natl Cancer Inst 2004;96:500-1.

- Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med 2004;140:189-202.

- Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PMM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol 2003;3:25-37.

- Whiting PF, Weswood ME, Rutjes AW, Reitsma JB, Bossuyt PN, Kleijnen J. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. BMC Med Res Methodol 2006;6:9-16.

- Wild C, Vineis P, Garte S, editors. Molecular Epidemiology of Chronic Diseases. Hoboken, NJ: John Wiley & Sons Inc: 2008

UNIVERSITAS
Miguel
Hernández

# PART 9

## ANNEXES

UNIVERSITAS
Miguel
Hernández

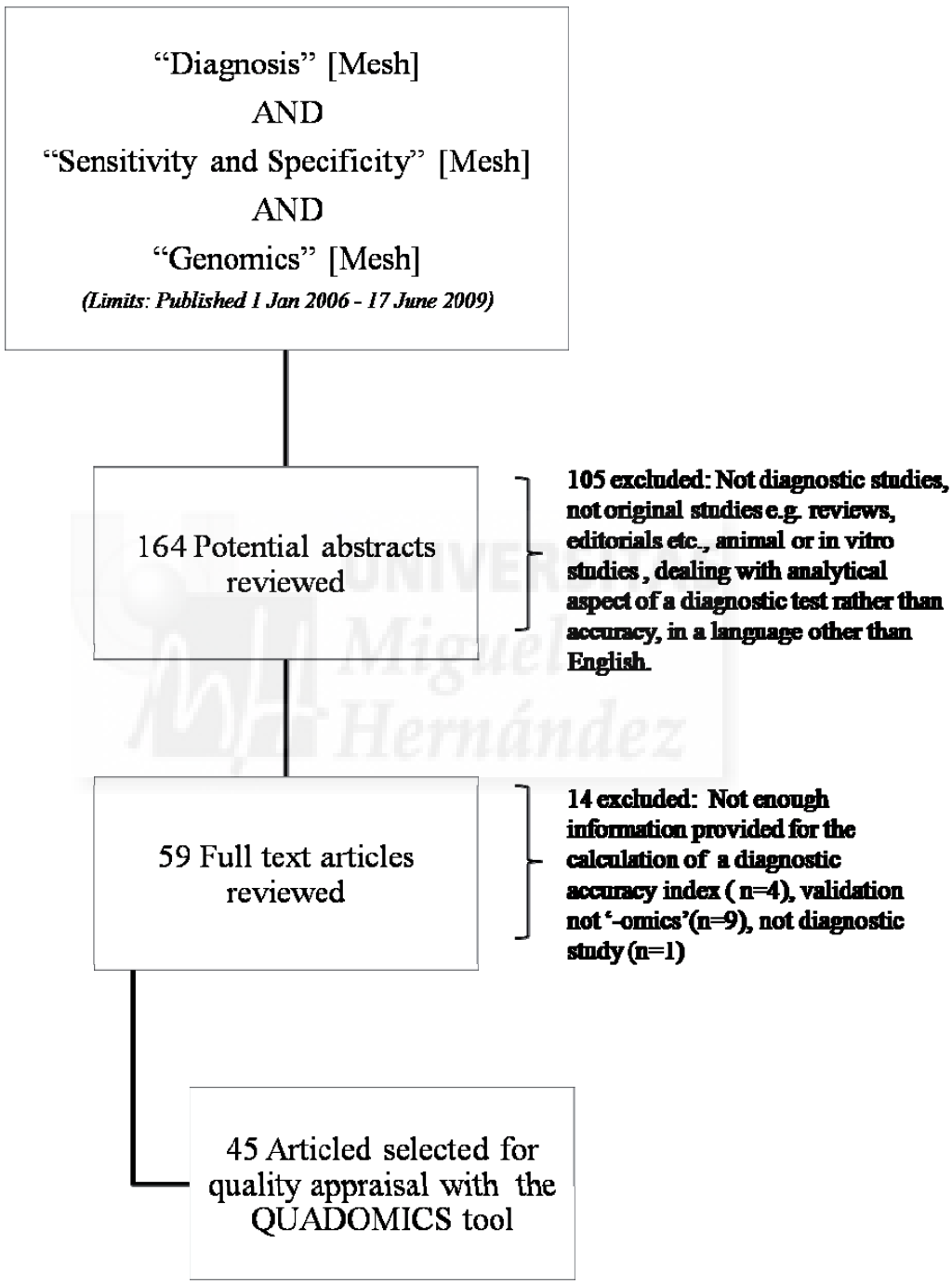
Annex 1: Spanish translation of QUADOMICS, presented at the XXVII meeting of the Spanish Society of Epidemiology in Zaragoza, Spain, 2009.

| QUADOMICS | Si | No | No es claro | No aplicado |
|---|---|---|---|---|
| Fase de estudio I_____ II_____ III_____ IV_____ | | | | |
| 1. ¿Se describieron claramente los criterios de selección? | | | | |
| 2. ¿El espectro de pacientes era representativo de los pacientes que recibirán la prueba en la práctica? | | | | |
| 3. ¿Se describió el tipo de muestra de manera completa? | | | | |
| 4. ¿Se describieron los procedimientos y los tiempos para la recogida de las muestras biológicas con respecto a los factores clínicos con suficiente detalle? | | | | |
| 4.1. ¿Factores clínicos y fisiológicos? | | | | |
| 4.2. ¿Procedimientos diagnósticos o tratamientos? | | | | |
| 5. ¿Se describieron los tratamientos y procedimientos pre-analíticos con suficiente detalle y fueron similares para todas las muestras? Y, si se mencionaron diferencias, ¿se evaluó su efecto en los resultados? | | | | |
| 6. ¿El periodo de tiempo entre la aplicación del estándar de referencia y la prueba de estudio fue suficientemente corto para garantizar que la condición no hubiera cambiado? | | | | |
| 7. ¿Es probable que la prueba de referencia clasifique la condición correctamente? | | | | |
| 8. ¿Toda la muestra o una selección aleatoria de la muestra recibió verificación con el estándar de referencia? | | | | |
| 9. ¿Los pacientes recibieron el mismo estándar de referencia a pesar del resultado de la prueba de estudio? | | | | |
| 10. ¿Se describió la ejecución de la prueba de estudio con suficiente detalle para permitir su replicación? | | | | |
| 11. ¿Se describió la ejecución del estándar de referencia con suficiente detalle para permitir su replicación? | | | | |
| 12. ¿Se interpretaron los resultados de la prueba de estudio sin conocimiento de los resultados obtenidos con el estándar de referencia? | | | | |
| 13. ¿Se interpretaron los resultados del estándar de referencia sin conocimiento de los resultados obtenidos con la prueba de estudio? | | | | |
| 14. ¿La información clínica de la que se disponía cuando se interpretaron los resultados de la prueba, estará presente cuando se aplique la prueba en la práctica? | | | | |
| 15. ¿Se informó sobre los resultados no interpretables o intermedios? | | | | |
| 16. ¿Es probable que se evitara la presencia de overfitting? | | | | |

UNIVERSITAS
Miguel
Hernández

Annex 2: Supplementary data from article 2, Parker LA et al, 2010.

**Figure 1: Flow diagram of the selection process.**



Annex 2 continued: Supplementary data from article 2, Parker LA et al, 2010.

**Annex 1: List of 45 articles evaluated.**


1. Belluco C, Petricoin EF, Mammano E, Facchiano F, Ross-Rucker S, Nitti D et al. Serum proteomic analysis identifies a highly sensitive and specific discriminatory pattern in stage 1 breast cancer. Ann Surg Oncol. 2007;14:2470-6
2. Bhattacharyya S, Epstein J, Suva LJ. Biomarkers that discriminate multiple myeloma patients with or without skeletal involvement detected using SELDI-TOF mass spectrometry and statistical and machine learning tools. Dis Markers. 2006;22:245-55
3. Bons JA, Drent M, Bouwman FG, Mariman EC, van Dieijen-Visser MP, Wodzig WK. Potential biomarkers for diagnosis of sarcoidosis using proteomics in serum. Respir Med. 2007;101:1687-95
4. Buhimschi CS, Bhandari V, Hamar BD, Bahtiyar MO, Zhao G, Sfakianaki AK et al. Proteomic profiling of the amniotic fluid to detect inflammation, infection, and neonatal sepsis. PLoS Med. 2007;4:e18
5. Buhimschi IA, Zambrano E, Pettker CM, Bahtiyar MO, Paidas M, Rosenberg VA et al. Using proteomic analysis of the human amniotic fluid to identify histologic chorioamnionitis. Obstet Gynecol. 2008 ;111:403-12
6. Cepek L, Brechlin P, Steinacker P, Mollenhauer B, Klingebiel E, Bibl M et al. Proteomic analysis of the cerebrospinal fluid of patients with Creutzfeldt-Jakob disease. Dement Geriatr Cogn Disord. 2007;23:22-8
7. Das S, Maeso PA, Becker AM, Prosser JD, Adam BL, Kountakis SE. Proteomics blood testing to distinguish chronic rhinosinusitis subtypes. Laryngoscope. 2008;118:2231-4
8. Finehout EJ, Franck Z, Choe LH, Relkin N, Lee KH. Cerebrospinal fluid proteomic biomarkers for Alzheimer's disease. Ann Neurol. 2007;61:120-9
9. Han KQ, Huang G, Gao CF, Wang XL, Ma B, Sun LQ et al. Identification of lung cancer patients by serum protein profiling using surface-enhanced laser desorption/ionization time-of-flight mass spectrometry. Am J Clin Oncol. 2008;31:133-9
10. He QY, Zhu R, Lei T, Ng MY, Luk JM, Sham P et al. Toward the proteomic identification of biomarkers for the prediction of HBV related hepatocellular carcinoma. J Cell Biochem. 2008;103:740-52
11. Hong M, Zhang X, Hu Y, Wang H, He W, Mei H et al. The potential biomarkers for thromboembolism detected by SELDI-TOF-MS. Thromb Res. 2009;123:556-64
12. Jacot W, Lhermitte L, Dossat N, Pujol JL, Molinari N, Daurès JP et al. Serum proteomic profiling of lung cancer in high-risk groups and determination of clinical outcomes. J Thorac Oncol. 2008;3:840-50.
13. Kyselova Z, Mechref Y, Kang P, Goetz JA, Dobrolecki LE, Sledge GW et al. Breast cancer diagnosis and prognosis through quantitative measurements of serum glycan profiles. Clin Chem. 2008;54:1166-75
14. Leiserowitz GS, Lebrilla C, Miyamoto S, An HJ, Duong H, Kirmiz C et al. Glycomics analysis of serum: a potential new biomarker for ovarian cancer? Int J Gynecol Cancer. 2008 ;18:470-5
15. Liang Y, Fang M, Li J, Liu CB, Rudd JA, Kung HF et al. Serum proteomic patterns for gastric lesions as revealed by SELDI mass spectrometry. Exp Mol Pathol. 2006;81:176-80
16. Lin YW, Lai HC, Lin CY, Chiou Jy, Shui HA, Chang CC et al. Plasma proteomic profiling for detecting and differentiating in situ and invasive carcinomas of the uterine cervix. Int J Gynecol Cancer. 2006; 16:1216-24
17. Martínez-Llordella M, Lozano JJ, Puig-Pey I, Orlando G, Tisone G, Lerut J et al. Using transcriptional profiling to develop a diagnostic test of operational tolerance in liver transplant recipients. J Clin Invest. 2008;118:2845-57
18. McLerran D, Grizzle WE, Feng Z, Bigbee WL, Banez LL, Cazares LH et al. Analytical validation of serum proteomic profiling for diagnosis of prostate cancer: sources of sample bias. Clin Chem. 2008;54:44-52
19. McLerran D, Grizzle WE, Feng Z, Thompson IM, Bigbee WL, Cazares LH et al. SELDI-TOF MS whole serum proteomic profiling with IMAC surface does not reliably detect prostate cancer. Clin Chem. 2008;54:53-60
20. Meuwis MA, Fillet M, Geurts P, de Seny D, Lutteri L, Chapelle JP et al. Biomarker discovery for inflammatory bowel disease, using proteomic serum profiling. Biochem Pharmacol. 2007;73:1422-33
21. Monzon FA, Lyons-Weiler M, Buturovic LJ, Rigl CT, Henner WD, Sciulli C et al. Multicenter validation of a 1,550-gene expression profile for identification of tumor tissue of origin. J Clin Oncol. 2009;27:2503-8
22. Mosley K, Tam FWK, Edwards RJ, Crozier J, Pusey CD, Lightstone L. Urinary proteomic profiles distinguish between active and inactive lupus nephritis. Rheumatology. 2006;45:1497-504

23. Ordway JM, Budiman MA, Korshunova Y, Maloney RK, Bedell JA, Citek RW et al. Identification of novel high-frequency DNA methylation changes in breast cancer. PLoS One. 2007;2:e1314
24. Pasinetti GM, Unger LH, Lange DJ, Yemul S, Deng H, Yuan X et al. Identification of potential CSF biomarkers in ALS. Neurology. 2006;66:1218-22
25. Petri AL, Simonsen AH, Yip TT, Hogdall E, Fung ET, Lundvall L et al. Three new potential ovarian cancer biomarkers detected in human urine with equalizer bead technology. Acta Obstet Gynecol Scand. 2009;88:18-26
26. Poon TCW, Sung JJY, Chow SM, Ng EKW, Yu ACW, Chu ESH et al. Diagnosis of Gastric cancer by serum proteomic fingerprinting. Gastroenterology. 2006;130:1858-64.
27. Reddy A, Wang H, Yu H, Bonates TO, Gulabani V, Azok J et al. Logical Analysis of Data (LAD) model for the early diagnosis of acute ischemic stroke. BMC Med Inform Decis Mak. 2008;8:30
28. Ren H, Du N, Liu G, Hu HT, Tian W, Deng ZP et al. Analysis of variabilities of serum proteomic spectra in patients with gastric cancer before and after operation. World J Gastroenterol. 2006;12:2789-92
29. Sanders ME, Dias EC, Xu BJ, Mobley JA, Billheimer D, Roder H et al. Differentiating proteomic biomarkers in breast cancer by laser capture microdissection and MALDI MS. J Proteome Res. 2008;7:1500-7
30. Scarlett CJ, Smith RC, Saxby A, Nielson A, Samra JS, Wilson SR et al Proteomic Classification of Pancreatic Adenocarcinoma Tissue Using Protein Chip Technology. Gastroenterology. 2006;130:1670–8
31. Sogawa K, Itoga S, Tomonaga T, Nomura F. Diagnostic values of surface-enhanced laser desorption/ionization technology for screening of habitual drinkers. Alcohol Clin Exp Res. 2007;31:S22-6.
32. Srinivasan R, Daniels J, Fusaro V, Lundqvist A, Killian JK, Geho D et al. Accurate diagnosis of acute graft versus host disease using serum proteomic pattern analysis. Exp Hematol. 2006;34:796-801.
33. Su Y, Shen J, Qian H, Ma H, Ji J, Ma H et al. Diagnosis of gastric cancer using decision tree classification of mass spectral data. Cancer Sci. 2007;98:37-43
34. Theodorescu D, Wittke S, Ross MM, Walden M, Conaway M, Just I et al. Discovery and validation of new protein biomarkers for urothelial cancer: a prospective analysis. Lancet Oncol. 2006;7:230-40.
35. Wada-Isoe K, Michio K, Imamura K, Nakaso K, Kusumi M, Kowa H et al. Serum proteomic profiling of dementia with Lewy bodies: diagnostic potential of SELDI-TOF MS analysis. J Neural Transm. 2007;114:1579-83
36. Wang L, Zheng W, Mu L, Zhang SZ. Identifying biomarkers of endometriosis using serum protein fingerprinting and artificial neural networks. Int J Gynaecol Obstet. 2008;101:253-8
37. Ward DG, Suggett N, Cheng Y, Wei W, Johnson H, Billingham LJ et al. Identification of serum biomarkers for colon cancer by proteomic analysis. Br J Cancer. 2006;94:1898-905
38. Wei YS, Zheng YH, Liang WB, Zhang JZ, Yang ZH, Lv ML et al. Identification of serum biomarkers for nasopharyngeal carcinoma by proteomic analysis. Cancer. 2008;112:544-51.
39. Weissinger EM, Schiffer E, Hertenstein B, Ferrara JL, Holler E, Stadler M et al. Proteomic patterns predict acute graft-versus-host disease after allogeneic hematopoietic stem cell transplantation. Blood. 2007;109:5511-9
40. Wu C, Wang Z, Liu L, Zhao P, Wang W, Yao D et al. Surface enhanced laser desorption/ionization profiling: New diagnostic method of HBV-related hepatocellular carcinoma. J Gastroenterol Hepatol. 2009;24:55-62
41. Wu SP, Lin YW, Lai HC, Chu TY, Kuo YL, Liu HS. SELDI-TOF MS profiling of plasma proteins in ovarian cancer. Taiwan J Obstet Gynecol. 2006;45:26-32
42. Yildiz PB, Shyr Y, Rahman JS, Wardwell NR, Zimmerman LJ, Shakhtour B et al. Diagnostic accuracy of MALDI mass spectrometric analysis of unfractionated serum in lung cancer. J Thorac Oncol. 2007;2:893-901
43. Zhang X, Wang B, Zhang XS, Li ZM, Guan ZZ, Jiang WQ. Serum diagnosis of diffuse large B-cell lymphomas and further identification of response to therapy using SELDI-TOF-MS and tree analysis patterning. BMC Cancer. 2007;7:235-46
44. Zhou L, Cheng L, Tao L, Jia X, Lu Y, Liao P.Detection of hypopharyngeal squamous cell carcinoma using serum proteomics. Acta Oto-laryngol. 2006;126:853-60
45. Zhu LR, Zhang WY, Yu L, Zheng YH, Hu J, Liao QP. Proteiomic patterns for endometrial cancer using SELDI-TOF-MS. J Zhejiang Univ Sci B. 2008;9:286-90

UNIVERSITAS
Miguel
Hernández

**Table S1: Characteristics of 45 studies evaluating the diagnostic use of an '-omics' based test.**

| 1ST AUTHOR | YEAR | JOURNAL | PHASE | N | TARGET DISORDER | INDEX TEST | REFERENCE STANDARD |
|---|---|---|---|---|---|---|---|
| Belluco | 2007 | Ann Surg Oncol | I | 310 | Breast cancer | Serum proteomic profiles using SELDI-TOF-MS | Pathologically proven disease and mammography negative controls |
| Bhattacharayya | 2006 | Dis Markers | I | 48 | Bone involvement in multiple myeloma | Serum proteomic profiles using SELDI-TOF-MS | Established diagnosis with radiological evidence of bone involvement |
| Bons | 2007 | Respir Med | I | 70 | Sarcoidosis | Serum proteomic profiles using SELDI-TOF-MS | Established diagnosis |
| Buhimschi | 2008 | Obstet Gynecol | IV | 158 | Chlorioamnionitis | Amniotic fluid proteomic fingerprint using SELDI-TOF-MS: Mass Restricted score | Histology |
| Buhimschi | 2007 | PLoS Med | IV | 169 | Neonatal sepsis | Amniotic fluid proteomic fingerprint using SELDI-TOF-MS: Mass Restricted score | Clinical symptoms and laboratory analysis |
| Cepek | 2007 | Dement Geriatr Cogn Disord | I | 28 | Creutzfeldt-jakob disease | CSF proteomic profile using 2DGE | Established diagnosis |
| Das | 2008 | Laryngoscope | IV | 42 | Chronic rhinosinusitis subtypes | Serum proteomic profiles using SELDI-TOF-MS | Fulfilment of established diagnostic criteria depending on subtype |
| Finehout | 2007 | Ann Neurol | I | 96 | Alzheimer's disease | CSF proteomic profile using 2DGE | Established diagnosis |
| Han | 2008 | Am J Clin Oncol | I | 253 | Lung cancer | Serum proteomic profiles using SELDI-TOF-MS | Established diagnosis and healthy controls |
| He | 2008 | J Cell Biochem. | I | 164 | HBV related hepatocellular carcinoma | Serum proteomic profiles using SELDI-TOF-MS | Established diagnoses and healthy controls |
| Hong | 2009 | Thromb Res | I | 69 | Thromboembolism | Plasma proteomic profiles usinf SELDI-TOF-MS | Established diagnosis and healthy controls |
| Jacot | 2008 | J Thorac Oncol | II | 170 | Lung cancer | Serum proteomic profiles | Established diagnoses |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | using SELDI-TOF-MS | |
| Kyselova | 2008 | Clin Chem | I | 109 | Breast cancer | Serum glycomic profile using MALDI-MS | Established diagnosis and healthy controls |
| Leiserowitz | 2008 | Int J Gynecol Cancer | I | 72 | Ovarian cancer | Serum glycomic profile using MALDI-FTMS | Established diagnosis and normal controls |
| Liang | 2006 | Exp Mol Pathol | I | 127 | Gastric lesions | Serum proteomic profiles using SELDI-MS | Established diagnosis and healthy controls |
| Lin | 2006 | Int J Gynecol Cancer | I | 129 | Cervical cancer | Plasma proteomic profiles usinf SELDI-TOF-MS | Established diagnosis and healthy controls |
| Martínez-Llordella | 2008 | J Clin Invest | I | 96 | Tolerance in liver transplants | Gene expression profiles in PBMC using oligonucleotide microarray | Established diagnosis |
| McLerran | 2008 | Clin Chem | II | 400 | Prostate cancer | Serum proteomic profiles using SELDI-TOF-MS | Established diagnoses |
| McLerran | 2008 | Clin Chem | I | 544 | Prostate cancer | Serum proteomic profiles using SELDI-TOF-MS | Established diagnoses and normal controls |
| Meuwis | 2007 | Biochem Pharmacol | II | 120 | Inflammatory bowel disease | Serum proteomic profiles using SELDI-TOF-MS | Established diagnosis |
| Monzon | 2009 | J Clin Oncol | IV | 547 | Tumour tissue of origin | Gene expression profiles in tissue using microarray | Histology |
| Mosley | 2006 | Rheumatology | I | 57 | Active lupus nephritis | Urinary proteomic profiles using SELDI-TOF-MS | Established diagnoses |
| Ordway | 2007 | PLoS One | I | 230 | Breast cancer | DNA methylation profiles is tissue using microarray | Established diagnosis and normal controls |
| Pasinetti | 2006 | Neurology | I | 102 | Amyotrophic Lateral Sclerosis | CSF proteomic profile using SELDI-MS | Established diagnoses and normal controls |
| Petri | 2009 | Acta Obstet Gynecol Scand | IV | 209 | Ovarian cancer | Urinary proteomic profiles using SELDI-TOF-MS | Surgery and histopathology |
| Poon | 2006 | Gastroenterology | I | 123 | Gastric cancer | Serum proteomic profiles using SELDI-MS | Established diagnosis and healthy controls |
| Reddy | 2008 | BMC Med Inform Decis Mak | I | 130 | Ischemic stroke | Logistic analysis of data applied to serum proteomic | Established diagnosis and healthy samples from blood |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | profiles | bank |
| Ren | 2006 | World J Gastroenterol | I | 86 | Gastric cancer | Serum proteomic profiles using SELDI-TOF-MS | Established diagnosis and healthy controls |
| Sanders | 2008 | J Proteome Res | I | 289 | Breast cancer | Tissue proteomic patterns using MALDI-MS | Established diagnosis and reduction mammoplasty specimens |
| Scarlett | 2006 | Gastroenterology | I | 50 | Pancreatic adenocarcinoma | Tissue proteomic patterns using SELDI-TOF-MS | Histology |
| Sogawa | 2007 | Alcohol Clin Exp Res | I | 75 | Alcoholism | Serum proteomic profiles using SELDI-TOF-MS | Questionnaire |
| Srinivasen | 2006 | Exp Hematol | I | 34 | Graft-versus-host disease | Serum proteomic profiles using SELDI-TOF-MS | Clinical findings and partial histopathological analysis |
| Su | 2007 | Cancer Sci | I | 245 | Gastric cancer | Serum proteomic profiles using SELDI-TOF-MS | Established diagnoses and healthy controls |
| Theodorescu | 2006 | Lancet Oncol | III | 655 | Urothelial cancer | Urinary proteomic profiles using CE-MS | Various established diagnoses and healthy controls |
| Wada-Isoe | 2007 | J Neural Transm | I | 52 | Dementia with Lewis bodies | Serum proteomic profiles using SELDI-TOF-MS | Established diagnosis |
| Wang | 2008 | Int J Gynaecol Obstet | I | 66 | Endometriosis | Serum proteomic profiles using SELDI-TOF-MS | Established diagnosis and healthy controls |
| Ward | 2006 | Br J Cancer | I | 93 | Colorectal cancer | Serum proteomic profiles using SELDI-MS | Established diagnosis and healthy controls |
| Wei | 2008 | Cancer | I | 168 | Naspharyngeal carcinoma | Serum proteomic profiles using SELDI-TOF-MS | Established diagnosis and healthy volunteers |
| Weissinger | 2007 | Blood | IV | 141 | Graft-versus-host disease | Urinary proteomic profiles using CE-MS | Histopathologic examination of tissue biopsies |
| Wu | 2006 | Taiwan J Obstet Gynecol | I | 65 | Ovarian cancer | Plasma proteomic profiles using SELDI-TOF-MS | Established diagnosis and healthy controls |
| Wu | 2009 | J Gastroenterol Hepatol | I | 59 | Hepatocellular carcinoma | Serum proteomic profiles using SELDI-TOF-MS | Established diagnoses |
| Yildiz | 2007 | J Thorac Oncol | I | 288 | Lung cancer | Serum proteomic profiles using MALDI-MS | Established diagnosis and healthy controls |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Zhang | 2007 | BMC Cancer | I | 207 | Large B-cell lymphomas | Serum proteomic profiles using SELDI-TOF-MS | Established diagnosis and healthy controls |
| Zhou | 2006 | Acta Oto-laryngol | I | 100 | Hypopharngeal squamous cell carcinoma | Serum proteomic profiles using SELDI-TOF-MS | Established diagnosis and healthy controls |
| Zhu | 2008 | J Zhejiang Univ Sci B | I | 100 | Endometrial cancer | Serum proteomic profiles using SELDI-TOF-MS | Established diagnosis and healthy volunteers |

**Abbreviations:** 2DGE, two-dimensional gel electrophoresis; CE-MS, capillary electrophoresis coupled online to mass spectrometry; CSF, cerebrospinal fluid; HBV, hepatitis B virus; MALDI-MS, matrix-assisted laser desorption ionization mass spectrometry; MALDI-FTMS, matrix-assisted laser desorption ionization fourier transformation mass spectrometry; PBMC, Peripheral blood mononuclear cell; SELDI-TOF-MS, surface enhanced laser desorption ionization time of flight mass spectrometry.

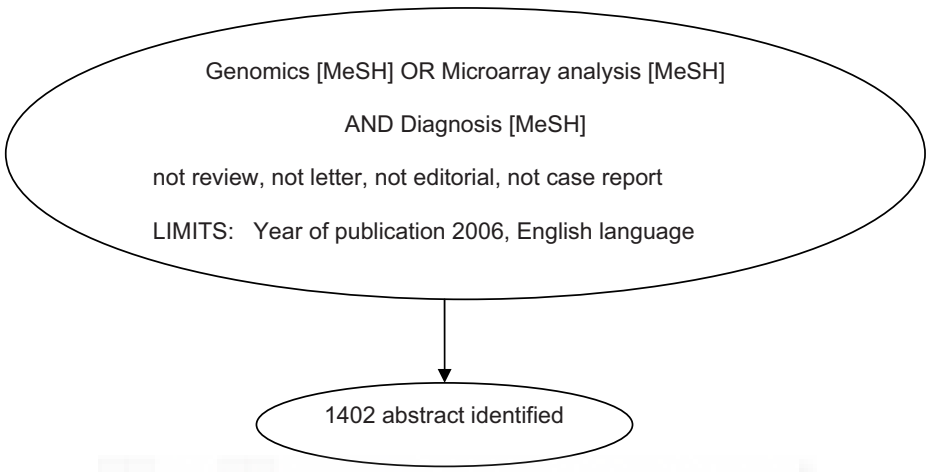
Annex 3: Supplementary data from article 3, Lumbreras B et al, 2009.

UNIVERSITAS
Miguel
Hernández

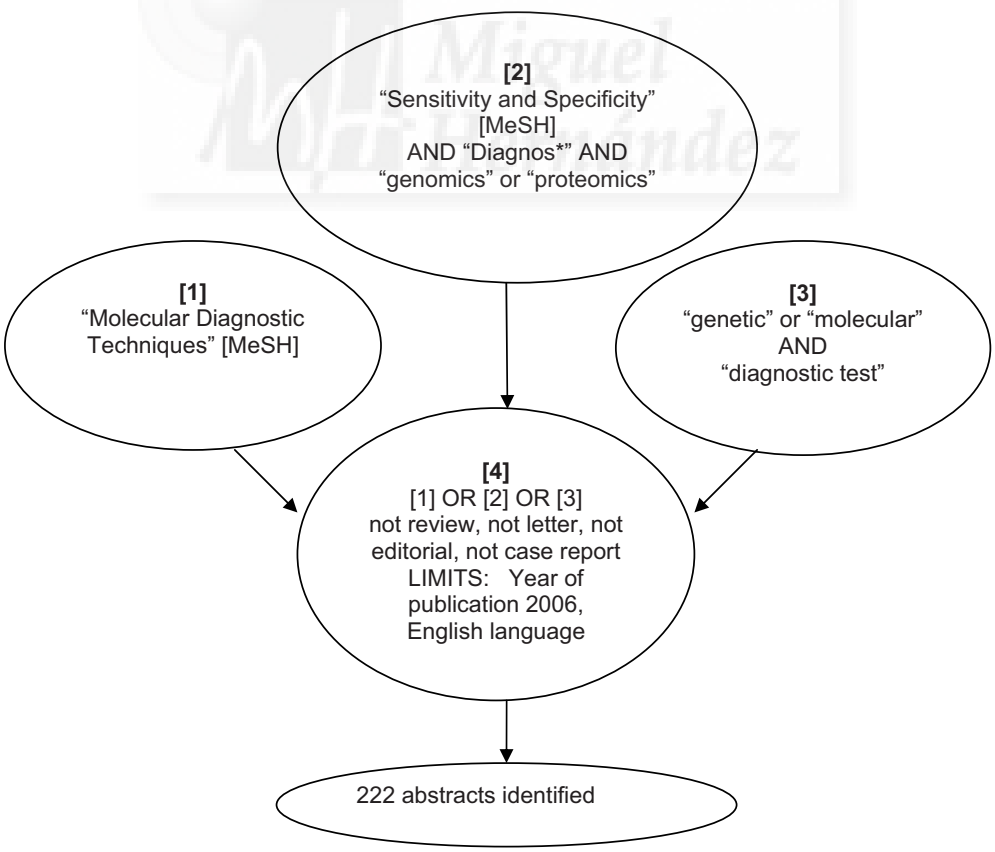
**Figure 1: Full search strategy carried out on 11th May 2007**

Aim: To identify all original articles published in English in 2006 investigating the diagnostic value of molecular, genetic or proteomic tests.

**MAIN SEARCH**




**ADDTIONAL SEARCH TO IMPROVE SENSITIVITY**



## Annex I: Transcriptions of selected articles

| Reference in annex 2 | Study design | Statements regarding use for clinical diagnosis | Statements regarding further validation studies | Overinterpretation |
|---|---|---|---|---|
| (25) | Alternative diagnosis control | **Definitively favourable**: 'MS-MA appears to be an efficient primary method to diagnose PWS/AS. Therefore, this rapid MS-MA is a good primary screening method that can be implemented in a diagnostic laboratory to determine the methylation patterns of patients with suspected PWS or A'. | **Further validation:** **'**Additional study with more samples and different types of pathogenesis for both PWS and AS may be necessary to determine the sensitivity and accuracy of both the MS-MA and MS-MLPA assays'. | Yes |
| (41) | Healthy control | **Definitively favourable**: 'The results suggest that a peripheral blood mononuclear cell-based gene expression signature can provide a molecular biomarker that can complement the standard diagnosis of UC and CD'. | **Further validation:** 'If prospectively validated in a larger population, may provide the basis for a molecular diagnosis of UC and CD and contribute to the diagnosis of patients classified as indeterminate IBD'. | Yes |
| (26) | Alternative diagnosis control | **Definitively favourable**: 'These assays may be reliably applied as a diagnostic test or large scale method for population screening'. | **Not mention of further validation.** | Yes |
| (48) | Alternative diagnosis control | **Definitively favourable**: 'We therefore conclude that elevations of MLCLs are specific for BTHS and that the MLCL/CL ratio in fibroblasts is a better diagnostic marker than CL alone' | **Not mention of further validation.** | Yes |
| (29) | Healthy control | **Promising:** 'These data indicated that these peak could be used as potential biomarkers for gastric cancer' | **Further validation:** 'Additional studies are required to validate these patterns as unique "malignant" protein signatures before they can be used with confidence to identify and screen populations at high risk for gastric cancer'. | No |

| Reference in annex 2 | Study design | Statements regarding use for clinical diagnosis | Statements regarding further validation studies | Overinterpretation |
|---|---|---|---|---|
| (40) | Healthy control | **Promising:** 'This study shows that free-circulating DNA can be detected in cancer patients compared with disease-free individuals, and suggests a new, non invasive approach for early detection of cancer'. | **Further validation:** 'Further studies are needed to understand the correlation of these new molecular markers with cancer diagnosis, outcome of disease, and eventually treatment response' | No |
| (30) | Healthy control | **Promising: '**The present study of the CSF proteins secreted in patients with INPH suggests that certain CSF proteins may be useful adjuncts in the clinical diagnosis of INPH'. | **Not mention of further validation.** | Yes |
| (35) | Alternative diagnosis control | **Promising:** 'The ability to demonstrate MSI in heterogenous endometrial samples suggests potential for the development of a novel EC screening tool for women in HNPCC kindreds' | **Not mention of further validation.** | Yes |
| (36) | Other: series of cases | **Unfavourable:** 'Screening for MMR deficiency should not be applied routinely in adenomas with the goal to identify HNPCC patients'. | **Not mention of further validation.** | No |
| (37) | Other: series of cases | **Promising: '**Real-time PCR on CSF samples seems a promising adjunct for diagnosis of mumps meningitis, especially in an age group with high incidence of mumps'- | **Not mention of further validation.** | Yes |
| (39) | Other: series of cases | **Promising:** 'It suggests that the assay will work on samples that are more likely to be poorly differentiated and more representative of the true clinical dilemma'. | **Further validation:** 'Further validation of the assay with larger numbers of true and resolved CUP samples will be needed to assess not only the true clinical value of such molecular techniques but also the ability of new information to impact survival and quality of life'. | No |

| Reference in annex 2 | Study design | Statements regarding use for clinical diagnosis | Statements regarding further validation studies | Overinterpretation |
|---|---|---|---|---|
| (43) | Other: highly selected cohort supplies by CDC to include all types. | **Definitively favourable:** 'The ability to rapidly identify new, potentially pandemic strains of influenza virus will allow health care officials to more rapidly respond and, potentially, reduce the spread and human impact of the disease'. | **Further validation:** 'Other plans include further studies with larger numbers and varieties of isolates and patient samples'. | Yes |
| (71) | Other | **Definitively favourable:** 'This study demonstrates that microarray is useful for simultaneous monitoring of several viruses and their subtypes'. | **Not mention of further validation.** | Yes |
| (50) | Clinically relevant population | **Unfavourable:** 'However, many challenges remain before PCR can be recommended for the diagnosis of sepses. | **Further validation:** 'The success of this approach must be proven on a much larger scale using multiple sites' | - No |
| (87) | Clinically relevant population | **Definitively favourable**: 'Component-based testing and the whole-allergen CAP are equally relevant in the diagnosis of grass-, birch- and cat-allergic patients'. | **Further validation:** 'The clinical relevance of each allergen needs to be validated separately prior to the implementation of multi-allergen panels into routine diagnostic settings'. | - No (according to study design)<br>- No (according to accuracy): Sensitivity: 72%; specificity: 92%. |
| (105) | Clinically relevant population | **Definitively favourable**: 'Use of array-CGH should increase the detection of abnormalities relative to the risk, and is an option for an enhanced level of screening for chromosomal abnormalities in high risk pregnancies'. | **Further validation:** 'Additional large-scale studies are required in order to determine whether array-CGH may eventually replace a karyotype in routine prenatal diagnoses. | - No (according to study design)<br>- Yes, due to lack of accuracy index. |
| (38) | Clinically relevant population | **Definitively favourable**: 'A semi-automated and simplified molecular diagnostic protocol for the rapid detection of Norovirus has been achieved'. | **Not mention of further validation.** | - No (according to study design)<br>- No (according to accuracy): Sensitivity: 100%; specificity: 66%. |
| (42) | Clinically relevant population | **Definitively favourable**: 'Our results indicated that the developed assay is reliable as well as time and cost effective for clinical diagnosis of chromosome 22q11.2 deletion'. | **Not mention of further validation.** | - No (according to study design)<br>- No (according to accuracy): Accuracy 100%. |

| Reference in annex 2 | Study design | Statements regarding use for clinical diagnosis | Statements regarding further validation studies | Overinterpretation |
|---|---|---|---|---|
| (54) | Clinically relevant population | **Definitively favourable**: 'This PCR assay detects a variety of strains exhibiting characteristics of the EAEC group, making it a useful tool for identifying both typical and atypical EAEC'. | **Not mention of further validation.** | - No (according to study design)<br>- Yes, due to lack of accuracy index. |
| (91) | Clinically relevant population | **Promising:** 'Using SELDI-TOF analysis of 195 unique specimens, we discovered with preliminary validation six distinct peaks that may potentially be useful in the detection and monitoring of ovarian cancer'. | **Further validation:** 'Additional studies are going on to further identify and validate these biomarkers' | No |

**Annex 2: References of 108 articles analysed.**


1- Villanueva J, Martorella AJ, Lawlor K et al. Serum peptidome patterns that distinguish metastatic thyroid carcinoma from cancer-free controls. Mol Cell Proteomics 2006;5:1840-52.
2- Poon LL, Wong BWY, Ma EHT et al. Sensitive and inexpensive molecular test for falciparum malaria: detecting Plasmodium falciparum DNA directly from heat-treated blood by loop-mediated isothermal amplification. Clin Chem 2006;52:303-6.
3- Benlloch S, Galbis-Caravajal JM, Martín C et al. Potential diagnostic value of methylation profile in pleural fluid and serum from cancer patients with pleural effusion. Cancer 2006;107:1859-65.
4- Ward DG, Cheng Y, N'Kontchou C et al. Changes in the serum proteome associated with the development of hepatocellular carcinoma in hepatitis C-related cirrhosis. Br J Cancer 2006; 94:287-292.
5- Grote HJ, Schmiemann V, Geddert H et al. Methylation of RAS association domain family protein 1A as a biomarker of lung cancer. Cancer 2006;108:129-34.
6- Kern W, Kohlman A, Schoch C et al. Comparison of mRNA abundance quantified by gene expression profiling and percentage of positive cells using immunophenotyping for diagnostic antigens in acute and chronic leukemias. Cancer 2006;107:2401-7.
7- Florent M, Kasahian S, Vekhoff A et al. Prospective Evaluation of a Polymerase Chain Reaction–ELISA Targeted to Aspergillus fumigatus and Aspergillus flavus for the Early Diagnosis of Invasive Aspergillosis in Patients with Hematological Malignancies. J Infec Dis 2006;193:741-7.
8- Dawson ED, Moore C, Smagala JA et al. MChip: A Tool for Influenza Surveillance. Anal Chem 2006;78:7610-5.
9- Meyer-Monard S, Parlier V, Passweg J et al. Combination of broad molecular screening and cytogenetic analysis for genetic risk assignment and diagnosis in patients with acute leucemia. Leukemia 2006;20:247-53.
10- Scarlett CJ, Smith RC, Saxby A et al. Proteomic Classification of Pancreatic Adenocarcinoma Tissue Using Protein Chip Technology. Gastroenterology 2006;130:1670-8.

11- Theodorescu D, Wittke S, Ross MM et al. Discovery and validation of new protein biomarkers for urothelial cancer: a prospective analysis. Lancet Oncol 2006;7:230-40.

12- Agranoff D, Fernandez-Reyes D, Papadopoulos MC et al. Identification of diagnostic markers for tuberculosis by fingerprinting of serum. Lancet 2006;368:1012-21.

13- Srinivasan R, Daniels J, Fusaro V et al. Accurate diagnosis of acute graft versus host disease using serum proteomic pattern analysis. Exp Hematol 2006;34:796-801.

14- Hye A, Lynham S, Thambisetty M et al. Proteome-based plasma biomarkers for Alzheimer's disease. Brain 2006;129: 3042-50.

15- Guillaud M, Benedet JL, Cantor SB, Staerkel G, Follen M, MacAulay C. DNA ploidy compared with Human Papilloma virus testing and conventional cervical cytology as a primary screening test for cervical high grade lesions and cancer in 1555 patients with biopsy confirmation. Cancer 2006;107:309-18.

16- Mosley K, Tam FWK, Edwards RJ, Crozier J, Pusey CD, Lightstone L. Urinary proteomic profiles distinguish between active and inactive lupus nephritis. Rheumatology 2006;45:1497-504.

17- Shi Q, Harris LN, Lu X et al. Declining plasma fibrinogen alpha fragment identifies HER2-positive breast cancer patients and reverts to normal levels after surgery. J Proteome Res 2006;5:2947-55.

18- Holloway AJ, Diyagama DS, Opeskin K et al. A molecular diagnostic test for distinguishing lung adenocarcinoma from malignant mesothelioma using cells collected from pleural effusions. Clin Cancer Res 2006;12:5129-35.

19- Poon TCW, Sung JJY, Chow SM et al. Diagnosis of Gastric cancer by serum proteomic fingerprinting. Gastroenterology 2006;130:1858-64.

20- Ward DG, Suggett N, Cheng Y et al. Identification of serum biomarkers for colon cancer by proteomic analysis. Br J Cancer 2006;94:1898-905.

21- Paterson PJ, Seaton S, McHugh TD et al. Validation and clinical application of Molecular methods for the identification of Molds in Tissue. Clin Infect Dis 2006;42:51-6.

22- Tester DJ, Will MS, Haglund CM, Ackerman MJ. Effect of clinical phenotype on yield of long QT syndrome genetic testing. J Am Coll Cardiol 2006;47:764-8.

23- Pasinetti GM, Unger LH, Lange DJ et al. Identification of potential CSF biomarkers in ALS. Neurology 2006;66:1218-22.
24- Zhou Y, Lum JMS, Yeo GH, Kiing J, Tay SKT, Chong SS. Simplified molecular diagnosis of fragile X syndrome by fluorescent methylation-specific PCR and GeneScan analysis. Clin Chem 2006;52:1492-500.
25- Procter M, Chou LS, Tang W, Jama M, Mao R. Molecular diagnosis of Prader-Willi and Angelman syndromes by methylation-specific melting analysis and methylation-specific multiplex ligation-dependent probe amplification. Clin Chem 2006;52:1276-83.
26- Alsmadi OA, Al-Kayal F, Al-Hamed M, Meyer BF. Frequency of common HFE variants in the Saudi population: a high throughput molecular beacon-based study. BMC Med Genet 2006;7:43.
27- Yannaraki M, Rebibou JM, Ducloux D et al. Urinary cytotoxic molecular markers for a noninvasive diagnosis in acute renal transplant rejection. Transpl Int 2006;19:759-68.
28- Ren H, Du N, Liu G et al. Analysis of variabilities of serum proteomic spectra in patients with gastric cancer before and after operation. World J Gastroenterol 2006;12:2789-92.
29- Liang Y, Fang M, Li J et al. Serum proteomic patterns for gastric lesions as revealed by SELDI mass spectrometry. Exp Mol Pathol 2006;81:176-80.
30- Li X, Miyajima M, Mineki R, Taka H, Murayama K, Ari H. Analysis of potential diagnostic biomarkers in cerebrospinal fluid of idiopathic normal pressure hydrocephalus by proteomics. Acta Neurochir (Wien) 2006;148:859-64.
31- Zhou L, Cheng L, Tao L, Jia X, Lu Y, Liao P. Detection of hypopharyngeal squamous cell carcinoma using serum proteomics. Acta Oto-laryngol 2006;126:853-60.
32- Satiroglu-Tufan NL, Tufan AC, Semerci CN, Bagci H. Accurate diagnosis of a homozygous G1138A mutation in the fibroblast growth factor receptor 3 gene responsible for achondroplasia. Tohoku J Exp Med 2006;208:103-7.
33- Zhu LR, Zhang WY, Yu L, Zhang JZ, Liao QP. Serum proteomic features for detection of endometrial cancer. Int J Gynecol Cancer 2006;16:1374-8.

34- Lin YW, Lai HC, Lin CY et al. Plasma proteomic profiling for detecting and differentiating in situ and invasive carcinomas of the uterine cervix. Int J Gynecol Cancer 2006;16:1216-24.
35- Hewitt MJ, Wood N, Quinton ND et al. The detection of microsatellite instability in blind endometrial samples--a potential novel screening tool for endometrial cancer in women from hereditary nonpolyposis colorectal cancer families? Int J Gynecol Cancer 2006;16:1393-400.
36- German HNPCC consortium, Müller A, Beckmann C et al. Prevalence of the mismatch-repair-deficient phenotype in colonic adenomas arising in HNPCC patients: results of a 5-year follow-up study. Int J Colorectal Dis 2006;21:632-41.
37- Krause CH, Eastick K, Ogilvie MM. Real-time PCR for mumps diagnosis on clinical specimens--comparison with results of conventional methods of virus detection and nested PCR. J Clin Virol 2006;37:184-9.
38- Antonishyn NA, Crozier NA, McDonald RR, Levett PN, Horsman GB. Rapid detection of Norovirus based on an automated extraction protocol and a real-time multiplexed single-step RT-PCR. J Clin Virol 2006;37:156-61.
39- Wang YC, Hsu HS, Chan TP, Chan JT. Molecular diagnostic markers for lung cancer in sputum and plasma. Ann NY Acad Sci 2006;1075:179-84.
40- Papadopoulou E, Davilas E, Sotiriou V et al. Cell-free DNA and RNA in plasma as a new molecular marker for prostate and breast cancer. Ann NY Acad Sci 2006;1075:235-43.
41- Burczynski ME, Peterson RL, Twine NC et al. Molecular classification of Crohn's disease and Ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells. J Mol Diagn 2006;8:51-61.
42- Chen YF, Kou PL, Tsai SJ et al. Computational analysis and refinement of sequence structure on chromosome 22q11.2 region: application to the development of quantitative real-time PCR assay for clinical diagnosis. Genomics 2006;87:290-7.
43- Townsend MB, Dawson ED, Mehlmann M et al. Experimental Evaluation of the FluChip diagnostic microarray for influenza virus surveillance. J Clin Microbiol 2006;44:2863-71.
44- Franco-Álvarez de Luna F, Ruiz P, Gutiérez J, Casal M. Evaluation of the GenoType Mycobacteria Direct assay for detection for mycobacterium

tuberculosis complex and four atypical mycobacterium species in clinical simples. J Clin Microbiol 2006;44:3025-7.

45- Deborggraeve S, Claes F, Laurent T et al. Molecular Dipstick Test for diagnosis of sleeping sickness. J Clin Microbiol 2006;44:2884-9.

46- Scarlett CJ, Saxby AJ, Neilson AQ et al. Proteomic profiling of cholangiocarcinoma: diagnostic potential of SELDI-TOF MS in malignant bile stricture. Hepatology 2006;44:658-66.

47- Wu SP, Lin YW, Lai HC, Chu TY, Kuo YL, Liu HS. SELDI-TOF MS profiling of plasma proteins in ovarian cancer. Taiwan J Obstet Gynecol 2006;45:26-32.

48- Van Werkhoven MA, Thorburn DR, Gedeon AK, Pitt JJ. Monolysocardiolipid in cultured fibroblasts is a sensitive and specific marker for Barth Syndrome. J Lipid Res 2006;47:2346-51.

49- Aulicino PC, Carrillo MG, Kopka J, Mangano AM, Ovejero M, Sen L. HIV-1 Genetic diversity in Argentina and early diagnosis of perinatal infection. Medicina (B Aires) 2006;66:319-26.

50- Jordan JA, Durso MB, Butchko AR, Jones JG, Brozanski BS. Evaluating the Near-term infant for early onset sepsis. J Mol Diagn 2006;8:357-63.

51- Talantov D, Baden J, Jatkoe T et al. A quantitative reverse transcriptase polymerase chain reaction assay to identify metastatic carcinoma tissue of origin. J Mol Diagn 2006;8:320-9.

52- Murphy KM, Zhang S, Geiger T et al. Comparison of the microsatellite instability analysis system and the Bethesda panel for the determination of Microsatellite instability in colorectal cancer. J Mol Diagn 2006;8:305-11.

53- Falk M, Vojtísková M, Lukás Z, Kroupová I, Froster U. Simple procedure for automatic detection of unstable alleles in the myotonic dystrophy and Huntington's disease loci. Genet Test 2006;10:85-97.

54- Jenkins C, Tembo M, Chart H et al. Detection of enteroaggregative Escherichia coli in faecal samples from patients in the community with diarrhoea. J Med Microbiol 2006;55:1493-7.

55- Bhattacharyya S, Epstein J, Suva LJ. Biomarkers that discriminate multiple myeloma patients with or without skeletal involvement detected using SELDI-TOF mass spectrometry and statistical and machine learning tools. Dis Markers 2006;22:245-55.

56- Abdi F, Quinn JF, Jankovic J et al. Detection of biomarkers with a multiplex quantitative proteomic platform in cerebrospinal fluid of patients with neurodegenerative disorders. J Alzheimers Dis 2006;9:293-348.
57- Varghese B, Rodrigues C, Beshmukh M et al. Broad-range bacterial and fungal DNA amplification on vitreous humor from suspected endophthalmitis patients. Mol Diagn Ther 2006;10:319-26.
58- Grossman HB, Soloway M, Messing E et al. Surveillance for recurrent bladder cancer using a point-of-care proteomic assay. JAMA 2006;295:299-305.
59- Chignard N, Shang S, Wang H et al. Cleavage of endoplasmic reticulum proteins in hepatocellular carcinoma: Detection of generated fragments in patient sera. Gastroenterology 2006;130:2010-22.
60- Huang LJ, Chen SX, Huang Y et al. Proteomics-based identification of secreted protein dihydrodiol dehydrogenase as a novel serum markers of non-small cell lung cancer. Lung Cancer 2006;54:87-94.
61- Belov L, Mulligan SP, Barber N et al. Analysis of human leukaemias and lymphomas using extensive immunophenotypes from an antibody microarray. Br J Haematol 2006;135:184-97.
62- Ziober AF, Patel KR, Alawi F et al. Identification of a gene signature for rapid screening of oral squamous cell carcinoma. Clin Cancer Res 2006;12:5960-71.
63- Sartain MJ, Slayden RA, Singh KK, Laal S, Belisle JT. Disease state differentiation and identification of tuberculosis biomarkers via native antigen array profiling. Mol Cell Proteomics 2006;5:2102-13.
64- Albrecht V, Chevallier A, Magnone V et al. Easy and fast detection and genotyping of high-risk human papillomavirus by dedicated DNA microarrays. J Virol Methods 2006;137:236-44.
65- Fujita Y, Nakanishi T, Hiramatsu M et al. Proteomics-based approach identifying autoantibody against peroxiredoxin VI as a novel serum marker in esophageal squamous cell carcinoma. Clin Cancer Res 2006;12:6415-20.
66- Sui G, Zhou S, Wang J et al. Mitochondrial DNA mutations in preneoplastic lesions of the gastrointestinal tract: A biomarker for the early detection of cancer. Mol Cancer 2006;5:73.
67- Gobel T, Vorderwülbecke S, Hauck K, Fey H, Háussinger D, Erhardt A. New multi protein patterns differentiate liver fibrosis stages and hepatocellular

carcinoma in chronic hepatitis C serum samples. World J Gastroenterol 2006;12:7604-12.

68- Takehara A, Eguchi H, Ohigashi H et al. Novel tumor marker REG4 detected in serum of patients with resectable pancreatic cancer and feasibility for antibody therapy targeting REG4. Cancer Sci 2006;97:1191-7.

69- Lee YM, Lee JY, Kim MJ et al. Hypomethylation of the protein gene product 9.5 promoter region in gallbladder cancer and its relationship with clinicopathological features. Cancer Sci 2006;97:1205-10.

70- Wang JX, Yu J, Wang L, Liu QL, Zhang J, Zheng S. Application of serum protein fingerprint in diagnosis of papillary thyroid carcinoma. Proteomics 2006;6:5344-9.

71- Jaaskelainen AJ, Maunula L. Applicability of microarray technique for the detection of noro- and astroviruses. J Virol Methods 2006;136:210-6.

72- Hever A, Roth RB, Hevezi PA, Lee J, Willhite D, White EC et al. Molecular characterisation of human adenomyosis. Mol Hum Reprod 2006;12:737-48.

73- Groathouse NA, Amin A, Marques MAM et al. Use of protein microarrays to define the humoral immune response in Leprosy Patients and Identification of disease-dtate.specific antigenic profiles. Infect Immun 2006;74:6458-66.

74- Tan X, Cai D, Wu Y et al. Comparative analysis of serum proteomes: discovery of proteins associated with osteonecrosis of the femoral head. Transl Res 2006;148:114-9.

75- Zhang H, Niu Y, Feng J, Guo H, Ye H, Cui H. Use of proteomic analysis of endometriosis to identify different protein expression in patients with endometriosis versus normal controls. Fertil Steril 2006;86:274-82.

76- Tsangaris GT, Karamessinis P, Kolialexi A et al. Proteomic analysis of amniotic fluid in pregnancies with Down syndrome. Proteomics 2006;6:4410-9.

77- Lin CY, Tsui KH, Yu CC, Yeh CW, Chang PL, Yung BYM. Searching cell-secreted proteomes for potential urinary bladder tumor markers. Proteomics 2006;6:4381-9.

78- Lee IN, Chen CH, Sheu JC et al. Identification of complement C3a as a candidate biomarker in human chronic hepatitis C and HCV-related hepatocellular carcinoma using a proteomics approach. Proteomics 2006;6:2865-73.

79- Bradford TJ, Wang X, Chinnaiyan AM. Cancer immunomics: using autoantibody signatures in the early detection of prostate cancer. Urol Oncol 2006;24:237-42.

80- Kebebew E, Peng M, Reiff E, Duh QY, Clark O, McMillan A. Diagnostic and prognostic value of cell-cycle regulatory genes in malignant thyroid neoplasms. World J Surg 2006;30:767-74.

81- Le Page C, Ouellet V, Madore J et al. From gene profiling to diagnostic markers: IL-18 and FGF-2 complement CA125 as serum-based markers in epithelial ovarian cancer. Int J Cancer 2006;118:1750-8.

82- Inoue M, Sakaguchi J, Sasagawa T, Tango M. The evaluation of human papillomavirus DNA testing in primary screening for cervical lesions in a large Japanese population. Int J Gynecol Cancer 2006;16:1007-13.

83- Li Y, Dang TA, Shen J et al. Identification of a plasma proteomic signature to distinguish pediatric osteosarcoma from benign osteochondroma. Proteomics 2006;6:3426-35.

84- Dave SS, Fu K, Wright GW et al. Molecular diagnosis of Burkitt's lymphoma. N Engl J Med 2006;354:2431-42.

85- Zhang H, Kong B, Qu X, Jia L, Deng B, Yang Q. Biomarker discovery for ovarian cancer using SELDI-TOF-MS. Gynecol Oncol 2006;102:61-6.

86- Wong YF, Cheung TH, Tsao GSW et al. Genome-wide gene expression profiling of cervical cancer in Hong Kong women by oligonucleotide microarray. Int J Cancer 2006;118:2461-9.

87- Wöhrl S, Vigl K, Zehetmayer S et al. The performance of a component-based allergen-microarray in clinical practice. Allergy 2006;61:633-9.

88- Zhu H, Hu S, Jona G et al. Severe acute respiratory syndrome diagnostics using a coronavirus protein microarray. Proc Natl Acad Sci U S A 2006;103:4011-6.

89- Mathelin C, Cromer A, Wendling C, Tomasetto C, Rio MC. Serum biomarkers for detection of breast cancers: A prospective study. Breast Cancer Res Treat 2006;96:83-90.

90- Xu W, Chen Y, Hu Y et al. Preoperatively molecular staging with CM10 ProteinChip and SELDI-TOF-MS for colorectal cancer patients. J Zhejiang Univ Sci B 2006;7:235-40.

91- Reichelt O, Múller J, von Eggeling F et al. Prediction of renal allograft rejection by urinary protein analysis using ProteinChip Arrays (surface-enhanced laser

desorption/ionization time-of-flight mass spectrometry). Urology 2006;67:472-5.

92- Kong F, White CN, Xiao X et al. Using proteomic approaches to identify new biomarkers for detection and monitoring of ovarian cancer. Gynecol Oncol 2006;100:247-53.

93- Ahn BY, Song ES, Cho YJ, Kwon OW, Kim JK, Lee NG. Identification of an anti-aldolase autoantibody as a diagnostic marker for diabetic retinopathy by immunoproteomic analysis. Proteomics 2006;6:1200-9.

94- Selman M, Pardo A, Barrera L et al. Gene expression profiles distinguish idiopathic pulmonary fibrosis from hypersensitivity pneumonitis. Am J Respir Crit Care Med 2006;173:188-98.

95- Sanchez-Carbayo M, Socci ND, Lozano JJ, Haab BB, Cordon-Cardo C. Profiling bladder cancer using targeted antibody arrays. Am J Pathol 2006;168:93-103.

96- Huang X, Wei Y, Li L et al. Serum proteomics study of the squamous cell carcinoma antigen 1 in tongue cancer. Oral Oncol 2006;42:26-31.

97- Yeh CS, Wang JY, Wu CH et al. Molecular detection of circulating cancer cells in the peripheral blood of patients with colorectal cancer by using membrane array with a multiple mRNA marker panel. Int J Oncol 2006;28:411-20.

98- Ali IU, Xiao Z, Malone W et al. Plasma proteomic profiling: search for lung cancer diagnostic and early detection markers. Oncol Rep 2006;15:1367-72.

99- Morgun A, Shulzhenko N, Perez-Diez A et al. Molecular profiling improves diagnoses of rejection and infection in transplanted organs. Circ Res 2006;98:74-83.

100- Adley BP, GuptaA, Lin F, Luan C, Teh BT, Yang XJ. Expression of kidney-specific cadherin in chromophobe renal cell carcinoma and renal oncocytoma. Am J Clin Pathol 2006;126:79-85.

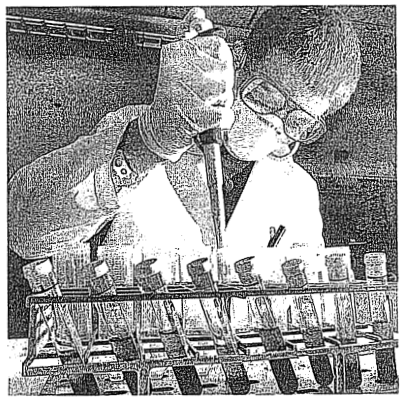
101- Huang HL, Stasky T, Morandell S et al. Biomarker discovery in breast cancer serum using 2-D differential gel electrophoresis/ MALDI-TOF/TOF and data validation by routine clinical assays. Electrophoresis 2006;27:1641-50.

102- Mérelle ME, Scheffer H, De Jong D, Dankert-Roelse JE. Extended gene analysis can increase specificity of neonatal screening for cystic fibrosis. Acta Paediatr 2006;95:1424-8.

Annex 4: Newspaper clipping referring to the third article included in the thesis: El Mundo , 9th April 2009

EL MUNDO / AÑO XVIII, NÚMERO 799 SALUD JUEVES 9 DE ABRIL DE 2009 S5

MEDICINA



Los test diagnósticos han experimentado un gran crecimiento. / EL MUNDO

INVESTIGACIÓN

# Algunos científicos pecan de exagerados

## MUCHOS ESTUDIOS VALORAN CON DEMASIADO OPTIMISMO LA APLICACIÓN CLÍNICA DE LOS MÉTODOS DE DIAGNÓSTICO MOLECULAR

MARÍA SÁNCHEZ-MONGE

La industria biotecnológica está creciendo muy rápidamente, invirtiendo ingentes cantidades de dinero en la investigación de nuevos métodos de diagnóstico molecular para un amplio abanico de enfermedades. Un equipo de epidemiólogos españoles publica en la revista *Clinical Chemistry* un estudio que, tras analizar 108 artículos recogidos en publicaciones biomédicas, demuestra que los investigadores exageraron sus posibles aplicaciones clínicas en el 56% de los casos.

Muchos de esos test llegan al mercado en un tiempo récord tras el inicio de los trabajos científicos. A esto hay que añadir que los protocolos sobre los pasos que hay que seguir para probar su utilidad clínica no se cumplen siempre, ente otras cosas porque no es obligatorio hacerlo y la normativa que regula la aprobación de estos métodos no es tan estricta como la que se encarga de dar vía libre a los nuevos medicamentos.

Los trabajos analizados valoraban la utilidad de procedimientos dedicados, sobre todo, al cáncer. Se consideró que se había producido una exageración cuando los autores calificaban como muy positivos los resultados obtenidos en su investigación y contemplaban su aplicación en los centros sanitarios en beneficio de los pacientes cuando la realidad era que los datos estadísticos recogidos en el propio artículo no avalaban afirmaciones tan categóricas.

Un ejemplo: un estudio sobre un test para el diagnóstico genético del riesgo de cáncer de endometrio resaltaba en las conclusiones que el nuevo método evaluado podría convertirse en un sistema de cribado de ese tipo de tumor en las mujeres que reuniesen una serie de condiciones. Sin embargo, en ningún momento se aludía a la necesidad de realizar más trabajos para corroborar esos resultados que, por otro lado, eran bastante preliminares.

«Si un estudio recoge las primeras fases de una investigación, sus conclusiones no pueden ser que el procedimiento se puede aplicar ya», subraya una de las autoras de la revisión crítica, Blanca Lumbreras, del departamento de Salud Pública de la Universidad Miguel Hernández de Alicante.

En esos casos, según la investigadora, lo que puede pasar es que, *ante las grandes esperanzas* que recogen las revistas científicas, los centros sanitarios gasten elevadas cantidades de dinero en procedimientos cuya utilización resulte, a la postre, decepcionante. Lumbreras cree que las consecuencias pueden ser todavía más graves: «A la hora de la verdad, estos test pueden llevar a la obtención de muchos falsos positivos o al sobrediagnóstico de ciertas patologías». Un dictamen incorrecto puede desembocar, en algunos casos, en la administración de tratamientos agresivos que en realidad no eran necesarios.

Un dato curioso del trabajo español es que, en general, las exageraciones tendían a ser mayores cuanto más alto era el factor de impacto (índice que mide la relevancia de las revistas) de la publicación en la que aparecían. También se constató que el exceso de optimismo era 18 veces superior cuando los autores trabajaban en laboratorios en comparación con los que ejercían su labor en centros asistenciales, en los que el contacto con los pacientes es mucho mayor.

UNIVERSITAS
Miguel
Hernández