VALORACIÓN DE PRUEBAS DIAGNÓSTICAS DE LABORATORIO

Blanca Lumbreras Lacarra

Departamento de Salud Pública, Historia de la Ciencia y Ginecología,

Universidad Miguel Hernández.

Alicante 2004

Directores:

Ildefonso Hernández Aguado

Eliseo Pascual Gómez

Dr. Enrique Perdiguero Gil, Director del Departamento de Salud Pública, Historia de la Ciencia y Ginecología de la Facultad de Medicina de la Universidad Miguel Hernández,

CERTIFICA

que **Blanca Lumbreras Lacarra** ha realizado bajo la coordinación de este Departamento su memoria de tesis doctoral titulada "*Valoración de pruebas diagnósticas de laboratorio*", cumpliendo todos los objetivos previstos, finalizando su trabajo en forma satisfactoria para su defensa pública y capacitándole para optar al grado de Doctor.

Lo que certifico en San Juan de Alicante, a de de dos mil cuatro.

Enrique Perdiguero Gil

Director del Departamento

ILDEFONSO HERNÁNDEZ AGUADO, Catedrático del área de Medicina preventiva y Salud Pública del Departamento de Salud Pública, Historia de la Ciencia y Ginecología de la Universidad Miguel Hernández y ELISEO PASCUAL GÓMEZ, Catedrático del área de Medicina del Departamento de Medicina Clínica de la Universidad Miguel Hernández.

CERTIFICAN

que la presente memoria para acceder al grado de Doctor, que lleva por título "*Valoración de pruebas diagnósticas de laboratorio*" de la que es autora Blanca Lumbreras Lacarra, ha sido realizada bajo su dirección.

Y para que así conste expido la presente certificación en San Juan de Alicante a de del 2004.

Fdo: Idefonso Hernández Aguado

Fdo: Eliseo Pascual Gómez

A mi familia.

AGRADECIMIENTOS

Deseo expresar mi más sincero agradecimiento a todas aquellas personas, equipos e instituciones que durante el tiempo de realización de este trabajo estuvieron implicadas de múltiples maneras y sin cuya inestimable colaboración no se hubiera podido llevar a cabo, especialmente:

Al Servicio de Reumatología del Hospital General Universitario de Alicante, por su gran colaboración en la consecución de las muestras.

A mis compañeros de residencia Fina González, Julia Frasquet y Enrique Rodríguez, por estar siempre disponibles para el análisis y apoyarme en mi trabajo.

A José Manuel Ramos por su contribución al desarrollo de esta tesis mediante la evaluación de artículos y la aportación de ideas.

A Inmaculada Jarrín, por sus análisis estadísticos.

A mis dos directores de tesis, por iniciarme en el camino de la investigación de calidad.

A mis padres, por su interés y seguimiento en el desarrollo de este trabajo.

A Diego, por animarme siempre a seguir adelante.

Cuando ante tí se abran muchos caminos y no sepas cuál recorrer, no te metas en uno cualquiera al azar: siéntate y aguarda. Respira con la confiada profundidad con que respiraste el día en que viniste al mundo, sin permitir que nada te distraiga: aguarda y aguarda más aún. Quédate quieta, en silencio, y escucha a tu corazón. Y cuando te hable, levántate y ve donde él te lleve.

(Susana Tamaro)

INDICE

1- INTRODUCCIÓN

El Diagnóstico Basado en la Evidencia (DBE) es una parte fundamental de la Medicina Basada en la Evidencia (MBE), estrategia cuyo objetivo es acercar los métodos epidemiológicos a la práctica clínica y que aporta como novedades una mejora en los procedimientos de localización de pruebas científicas y una mayor atención a la solución de problemas en pacientes concretos. Este último aspecto es clave en el campo diagnóstico, donde se parte del problema tratando de trasformarlo en una pregunta susceptible de respuesta, la cual puede responderse mediante la investigación disponible [1-3].

Si queremos que el uso de las pruebas diagnósticas en la clínica mejore, es decir, que se produzcan avances en este campo del Diagnóstico Basado en la Evidencia, es necesario realizar investigación diagnóstica de calidad y comunicarla de manera adecuada, a fin de responder a preguntas diagnósticas oportunas en el manejo clínico.

Sabemos que cuando un médico quiere diagnosticar a un paciente, observa y reconoce sus síntomas y signos, identifica las enfermedades o síndromes que de acuerdo con el cuerpo de información médica disponible y con su propia experiencia más se asemejan al cuadro clínico del paciente, y de entre ellos selecciona como más probables aquellos que se dan con más frecuencia en el contexto clínico donde desarrolla su trabajo. A continuación, hace uso de las pruebas diagnósticas complementarias que le parecen más eficaces para confirmar o descartar los diagnósticos provisionales, o para aumentar o disminuir su verosimilitud. Cuando existan varias hipótesis diagnósticas se realizará el diagnóstico diferencial y las pruebas complementarias tratarán de aclarar las dudas existentes; si solamente hay una sospecha diagnóstica, las pruebas complementarias tratarán de confirmarla o excluirla [4]. Para aplicar adecuadamente los resultados obtenidos en la investigación diagnóstica a la toma de decisiones clínicas, es necesario que dicha investigación se ejecute y comunique con rigurosidad. La baja calidad de la investigación diagnóstica determina su escasa aplicabilidad en el entorno clínico [5-8].

Si comparamos la investigación diagnóstica con la terapéutica o etiológica, comprobamos como en estas últimas se conocen y aplican los requisitos que deben reunir los estudios para alcanzar una mejor calidad, mientras que en la investigación diagnóstica el conocimiento sobre la metodología está menos estructurado y por tanto, afecta negativamente a la calidad de la investigación [9]. Llama la atención por ejemplo, la diferencia existente entre los estándares de

calidad empleados en la investigación terapéutica en comparación con los estándares diagnósticos: ante la comercialización de un nuevo fármaco son necesarios una serie de estudios, como son los ensayos clínicos desde la fase experimental en animales hasta su aplicación al paciente. Posteriormente se realiza un seguimiento postcomercialización para estudiar sus efectos positivos y la posible aparición de reacciones adversas. En cambio, cuando una prueba diagnóstica se incorpora a la práctica clínica únicamente se valora si cumple requisitos técnicos, sin tener en cuenta su eficacia para distintos grupos de población, su ámbito de aplicación, ni su repercusión sobre la salud de los pacientes.

La escasa calidad metodológica no es la única razón que impide la aplicación de los resultados de las evaluaciones diagnósticas a la práctica clínica. Gran parte de los resultados que se obtienen en la investigación no son aplicables a la realidad clínica, no por problemas de validez interna, sino porque se diseñan sin el objetivo de contestar preguntas clínicas relevantes en el área del diagnóstico [10]. En el campo diagnóstico cuando partimos de un problema concreto de identificación de la enfermedad que padece una persona y lo trasformamos en pregunta siguiendo los pasos recomendados por Sackett y colaboradores [3], comprobaremos que esta pregunta nos dará una indicación bastante precisa de cómo debe ser la investigación diagnóstica necesaria para contestarla. Es posible que la actual difusión de la Medicina Basada en la Evidencia haya contribuido a un mejor enfoque en la investigación diagnóstica justo en uno de los aspectos más deficientes hasta ahora, el diseño dirigido a contestar preguntas clínicas relevantes.

Aunque en los últimos años se han hecho algunos avances para aumentar la relación entre la medicina clínica y los métodos científicos mediante la mejora de la validez y utilidad de las pruebas diagnósticas, todavía se está lejos de lo realizado en otras áreas de investigación clínico-epidemiológica. La producción de una investigación diagnóstica válida y que responda a problemas clínicos reales puede facilitar la práctica médica, permitiendo una mejor elección de pruebas diagnósticas y por ende, un uso más costo-efectivo de los recursos sanitarios [4].

Alguno de estos avances en el campo del diagnóstico han consistido en la realización de revisiones de los estudios que evalúan pruebas diagnósticas [5-8]. Éstas han reflejado una escasa calidad en la metodología empleada, lo que ha conducido a la puesta en marcha de distintas iniciativas con el fin de mejorarla. Este trabajo pretende unirse a dichas sugerencias, mediante la valoración del estado actual de la investigación específica de pruebas de laboratorio y el análisis del impacto de las recomendaciones efectuadas, determinando los aspectos más deficientes y las mejoras encontradas.

Para obtener una visión global del estado de la investigación sobre pruebas diagnósticas, en este estudio nos vamos a centrar en dos aspectos importantes: por un lado valoraremos la calidad de los trabajos publicados que exponen la evaluación de pruebas diagnósticas de laboratorio, y por otro, y por tratarse de un aspecto más olvidado en el campo del laboratorio, realizaremos un estudio de concordancia para determinar la reproducibilidad de una prueba diagnóstica concreta y comprobaremos la aplicabilidad de las recomendaciones de Feinstein [11] para este tipo de estudios.

En el primer apartado, nos centramos en los estudios que calculan la exactitud diagnóstica de un procedimiento de laboratorio en términos de sensibilidad y especificidad. Evaluamos su calidad metodológica, analizamos los aspectos más susceptibles de mejora y comprobamos si se ha producido alguna mejora en su metodología a lo largo del periodo de estudio.

Respecto al segundo apartado, únicamente Feinstein ha publicado unas recomendaciones sobre la realización de estudios de variabilidad observacional en pruebas de laboratorio [11] y apenas se han puesto en práctica. Nosotros hemos realizado un estudio que determine la variabilidad entre varios observadores de una prueba diagnóstica de laboratorio determinada, mediante la aplicación de esta guía y de esta manera poder determinar a su vez, la aplicabilidad de sus criterios.

Es importante constatar que este estudio ha sido realizado en el contexto de los Servicios de Análisis Clínicos hospitalarios. Las razones de su elección son dos fundamentalmente: por un lado está su papel clave en la investigación de nuevas pruebas diagnósticas, ya que es en estos servicios donde se van a realizar la mayoría de los análisis, y por otro, que estos servicios suelen estar más alejados de la práctica clínica, hecho que puede dificultar el desarrollo de una investigación de

calidad, ya que impide una adecuada información sobre los problemas a resolver, así como sobre variables relevantes de la población.

Podemos resumir en tres principalmente las consecuencias de esta falta de comunicación entre servicios. Por un lado, los laboratorios ignoran los problemas diagnósticos a resolver, por lo que no pueden plantear preguntas de investigación relevantes. Por otro, los laboratorios tienen un menor acceso a información referente a los motivos de demanda de la prueba diagnóstica y las características clínicas y sociodemográficas de la población en estudio. Por último, los laboratorios conocen mejor los procedimientos, las técnicas y la interpretación de las pruebas, información que muchas veces no llega al clínico, por lo que no puede interpretar debidamente los resultados obtenidos.

Respecto al primer apartado del presente estudio referente al análisis de los trabajos que calculan la exactitud diagnóstica de un procedimiento de laboratorio, iniciamos la introducción, describiendo las principales características que estos estudios de sensibilidad y especificidad diagnóstica deben cumplir, como son los rasgos relacionados con la pregunta de investigación planteada y los relacionados con la manera en que se realiza la evaluación de la prueba de laboratorio a estudio.

1.1- DESCRIPCIÓN DE LOS ESTUDIOS QUE DETERMINAN LOS PARÁMETROS DE SENSIBILIDAD Y ESPECIFICIDAD DIAGNÓSTICA DE UNA PRUEBA DE LABORATORIO.

El tipo de investigación sobre pruebas diagnósticas que con más frecuencia se encuentra en la literatura médica es aquel que trata de determinar la exactitud de un procedimiento diagnóstico [10]. En estos estudios de exactitud diagnóstica, los resultados de una o más pruebas diagnósticas realizadas en un grupo de pacientes sospechosos de tener la condición de interés se comparan con un estándar de referencia. El término exactitud se refiere al grado de acuerdo entre los

procedimientos a estudio y la prueba de referencia [12]. Puede ser expresada de muchos modos: sensibilidad-especificidad, cocientes de probabilidad, razón de odds (odds ratio) y área bajo la curva (curva ROC) [13-16].

La información derivada de la sensibilidad y especificidad de una prueba diagnóstica es de utilidad en los momentos iniciales del proceso diagnóstico, cuando se trata de decidir qué estudios complementarios serán más eficaces para confirmar o descartar una sospecha diagnóstica. Si se usa una prueba con el propósito de excluir una posibilidad diagnóstica o de hacer un escrutinio inicial, debe usarse una prueba sensible, ya que la sensibilidad se define como la capacidad del test para detectar la enfermedad. Por el contrario, si lo que se pretende es confirmar una sospecha diagnóstica, la prueba más específica (con mayor capacidad para detectar a los sanos) será la de mayor eficacia [4].

Estos estudios se deben distinguir de aquellos que están enfocados al estudio de las características analíticas de los ensayos, tales como la exactitud analítica, y de aquellos referidos a factores no analíticos, tales como la variación biológica interpersonal [12].

En la realización de estos estudios es importante que la investigación diagnóstica reúna una serie de requisitos para poder aplicar los resultados obtenidos a la práctica clínica. Entre ellos se encuentran las características relacionadas con la pregunta de investigación y con el método de realización del estudio de la prueba diagnóstica:

a) Relacionados con la pregunta de investigación:

Como he mencionado previamente, la Medicina Basada en la Evidencia ha contribuido al conocimiento de que la clave en una investigación sobre valoración de pruebas diagnósticas de laboratorio está en partir del problema a solucionar y no de preguntas acerca del valor de la prueba frente a una enfermedad concreta. Es decir, el investigador debería definir claramente qué pretende estudiar: no es lo mismo, por ejemplo, tratar de comprobar si un nuevo marcador tumoral puede tener algún valor diagnóstico hasta el momento desconocido, (estaríamos en una fase inicial de la investigación de una prueba diagnóstica, en la que se trata de comprobar si el marcador proporciona resultados positivos en pacientes con cáncer y resultados negativos en pacientes sin cáncer) que

definir su uso clínico para solventar problemas de incertidumbre diagnóstica en pacientes con determinados signos y síntomas (en este caso, se trataría de comprobar si los resultados del marcador son útiles para detectar o descartar la presencia de cáncer, cuando esta enfermedad se plantea en el diagnóstico diferencial). A partir de este supuesto es sencillo proponer un objetivo de investigación adecuado y además justificar la necesidad de realizar la investigación [10].

De la pregunta de investigación a responder depende el diseño a emplear y muy particularmente la población de estudio a elegir. Esta última, debería estar constituida por pacientes sospechosos de padecer la enfermedad, la cual quiere diagnosticar la prueba que estamos evaluando. Por ejemplo, si el objetivo de nuestro estudio es determinar el valor de los anticuerpos antinucleares en el diagnóstico del lupus eritematoso sistémico, no estamos identificando el problema diagnóstico y el estudio puede consistir por ejemplo, en la comparación de los resultados que arroja la prueba diagnóstica en un grupo de pacientes tratados por lupus eritematoso sistémico y un grupo de pacientes con otra patología, o incluso un grupo de donantes de sangre; en ambos supuestos la investigación tendría un interés muy limitado. Si el objetivo del estudio es la determinación del valor diagnóstico de los anticuerpos antinucleares en pacientes con sospecha de lupus eritematoso, estamos indicando que se debe aplicar la prueba a valorar a pacientes todavía no diagnosticados pero en los que por sus características puede sospecharse la patología de lupus, como es el caso de mujeres de edad media con fiebre, artritis y artralgias [10].

b) Relacionados con el método de realizar la evaluación de la prueba:

El mejor diseño de un estudio para determinar la exactitud de una prueba, consiste en una población compuesta de pacientes con un determinado problema diagnóstico a los que se les aplica la prueba problema, e independientemente de los resultados obtenidos se les practica la prueba considerada patrón de referencia. El patrón de referencia (en inglés, “gold standard”) es el mejor método para establecer la presencia o ausencia de enfermedad. Puede ser una sola prueba o una combinación de métodos y técnicas, incluyendo el seguimiento clínico de los sujetos [13].

Se debe añadir información sobre el contexto clínico y sobre los criterios empleados para la inclusión o exclusión definitiva de cada paciente en la población a estudio, para poder considerar la

aplicabilidad de los resultados a los propios pacientes. De esta forma, cuando se deseen hacer predicciones sobre otros pacientes a partir de los resultados obtenidos en los pacientes de un estudio, además de saber que acudieron al servicio sanitario en las mismas condiciones que los pacientes sobre los que desea hacer la predicción de los resultados, también sabremos si se presentaron a su médico de cabecera o al Servicio de Urgencias del hospital, si son pacientes atendidos en un hospital especializado, en uno comarcal... Todos ellos son datos que facilitarán nuestra tarea de predecir cómo funcionará la prueba con otro grupo de pacientes y cómo debemos interpretar sus resultados. La falta de estas características es uno de los problemas que con más frecuencia encontramos en los estudios de diagnóstico.

La población a estudio debe estar suficientemente descrita: características demográficas básicas, forma de presentación clínica, gravedad y comorbilidad de la enfermedad... [17]. La presentación de los pacientes en distintos estratos demográficos debe seguirse de la estimación de la sensibilidad y especificidad de la prueba en cada uno de los estratos que tengan relevancia clínica, para poder comprobar como la exactitud de la prueba varía de acuerdo a características importantes de los pacientes [10] [18-20].

Aunque el metaanálisis ya se está aplicando a la revisión de artículos de valoración de pruebas diagnósticas y se dispone de metodología y recomendaciones para su aplicación [21], su valor está limitado por la heterogeneidad de las poblaciones estudiadas y sobre todo por la carencia en la mayoría de las investigaciones primarias de una especificación de la situación clínica y del problema diagnóstico que se pretende resolver. Mientras no sea corriente que los estudios de diagnóstico se realicen en poblaciones clínicas reales, los resultados de los metaanálisis sobre investigación diagnóstica serán de escasa utilidad clínica.

Uno de los problemas más difíciles de reconocer y que a menudo no se incluye en las listas de criterios es el sesgo de secuencia o verificación diagnóstica (en inglés, “workup bias”). Ocurre cuando los resultados de la prueba a evaluar no son igualmente confirmados con la prueba estándar, sino que dependen del resultado de la prueba. Cuando esto ocurre en estudios diseñados para evaluar pruebas diagnósticas, la sensibilidad y especificidad de estas pruebas pueden ser estimadas incorrectamente. Si pacientes con resultados negativos de la prueba se asumen como sanos cuando no lo son, se producirá una estimación falsamente elevada de la sensibilidad y la especificidad. Este

sesgo también puede ocurrir si pacientes con resultados negativos a los que no se les ha sometido a la prueba estándar son excluidos de los cálculos de la prueba diagnóstica. En esta situación la sensibilidad puede verse falsamente elevada y la especificidad aparentemente disminuida. Este sesgo es común en estudios retrospectivos donde solo los pacientes testados y sometidos a la evaluación de la prueba estándar son normalmente incluidos y la prueba estándar está más a menudo aplicada a pacientes con resultados positivos. Los estudios retrospectivos pueden ser objeto de este sesgo, a no ser que en el estudio se indique que los pacientes que recibieron la prueba diagnóstica tienen, igual oportunidad de recibir la prueba estándar.

Otro sesgo frecuente y que hay que prevenir es el sesgo de revisión o comparación ciega (en inglés, “review bias”): si la prueba diagnóstica se realiza primero, el conocimiento de su resultado puede conducir a un excesivo grado de acuerdo con la interpretación del patrón de referencia.

Al planificar el estudio hay que tener en cuenta el tamaño muestral. Es decir, el número de sujetos necesarios incluir con el fin de obtener la precisión deseada en términos de intervalos de confianza. El número final de sujetos vendrá determinado por la frecuencia esperada de la enfermedad.

El valor de cualquier prueba depende de su capacidad de dar el mismo resultado cuando se aplica a los mismos pacientes en igualdad de condiciones. Una escasa reproducibilidad puede ser debida a problemas con la prueba en sí misma o a la interpretación del observador [22]. El estudio de la variabilidad interobservador en la evaluación de una prueba diagnóstica se descuida con excesiva frecuencia cuando se considera la investigación diagnóstica, aunque se ha visto el papel importante que desempeña en la mejora de la calidad del estudio.

Como he comentado, se han realizado algunas revisiones previas de la metodología diagnóstica empleada en los trabajos que evalúan la sensibilidad y especificidad diagnóstica de pruebas de laboratorio, mostrando una baja calidad en su sistemática. Esto ha llevado a diversos grupos al planteamiento de recomendaciones que incidan en una mejora de la calidad diagnóstica. En el siguiente apartado se van a describir las características más importantes de estos trabajos previos y las principales iniciativas desarrolladas para mejorar la investigación en este campo.

1.2- ANTECEDENTES DE TRABAJOS QUE EVALÚAN LA CALIDAD METODOLÓGICA UTILIZADA EN LOS ESTUDIOS DE SENSIBILIDAD Y ESPECIFICIDAD DIAGNÓSTICA DE PRUEBAS DE LABORATORIO, Y RECOMENDACIONES LLEVADAS A CABO EN ESTE CAMPO DEL DIAGNÓSTICO.

Los estándares metodológicos anteriormente mencionados, todavía no se han incorporado a la investigación sobre pruebas diagnósticas, lo que contribuye a la peor aplicabilidad de los resultados obtenidos a la práctica clínica. Si analizamos algunas de las revisiones realizadas en este campo podemos comprobar cómo la investigación sobres pruebas diagnósticas ha sido menos permeable a la mejora en el rigor científico, a pesar de los esfuerzos recientes antes indicados [23].

Una revisión de los trabajos sobre valoración de pruebas diagnósticas publicados en las mejores revistas de medicina general internacionales mostraba un indudable déficit metodológico [5]. En ella se analizaron los artículos que evaluaban pruebas diagnósticas, publicados en cuatro de las mejores revistas de ámbito internacional desde el año 1978 al 1993: *New England Journal of Medicine*, *Journal of the American Medical Association, British Medical Journal* y *Lancet* (un total de 112 estudios). Se aplicaron siete estándares metodológicos que debían contener los trabajos (Anexo I) y se comprobó que ningún estudio cumplía más de cinco de sus ítems y la mayoría solo seguía tres o cuatro de ellos. Aunque se veía una tendencia a la mejora en los últimos años.

Otro estudio demostró en el año 1999, cómo el fallo de la adherencia a los estándares metodológicos produce sobreestimaciones en la exactitud diagnóstica [6]. Esta evidencia indica la necesidad de mejorar las evaluaciones clínicas de los tests diagnósticos.

En España se han tratado adecuadamente los aspectos más cuantitativos de la investigación sobre pruebas diagnósticas, en particular la presentación de resultados y su interpretación. No ocurre igual en lo relativo a la calidad de la metodología empleada. Una revisión de los artículos sobre

pruebas diagnósticas publicada en la revista *Medicina Clínica* ponía en evidencia un grado similar de problemas de método [7]. Se estudiaron los artículos publicados durante los años 1992-1995 que versaban sobre pruebas diagnósticas, aplicándose una serie de criterios metodológicos (Anexo II). La conclusión principal obtenida fue que quedaban aspectos por mejorar en el diseño de estudios sobre tests diagnósticos, con resultados muy similares a los obtenidos en la revisión internacional antes mencionada [5]. A destacar la falta del estudio de la reproducibilidad de las pruebas estudiadas (solo un 19% de ellos la evaluaba), la escasa comunicación de la especificación de la fuente de la población a estudio, los criterios de inclusión y la descripción del espectro de sujetos (en el 31, 36 y 31% respectivamente del total de trabajos analizados) y el cálculo de la precisión estadística de las estimaciones en solo un 17% de los estudios.

También aparece publicada otra revisión en la revista *Enfermedades Infecciosas y Microbiología Clínica* sobre métodos de valoración de pruebas diagnósticas [8]. Se incluyeron en el estudio 45 artículos que estudiaban sensibilidad y especificidad de pruebas diagnósticas, publicados durante los años 1990-1996 y se les aplicó la misma guía que en la revisión anterior (Anexo II). Aunque la calidad de los trabajos era aceptable, los apartados referentes al diseño y presentación de resultados quedaban por perfeccionar: solo en el 11% de los trabajos se evaluaba la reproducibilidad de la prueba diagnóstica; la especificación de la fuente de la población de referencia, criterios de inclusión y espectro de sujetos se presentaron en el 58, 33 y 40% de los artículos respectivamente; solo un 6% especificaba el análisis independiente de los resultados y solo un 11% tuvieron en cuenta los posibles resultados indeterminados de las pruebas diagnósticas.

Esta falta patente de calidad en el campo del diagnóstico de laboratorio ha llevado a diversos grupos editoriales e investigadores a plantearse la elaboración de numerosas recomendaciones sobre investigación diagnóstica. Estas indicaciones se diseñan con el doble objetivo de mejorar el rigor metodológico de la investigación y la calidad de la comunicación de los resultados, y están dirigidas tanto a los investigadores para su aplicación en el diseño de sus estudios, como a los revisores de las revistas científicas.

Desde 1996 una revista internacional de laboratorio, *Clinical Chemistry*, incluye en las instrucciones para los autores los siete criterios antes mencionados, usados por Reid y colaboradores [5]. En 1997 y en la misma publicación, aparecen una serie de recomendaciones que deben seguir los estudios que versen sobre eficacia diagnóstica de las pruebas [24]. Esta guía se completa con otra revisión de la misma revista del año 2000 y que se basa en comentarios aportados entre otros, por estadísticos y editores [25].

Con el fin de ayudar a mejorar la calidad de los trabajos de los ensayos controlados aleatorios, se constituye el Grupo de Estándares de Ensayos Clínicos (Standards of Reporting Trials, SORT) y el Grupo de Trabajo de Recomendaciones para la Divulgación de Ensayos Clínicos en la Literatura Biomédica. Ambos grupos desarrollaron el CONSORT (Consolidated Standard of Reporting Trials). Esta iniciativa pretende aumentar la transparencia en la divulgación de los métodos y resultados para que los ensayos clínicos puedan ser interpretados fácilmente y de manera exacta [26-31].

En 1999, se reúne la asociación Cochrane en el Grupo de Trabajo de Métodos de Screening y Diagnóstico de Cochrane y discuten la baja calidad metodológica de los trabajos que evalúan los tests diagnósticos. Consideran que el primer paso en la corrección de estos problemas es la mejora en la divulgación de los artículos. Teniendo como referente la exitosa iniciativa CONSORT (Consolidated Standards of Reporting Trials), el Grupo de Trabajo se propuso el desarrollo de una lista de criterios que deberían ser incluidos en un artículo de exactitud diagnóstica. Esta iniciativa se denominó STARD, (Standards for Reporting of Diagnostic Accuracy) y su objetivo fue mejorar la calidad de los trabajos que tratan la exactitud diagnóstica.

Para el desarrollo de la iniciativa STARD se estableció un comité en Diciembre del 1999. En Septiembre del 2000 en Ámsterdam, un grupo de trabajo multidisciplinar preparó un documento de consenso y una plantilla de los criterios metodológicos publicados. En Noviembre del 2000, un boceto de la iniciativa STARD y una lista de 25 criterios se circuló entre el grupo. Finalmente en el año 2003, aparece publicada la lista de los 25 criterios de STARD que debe cumplir un artículo que evalúa

pruebas diagnósticas en las mejores revistas internacionales, tanto de laboratorio, como de ámbito médico (Anexo IV) [12, 13][32]. Estos criterios todavía no han sido evaluados en la práctica, y por lo tanto no tenemos información acerca de su posible aplicabilidad.

En el año 2002 aparece publicado el primer libro sobre diseño de investigación en diagnóstico clínico [33], lo que pone de manifiesto la importancia que está adquiriendo este tema y la preocupación por mantener una metodología diagnóstica de calidad.

Todas estas iniciativas pueden haberse reflejado en un cambio positivo de la calidad de la investigación diagnóstica.

Como menciono anteriormente, dentro de los estudios de exactitud diagnóstica un aspecto muy importante e imprescindible antes de la utilización de la prueba diagnóstica en la práctica clínica, es el cálculo de su reproducibilidad, siendo uno de los estándares metodológicos exigibles para aumentar la calidad de los procedimientos en el laboratorio de Análisis Clínicos. Dicha reproducibilidad puede ser de dos tipos: del instrumento con el que se realiza la determinación de laboratorio o el cálculo de la variabilidad entre los distintos observadores que realizan la prueba diagnóstica. Hemos centrado el segundo apartado de este trabajo en la realización de un estudio que determine la variabilidad entre varios observadores que llevan a cabo una prueba diagnóstica de laboratorio, mediante la aplicación de los criterios que recoge una guía previa de Feinstein [11].

La reproducibilidad del instrumento con el que se realiza la determinación analítica expresada como coeficiente de variación, es una característica que se estudia en la mayor parte de las evaluaciones diagnósticas y se ha incorporado a los programas de control de calidad ya existentes en los laboratorios. Sin embargo, aunque es necesario que se realicen estudios que evalúen la variabilidad entre los observadores que realizan una prueba diagnóstica antes de que ésta se introduzca en la práctica clínica, este tipo de análisis no se ha introducido a los programas de control de calidad de los Servicios de Análisis Clínicos.

Aunque se han realizado distintos esfuerzos para aumentar la calidad del diseño de los estudios de variabilidad observacional, en la práctica se ha prestado poca atención a este tipo de trabajos.

En el año 1985, Feinstein en su libro de Epidemiología Clínica [11], dedica un capítulo al estudio de la variabilidad observacional en la valoración de pruebas diagnósticas. En él detalla una guía de recomendaciones que se deben seguir para la realización de estudios de consistencia entre distintos observadores con suficiente calidad.

En los años 1985 y 1992 aparecen dos revisiones que recogen los estudios que evalúan la reproducibilidad entre distintos observadores [34, 35]. Se puede comprobar cómo apenas aparecen estudios en el área de las pruebas diagnósticas de laboratorio [36-42], aunque sí se han hecho numerosas valoraciones en áreas relacionadas con la Demografía y estudios de la población, el área del Radiodiagnóstico y de Anatomía Patológica, entre otras.

Debido a la importancia de este tipo de valoraciones cuando se realiza una determinación analítica y al escaso interés desarrollado en este campo, en este trabajo hemos querido realizar un estudio de variabilidad observacional siguiendo las recomendaciones establecidas por Feinstein [11]. De esta manera, se puede determinar la aplicabilidad y posibles problemas en la utilización de esta guía, e identificar con mayor certeza los retos que supone el realizar un estudio de variabilidad interobservador en una prueba diagnóstica de laboratorio.

1.3- CARACTERÍSTICAS PRINCIPALES DE LOS ESTUDIOS DE VARIABILIDAD ENTRE LOS DISTINTOS OBSERVADORES QUE ANALIZAN UNA PRUEBA DIAGNÓSTICA.

En la guía publicada por Feinstein de recomendaciones para la realización de estudios de variabilidad observacional [11], aparecen los aspectos más importantes para alcanzar un trabajo de calidad. Se basa en seis aspectos metodológicos clave que se detallan a continuación, como son la elección de un objetivo de calidad previo a la realización del estudio, la descripción del procedimiento del estudio, la elección de la muestra a analizar, la manera de expresar los resultados y el desacuerdo o acuerdo entre los observadores, y los índices estadísticos que se van a calcular para dichas expresiones de acuerdo.

1- Elección del objetivo de evaluación:

El objetivo de un estudio de reproducibilidad debe ser el análisis de los métodos instrumentales, la capacidad de los observadores o una combinación de ambos. En este caso, nosotros vamos a realizar un estudio de variabilidad observacional, por lo que hay que tener en cuenta que los análisis que se realizan en él dependen del tipo de objetivo que se ha elegido para la investigación:

A- Demostración de la consistencia de un nuevo índice: En un artículo original que trata del desarrollo de un nuevo índice, se incluye un estudio de variabilidad observacional para demostrar que da resultados consistentes. No obstante, muchas veces una nueva prueba se presenta sin la realización de estos estudios y el procedimiento puede llegar a aceptarse y ser ampliamente utilizado antes de que se investigue su variabilidad formalmente.

B- Demostración de la falta de fiabilidad: Ciertos estudios de variabilidad observacional pueden llevarse a cabo para demostrar que un determinado procedimiento es impreciso. Estudios de este tipo pueden ser necesarios para alertar a la comunidad científica de problemas que pueden haber sido pasados por alto.

C- Detalle del espectro de variabilidad: Aunque muchos estudios de variabilidad observacional parecen estar hechos por la única razón de demostrar y cuantificar su

existencia, otros pueden incluir un análisis más detallado del espectro de la variabilidad. En este caso, los investigadores pueden demostrar cómo la presente o ausente variabilidad está relacionada con el tipo de observaciones realizadas, por ejemplo, peculiaridades de los sujetos tales como la edad, el género o las características clínicas.

D- Reducir la variabilidad y mejorar la calidad: La característica más útil de un estudio de variabilidad observacional es el papel que juega en la reducción de la imprecisión y la mejora de la calidad del procedimiento en estudio. El investigador a menudo necesita que los observadores repitan el análisis que produce desacuerdo y entonces intentar identificar los puntos de error en la fase de preparación de las muestras o en la de interpretación de los resultados. Este tipo de estrategia puede servir por lo tanto, para mejorar la calidad de un proceso observacional sin ser necesario ningún estudio formal ni publicación de los resultados.

2- Procedimiento:

Uno de los problemas que puede encontrar el investigador que realiza este tipo de estudios es contar con los observadores necesarios y persuadirlos para que participen en la investigación: muchos de ellos pueden no querer participar debido a que no les guste la idea de someter sus conocimientos a verificación. Además, otra dificultad es el esfuerzo y el tiempo que deben invertir los observadores, por lo que el investigador debe ser el encargado de proveer las muestras en el formato necesario y de manera ciega e independiente.

Un estudio de variabilidad entre observadores se compone de dos pasos principalmente. El primero se refiere al proceso de obtención y preparación de las muestras para su observación. Este paso que debe ser explicado de manera detallada, en muchos estudios de variabilidad observacional se omite. El segundo comprende el proceso por el que se transforman los datos en categorías de interpretación.

3- Elección de la muestra:

Debe representar a la población clínica real. Si la muestra elegida tiene mayor proporción de casos “difíciles” que los que ocurren en la práctica habitual, los observadores pueden mostrar una alta tasa de desacuerdo. Este problema ocurre frecuentemente en los estudios de variabilidad entre observadores. Si se da el caso contrario, se obtienen unos resultados falsamente elevados del porcentaje de acuerdo que se calcula. El mejor método para evitar estos sesgos es analizar una serie consecutiva de pacientes candidatos a la evaluación diagnóstica, ampliada con pacientes adicionales elegidos de los casos extremos que se producen en la clínica. El número de estos casos inusuales se adecuará para que se obtenga una idea de la capacidad de evaluación de los observadores.

4- Escala de expresión de resultados:

El tipo de problema que ocurre con más frecuencia cuando se estudia la variabilidad observacional es la falta de estandarización de la escala de expresión de los resultados. Si dos observadores usan exactamente la misma escala para expresar sus resultados, el investigador puede haber evitado la primera causa de variabilidad. En este caso la fuente de imprecisión será fácilmente analizable ya que estará relacionada únicamente con el proceso observacional.

Una vez se ha elegido la escala, el investigador deberá discutir con los observadores acerca de su idoneidad y si es aceptada por todos los participantes en el estudio. Hay que tener en cuenta que un aumento en las categorías de expresión de los resultados permitirá a los observadores expresarse con más libertad, pero también aumentarán las posibilidades de variabilidad; un número pequeño de categorías puede reducir la variabilidad, pero si el número es demasiado pequeño la variabilidad puede aumentar debido a que los observadores no pueden expresar importantes distinciones.

5- Expresiones de acuerdo:

Además de la escala usada para expresar las observaciones, un primer objetivo de la investigación debe ser medir la magnitud del acuerdo. Dos observadores obviamente están en desacuerdo si usan diferentes categorías de expresión, pero el grado del mismo puede ser trivial, menor o mayor, dependiendo de la prueba que se esté realizando. Por ejemplo, si para una misma

muestra el laboratorio expresa un resultado de anticuerpos antinucleares positivos a dilución 1/10 y a dilución 1/20, la diferencia se puede ignorar; si los valores son positivos a dilución 1/10 y a 1/80, hay que preocuparse, y si los resultados son negativos y a dilución 1/160, evitaremos que el laboratorio realice esa determinación.

6- Índices estadísticos para expresión del acuerdo:

El acuerdo o desacuerdo entre varios observadores se debe expresar con índices estadísticos de concordancia y no con los índices habituales de correlación, como a menudo se han usado en los estudios de variabilidad observacional. Los índices apropiados son el porcentaje de acuerdo y el índice kappa.

Para una mejor comprensión de estas recomendaciones la guía incluye a su vez un resumen de los componentes de los que debe constar un estudio de variabilidad entre distintos observadores:

a- Objetivo: Debe aparecer claramente especificado, detallando si el estudio consiste en una demostración o una mejora de la variabilidad.

b- Muestra a estudio: Los sujetos o muestras que participan en el estudio deben ser representativos de la práctica clínica.

c- Partes del análisis: Especificación de si el objetivo de la investigación son los métodos instrumentales, la variabilidad observacional, o ambas. Si la investigación se centra en uno solo de los componentes debe haber una aclaración de que el otro está adecuadamente estandarizado.

d- Observaciones: Destacar si se realizan de manera independiente y "ciega".

e- Observadores: Describir si son suficientemente competentes y adecuados para llevar a cabo el estudio.

f- Escala de expresión de los resultados: Comprobar que la escala está expresada de manera satisfactoria y si se ha establecido ante de iniciar la investigación.

g- Escala de expresión del acuerdo: Comprobar que sea una escala adecuada y necesaria para expresar el grado de acuerdo entre varias observaciones.

h- Índice de concordancia: Deben indicar si los resultados se han expresado con índices adecuados. También debe aparecer el grado de acuerdo esperado por azar.

i- Procedimiento: Verificación de si se han establecido los criterios tanto de la primera fase de provisión y preparación de las muestras, como de la segunda fase de análisis de los resultados.

j- Interpretación: Descripción del procedimiento de transformación de los resultados en la escala de expresión.

k- Análisis: Identificación de las fuentes de variabilidad que pueden explicar la falta de precisión de la técnica.

l- Mejoras: Los observadores ponen en común sus desavenencias para intentar mejorar sus causas de variabilidad.

m- Recomendaciones: Presencia de sugerencias sobre cómo mejorar los defectos que se han detectado.

En resumen, en el contexto de esta investigación sobre la calidad de la investigación de pruebas de laboratorio, consideramos prioritario describir los retos del procedimiento que suponen los estudios de consistencia, en particular los estudios de variabilidad observacional en una prueba diagnóstica de laboratorio.

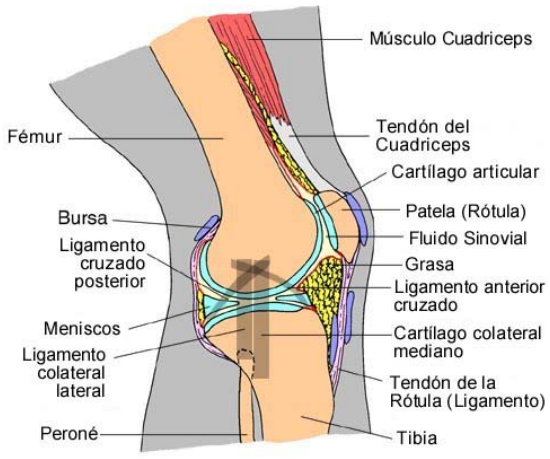
Para poner en práctica las recomendaciones de Feinstein, diseñamos y ejecutamos un estudio a cerca de la consistencia de una prueba concreta de laboratorio: el análisis de cristales de urato monosódico y de pirofosfato cálcico en muestras de líquido sinovial mediante el empleo sucesivo del microscopio óptico y polarizado. Se escogió esta evaluación porque aún tratándose de un método imprescindible para el diagnóstico de artropatías relacionadas con cristales, como son la gota y la pseudogota, se ha demostrado cómo su utilidad diagnóstica puede verse amenazada por la interpretación subjetiva del observador [43-47].

Se describe a continuación la anatomía de la articulación y las características de las dos patologías relacionadas con la deposición de cristales que se van a estudiar: la gota y la artropatía relacionada con cristales de pirofosfato cálcico. También es importante referir los estudios previamente realizados, para poder analizar sus posibles deficiencias que podrían explicar la inconsistencia de sus resultados.

1.4- ESTUDIO DE LA VARIABILIDAD INTEROBSERVACIONAL EN EL ANÁLISIS DE CRISTALES DE URATO MONOSÓDICO Y PIROFOSFATO CÁLCICO EN MUESTRAS DE LÍQUIDO SINOVIAL.

La articulación es una cavidad formada por el cartílago articular que actúa como amortiguador de los impactos producidos por los movimientos y torsiones, y por la membrana sinovial que produce el líquido sinovial, auténtico lubricante de la articulación. (Figura 1)

Figura 1: Anatomía de la articulación de la rodilla.



El líquido sinovial se forma a partir del ultrafiltrado del plasma por la membrana sinovial y excluye las proteínas de alto peso molecular. El volumen de líquido sinovial contenido en la cavidad articular es suficiente para cubrir sus superficies y en condiciones normales es como máximo de 3,5 ml (en condiciones de inflamación puede llegar a los 25 ml). Sus funciones son: reducir la fricción entre los huesos, lubricar las articulaciones y aportar nutrientes al cartílago.

Los elementos de sujeción de la articulación lo forman: la cápsula, envoltura muy resistente que rodea y sujeta toda la articulación; los ligamentos, que dan estabilidad a la articulación, y los músculos y tendones que además de contribuir a la estabilidad articular, permiten el movimiento.

Cuando se produce algún traumatismo o proceso inflamatorio o infeccioso, se acumula líquido sinovial en la cavidad articular, aumentando de tamaño muchas veces con dolor y dificultad para realizar los movimientos normales. Es de gran trascendencia la determinación de la causa que ha producido este exceso de líquido sinovial a través de la historia médica, el examen físico, la analítica sanguínea, análisis de orina de 24 horas, radiografías y el análisis del líquido sinovial del paciente para evitar un diagnóstico erróneo. El análisis del líquido sinovial es la prueba esencial en el establecimiento de un diagnóstico seguro, cobrando una mayor relevancia en el diagnóstico de artritis séptica y sobre todo en las artritis inducidas por cristales, donde es patognomónico.

Las artropatías producidas por cristales son un grupo heterogéneo de enfermedades bioquímicas que tienen en común la deposición de cristales en la articulación. Los mecanismos por los que ocurre la acumulación articular y la inflamación son similares en todas ellas. La formación de cristales puede ser debida a desórdenes metabólicos, disminución de la excreción renal, degeneración de hueso y cartílago o inyección terapéutica (por ejemplo de corticoesteroides).

Las dos patologías principales producidas por acumulación de cristales son gota (presencia de cristales de urato monosódico en líquido sinovial) y pseudogota o condrocalcinosis (presencia de cristales de pirofosfato cálcico dihidratado en líquido sinovial).

A- GOTA:

La gota es una enfermedad que se manifiesta clínicamente por episodios inflamatorios, habitualmente monoarticulares y recurrentes que suelen ser intensos y autolimitados; en fases más avanzadas de la enfermedad la inflamación articular puede ser crónica.

Está directamente relacionada con la presencia de cristales de urato monosódico en el interior de las articulaciones, tanto durante los episodios inflamatorios como fuera de ellos. Los cristales de urato monosódico se encuentran en rodillas de pacientes asintomáticos con gota y una vez depositados, permanecen ahí hasta que disminuyen los niveles de ácido úrico por acción del tratamiento. Por lo tanto la identificación de cristales en líquido sinovial puede usarse para el diagnóstico de gota en los períodos intercríticos [48, 49]. El diagnóstico de gota se establece mediante la identificación de cristales de urato monosódico en líquido sinovial o formando parte de tofos [50].

El límite de solubilidad del ácido úrico en sangre es de 7 mg/dl y a partir de ese valor precipita. Aunque la gota refleja la existencia de hiperuricemia, solamente un pequeño porcentaje de las personas que la presentan llega a padecer gota. En el resto de las ocasiones la hiperuricemia se considera asintomática y no precisa tratamiento a no ser que se asocie a litiasis renal por cálculos de ácido úrico. Los factores que determinan que esta inflamación dé lugar a los ataques de gota y los elementos de los que depende su cese espontáneo son de momento poco conocidos.

En el diagnóstico de esta patología hay que destacar las pruebas básicas de laboratorio que incluyen análisis de orina y análisis sanguíneo; recuento celular, determinación de creatinina, urea, nitrógeno y ácido úrico; radiografías, que no son muy útiles en el diagnóstico de los ataques iniciales de la artritis gotosa aguda ya que los hallazgos son generalmente inespecíficos y el análisis del líquido sinovial, que incluye la identificación de cristales y el recuento celular. La identificación de estos cristales como he comentado previamente, es patognomónico de gota y permite un diagnóstico exacto de la enfermedad [51, 52].

Ocasionalmente, en los pacientes con gota no se detectan cristales de urato monosódico en muestras de líquido sinovial, sin embargo la aspiración repetida veinticuatro horas después muestra cristales en las muestras de la mayoría de estos pacientes [53].

El recuento leucocitario es normalmente proporcional a la concentración de cristales presentes en la muestra de líquido sinovial. Durante el ataque agudo alcanza un valor entre 2.500 y 25.000 células/mm3, de los cuales más del 90% son neutrófilos.

B- ARTROPATIA POR CRISTALES DE PIROFOSFATO CÁLCICO DIHIDRATADO:

Es una enfermedad caracterizada por el depósito de cristales de pirofosfato cálcico dihidratado en la cavidad articular. Con el tiempo el cartílago llega a calcificarse, causando artrosis y en ocasiones ataques de artritis [54].

El depósito de cristales de pirofosfato cálcico dihidratado, casi restringido a tejidos articulares y pararticulares, representa la primera fase identificable de esta enfermedad y está relacionada con la sobreproducción de pirofosfato inorgánico por los condrocitos o células del cartílago [55]. En el 90% de los enfermos no se encuentra una causa para la enfermedad, aunque se puede relacionar con enfermedades metabólicas o endocrinas y con trastornos hereditarios.

La mayoría de las veces este trastorno no provoca ningún tipo de síntoma, tratándose entonces de un hallazgo radiológico casual (la calcificación forma una línea tenue paralela al hueso, que puede verse mediante radiografías). Sin embargo, en algunas personas sí produce molestias o dolor persistente, e incluso puede desencadenar un ataque agudo parecido a los producidos por la gota, por ello también se han denominado ataques de pseudogota [56].

Se dispone de dos herramientas fundamentales para asegurar el diagnóstico: la radiografía y el análisis de líquido sinovial. La radiografía puede conducir a error en calcificaciones que todavía no se han formado o son difíciles de apreciar, por lo que el diagnóstico se basa en la detección de cristales de pirofosfato cálcico dihidratado en una muestra de líquido sinovial [57]. La media del recuento leucocitario en líquido sinovial es normalmente más baja que en el caso de la gota, pero el diferencial es también de predominio neutrófilo.

El análisis del líquido sinovial, objeto de nuestro estudio, se realiza a partir de muestras extraídas de las articulaciones accesibles como son la rodilla y el codo, mientras que otras como la cadera no son viables a este procedimiento. Además de la identificación de cristales, en una muestra de líquido sinovial se pueden realizar otras determinaciones como son:

1- Análisis básico: color y viscosidad:

El líquido sinovial normal es de color amarillo claro, viscoso, (los sinoviocitos, células de la membrana sinovial, secretan mucopolisacárido, que contiene entre 0,2 y 0,5% de ácido hialurónico polimerizado formando cadenas de alto peso molecular y proteínas) pero no coagula.

2- Recuento y diferencial leucocitario:

Los leucocitos son el hallazgo celular más frecuente en el análisis de líquido sinovial de un sujeto no sano y da información acerca de la inflamación presente en las articulaciones (con la excepción de los casos de trauma donde el recuento de hematíes es muy alto).

El nivel aceptado de leucocitos que diferencia entre inflamatorio y no inflamatorio es de 2.000 x 10^6 células/mm^3. La clasificación del líquido sinovial en categorías de no inflamatorio, inflamatorio y séptico también se basa en la transparencia de la muestra y el porcentaje de neutrófilos. (Tabla 1)

Tabla 1: Clasificación de los distintos tipos de líquido sinovial en función de su transparencia, recuento de leucocitos y porcentaje de neutrófilos.

CLASIFICACIÓN	ASPECTO	LEUCOCITOS / mm3	% NEUTRÓFILOS
NORMAL	Transparente	<200	<25
NO INFLAMATORIO	Transparente	<2000	<25
INFLAMATORIO	Translúcido	<75000	>50
SÉPTICO	Opaco	>75000	>75

3- Determinaciones bioquímicas:

Algunos análisis bioquímicos adicionales realizados en muestras de líquido sinovial tales como la determinación de proteínas (valores normales <3g/dl: un aumento se puede deber a cambios en la permeabilidad de la membrana sinovial, aumento en la síntesis de articulación y a un proceso inflamatorio), factor reumatoide, niveles de complemento y glucosa (el nivel en líquido sinovial es similar al encontrado en sangre y se evalúan las diferencias entre ambos valores: normal < 10 mg/dl, proceso inflamatorio > 25mg/dl y sepsis >40 mg/dl), tienen escasa utilidad clínica [58].

Como he comentado previamente, el diagnóstico definitivo de las artropatías por cristales, gota y artropatía por pirofosfato cálcico, requiere aspiración y posterior examen del líquido sinovial para la detección y posterior identificación con microscopio de luz ordinaria y polarizada de los cristales de urato monosódico y de pirofosfato cálcico [59, 60].

Del resultado de este análisis depende el diagnóstico y por lo tanto el tratamiento de la enfermedad, que en el caso de la gota puede ser farmacológico de por vida. La identificación de cristales no debería excluir otras causas de inflamación articular tales como la artritis séptica, ya que estas enfermedades pueden coexistir.

Se han publicado varios casos de pacientes con gota y artropatía por cristales de pirofosfato cálcico en la misma articulación, por lo que cuando se examina una muestra de líquido sinovial y se detecta la presencia de un tipo de cristal se debe investigar también la de otro tipo de cristales [61].

Aunque solo estos dos tipos de cristales estudiados tienen relevancia diagnóstica, en una muestra de líquido sinovial pueden aparecer otros cristales [62]:

- *Cristales de hidroxiapatita (fosfato cálcico):* Se encuentran en líquido sinovial de pacientes con osteoartritis asociadas con deposiciones de calcio. Para su detección es necesario el microscopio electrónico, ya que al no ser birrefringentes no se

observan con el microscopio de luz polarizada. Son intracelulares, pequeños y de forma acicular.

- *Cristales de corticoides*: Se pueden encontrar en pacientes que han recibido recientemente una inyección terapéutica intraarticular. Pueden confundirse con los de urato monosódico ya que son también de forma acicular e intracelulares, pero exhiben birrefringencia positiva y negativa.
- *Cristales de oxalato cálcico* en pacientes con insuficiencia renal después de diálisis.
- *Cristales de colesterol*, aunque no tienen ninguna significación clínica, se asocian con inflamación crónica. Muestran dirección de birrefringencia negativa, bajo la luz polarizada son fuertemente birrefringentes y normalmente son extracelulares y de forma rómbica.

En la tabla 2, se muestran las características más relevantes de los cristales que se pueden hallar en una muestra de líquido sinovial.

Tabla 2: Características de los posibles cristales presentes en muestras de líquido sinovial:

CRISTAL	TAMAÑO (μm)	MORFOLOGÍA	BIRREFRINGENCIA	PATOLOGÍA
Urato monosódico	2-20	Barra, aguja	Negativa, intensa	Gota aguda y crónica.
Pirofosfato cálcico	2-10	Barra, romboide	Positiva, débil	Enfermedad por deposición de CPPD aguda o crónica.
Hidroxiapatita	5-20	Redondo o irregular	No	Artritis aguda o periartritis.
Corticoesteroides	4-15	Barra, irregular	Positiva o negativa intensa.	Postinyección

En una muestra de líquido sinovial podemos encontrar también distinto material birrefringente que es necesario distinguir de los cristales: cristales de anticoagulante como oxalato cálcico o heparina de litio, gránulos de almidón, fragmentos de prótesis, fibras de colágeno y partículas de polvo.

Los cristales de urato monosódico y pirofosfato cálcico muestran características diferentes de birrefringencia y el microscopio de luz polarizada constituye el método estándar para el análisis de líquido sinovial en la búsqueda de cristales [62-65]: aunque los cristales de urato monosódico son fuertemente birrefringente y fácilmente identificables con el microscopio de luz polarizada, los cristales de pirofosfato cálcico son escasamente birrefringentes y muchos no lo son [66].

Diversos estudios han demostrado cómo la mencionada utilidad diagnóstica puede verse amenazada por la falta de consistencia entre los observadores que analizan muestras de líquido sinovial. Estos resultados ponen de manifiesto la falta de consenso sobre la rutina del análisis de líquido sinovial.

En un primer estudio, se remiten de manera ciega once alícuotas de muestras de líquido sinovial distintas a 3 laboratorios que participan en un estudio de variabilidad observacional. Aunque los análisis se realizan de manera independiente y sin conocimiento del diagnóstico definitivo, no se especifica la experiencia de los observadores en el análisis de cristales ni se protocoliza el mismo procedimiento para la realización de las determinaciones. Los resultados se expresan como ausencia de cristales, cristales de urato monosódico o cristales de pirofosfato cálcico. Cuando se compara los resultados obtenidos con el laboratorio de referencia se observan discrepancias en 7 de las 11 muestras analizadas, pero no se calcula ningún índice que exprese estos resultados. Se obtiene una sensibilidad del 62.5% para el análisis de cristales de urato monosódico [43].

En otro trabajo acerca del estudio de variabilidad de los resultados obtenidos al analizar muestras de líquido sinovial, se envían alícuotas de cuatro líquidos sinoviales distintos a 25 laboratorios hospitalarios que no conocían el contenido de las muestras. Todos los hospitales participantes en el estudio realizaban habitualmente este tipo de análisis, pero no se indica en el trabajo el grado de experiencia de sus analistas. Se pregunta por presencia o ausencia de cristales, tipo de cristales, recuento celular y diferencial. En los resultados se comprueba como una muestra fue identificada correctamente por 4 laboratorios; otra muestra por 20 centros; la tercera muestra por 1 laboratorio y la última muestra por 24 hospitales, obteniendo una sensibilidad del 78% para urato monosódico y del 12% para pirofosfato cálcico dihidratado. Aunque las determinaciones se llevaron a

cabo de manera ciega e independiente, no se unificaron los procedimientos de análisis para todos los centros participantes. Para la determinación de cristales de urato monosódico se obtiene una sensibilidad del 78% y un 24% de resultados falsos positivos y para los cristales de pirofosfato cálcico una sensibilidad de 12% [44].

En un nuevo trabajo, se preparan alícuotas con cristales de urato monosódico y otros posibles materiales de confusión, como son cristales de colesterol o partículas de almidón, para así estudiar la precisión obtenida entre 25 laboratorios. La determinación se realiza de manera ciega e independiente. En este caso tampoco se conoce el grado de formación de los observadores, ni se calculan los índices de concordancia entre ellos. Se obtienen un 24% de falsos positivos [45].

Queriendo analizar una vez más la reproducibilidad entre varios observadores que analizan muestras de líquido sinovial, se envían a seis observadores 99 portaobjetos que contienen muestras con distintas concentraciones de cristales (41 con cristales de urato monosódico y 42 con cristales de pirofosfato cálcico) y 16 muestras sin cristales. Aunque la correcta identificación de los cristales aumenta con su concentración, la sensibilidad para cristales de urato monosódico es del 69% (especificidad 97%) y la sensibilidad para cristales de pirofosfato cálcico dihidratado del 82% (especificidad 78%). En este trabajo, se describe correctamente la primera fase del estudio de provisión y preparación de muestras y se define previamente la escala de expresión de resultados. Aunque se determinan la sensibilidad y especificidad para el análisis de los cristales (para urato monosódico 69 y 97% respectivamente y para pirofosfato cálcico 82 y 78%), no se calcula ningún índice que exprese el grado de concordancia obtenido por los evaluadores [46].

En un último estudio, se distribuyen cuatro preparaciones de líquido sinovial entre 25 y 47 laboratorios clínicos entre 1989 y 1996. Cuando los cristales de urato monosódico son abundantes, la tasa de falsos positivos es entre 0 y 38% y aumenta a 67% cuando estos cristales son escasos. En este caso, se describe la manera de proceder para la preparación de muestras y el modo en el que los observadores deben analizar las muestras, pero no se describe la escala de expresión de los

resultados, ni se calcula ningún índice de concordancia en la observación. Tampoco conocemos el grado de formación de los participantes [47].

Esta falta de consistencia en el análisis del líquido sinovial se puede explicar por a) la incapacidad de la técnica para dar resultados de calidad o b) puede ser atribuida al observador. Debido a que el uso del microscopio para el análisis de cristales en muestras de líquido sinovial es una herramienta estandarizada internacionalmente, nos planteamos si se pueden obtener unos resultados más adecuados cuando realizamos un estudio de variabilidad observacional en las condiciones metodológicamente adecuadas, antes mencionadas.

2- JUSTIFICACIÓN

Esta tesis doctoral se centra en el estudio de la investigación de la metodología diagnóstica en los Servicios de Análisis Clínicos y se aborda desde una doble perspectiva.

En la primera parte, nos hemos querido centrar en la valoración de los estudios publicados, que calculan la exactitud diagnóstica de pruebas de laboratorio. Diversas revisiones previas han mostrado un indudable déficit metodológico en este tipo de estudios, sobre todo en determinados parámetros como puede ser el cálculo de la reproducibilidad de la técnica o la descripción del espectro de los sujetos que participan en los estudios. Esto conlleva a que los resultados obtenidos en estos estudios no sean de suficiente calidad, y que el clínico no pueda juzgar la validez interna y la aplicabilidad de la prueba diagnóstica en estudio a otros contextos (validez externa). Tras la publicación recientemente de diversas iniciativas que pretenden mejorar estos aspectos más deficientes, creemos necesario valorar cómo han influido en la investigación diagnóstica y por lo tanto, conocer el estado actual de la misma. De esta manera, si comparamos la calidad de los artículos actuales con aquellas revisiones que se realizaron previamente a la aparición de las recomendaciones, podremos evaluar si se ha producido un cambio en la calidad metodológica de las pruebas de laboratorio, y por tanto, si han sido útiles dichas recomendaciones.

Este trabajo puede permitir además, constatar la aplicabilidad de las guías existentes a la evaluación de la calidad de los estudios de exactitud diagnóstica de pruebas de laboratorio. El único modo de poder comprobar si realmente estos criterios son suficientes para la valoración metodológica, es mediante su aplicación a artículos concretos, como hemos llevado a cabo en este trabajo. Por otro lado, actualmente está vigente la guía de recomendaciones STARD (Standards for Reporting of Diagnostic Accuracy) [32], la cual se está difundiendo ampliamente. Muchos de los criterios que nosotros vamos a emplear coinciden con los contenidos en STARD, por lo que podremos añadir aclaraciones a la utilización de esta guía.

Respecto a la segunda parte del trabajo, uno de los puntos de que adolecen los principales trabajos de exactitud diagnóstica, es el estudio de la variabilidad entre los observadores que verifican una prueba diagnóstica. Actualmente hay pocos estudios de calidad realizados en este campo, y no es un aspecto que se incluya habitualmente en los controles de calidad de los laboratorios clínicos.

Hay que tener en cuenta que solo se ha desarrollado la guía de Feinstein [11] de recomendaciones para este tipo de estudios, y que todavía no se ha aplicado, por lo que no podemos valorar su utilidad. Por lo tanto, hemos creído necesario realizar un estudio de variabilidad observacional de una prueba de laboratorio concreta, mediante la aplicación de estas recomendaciones para, por un lado aumentar la calidad de dicha prueba, y por otro comprobar la aplicabilidad de los criterios de la guía.

Elegimos por la presencia de resultados contradictorios previos, la determinación de cristales de urato monosódico y de pirofosfato cálcico en muestras de líquido sinovial. Se trata de una determinación patognomónica de gota y de artropatía por cristales de pirofosfato, cuyo valor se ha visto reducido en anteriores estudios por una gran variabilidad observacional. Pensamos que si realizamos un estudio de elevada calidad metodológica, mediante el seguimiento de los principios de Feinstein, lograremos aumentar la concordancia de los observadores, y por lo tanto, la validez de la prueba. Además con este procedimiento vamos a poder determinar la aplicabilidad y utilidad de los criterios de Feinstein, para futuras aplicaciones en otros estudios de concordancia observacional.

Por último, cabe destacar que la elección del Servicio de Análisis Clínicos como ámbito de estudio se ha debido a las diferentes características de estos servicios con otros de tradición clínica, y a que concentran gran parte del trabajo diagnóstico que se efectúa en los hospitales. Como he comentado previamente, el desarrollo de una nueva prueba diagnóstica de laboratorio se suele realizar en los Servicios de Análisis Clínicos hospitalarios, los cuales se encuentran alejados del resto de los departamentos clínicos. Esto podría conllevar a una menor calidad metodológica de las evaluaciones que realizan los nuevos tests diagnósticos, ya que no se tiene acceso a las características demográficas y clínicas clave de los pacientes que participan. Pero por otro lado, estos servicios tienen la ventaja de una buena puesta en práctica de la prueba diagnóstica, con un mayor cuidado de los aspectos analíticos y técnicos. Con la realización de esta valoración, podremos evaluar las principales deficiencias en investigación sobre diagnóstico, específicamente en pruebas de laboratorio, y establecer recomendaciones para subsanarlas.

3- HIPÓTESIS

Respecto al estudio de la calidad metodológica de los artículos que determinan la exactitud diagnóstica de pruebas de laboratorio, cabe suponer que la reciente publicación de guías sobre la metodología que debe seguir este tipo de estudios [5, 7, 8, 24, 25], junto con la aparición del primer libro sobre el tema [33] y la difusión del metaanálisis [21], deben haber influido en la calidad de las investigaciones publicadas.

Por otra parte, es verosímil que la investigación sobre diagnóstico publicada en revistas del área de laboratorio de Análisis Clínicos, a pesar de la escasez de antecedentes que orientan sobre la cuestión, tenga unas carencias específicas por la falta de integración entre los equipos clínicos y los de laboratorio.

Por último, la escasa experiencia en la aplicación de algunos criterios, parte de los cuales están incluidos en STARD [32], así como la ausencia de información sobre su aplicabilidad general, indica que es posible que haya criterios metodológicos de difícil interpretación que exijan una nueva consideración.

Por todo esto nos plantemos las siguientes hipótesis:

Los artículos que evalúan pruebas diagnósticas de laboratorio aparecidos tanto en revistas nacionales como internacionales:

1- Muestran una apreciable mejora de la calidad metodológica, particularmente después de la publicación de las recomendaciones sistemáticas.

2- Tienen carencias específicas en comparación a las áreas diagnósticas más cercanas al ámbito clínico, como pueden ser las relacionadas con las características clínicas y demográficas de los pacientes que se incluyen en una investigación de pruebas diagnósticas de laboratorio.

Respecto al estudio de la consistencia en la detección e identificación de cristales de urato monosódico y pirofosfato cálcico dihidratado en muestras de líquido sinovial, la falta de precisión reflejada en estudios previos puede derivar de la falta de realización de un estudio de concordancia de suficiente calidad como para dar resultados consistentes. Sobre todo, no se hace referencia en los trabajos anteriores al grado de formación y experiencia de los participantes en el estudio, siendo éste un aspecto clave para la realización de un estudio de variabilidad diagnóstica según las recomendaciones de Feinstein [11].

Pensamos que si los analistas reciben una adecuada formación, se puede alcanzar una reproducibilidad elevada en la detección e identificación de cristales en muestras de líquido sinovial con el empleo sucesivo del microscopio óptico y el de luz polarizada. Con todo esto se plantea la siguiente hipótesis de trabajo:

Si los analistas que van a evaluar muestras de líquido sinovial para la detección e identificación de cristales reciben una formación previa adecuada, se pueden obtener unos valores aceptables de concordancia y validez para el análisis de cristales de urato monosódico y pirofosfato cálcico dihidratado en muestras de líquido sinovial.

Hay otra cuestión que no precisa de la formulación de una hipótesis previa y que esperamos obtener al finalizar el presente trabajo, como es la confirmación de las normas necesarias para la realización de un correcto estudio de variabilidad entre observadores, señalando los aspectos de la guía metodológica de Feinstein[11] con más dificultad en su aplicación.

4- OBJETIVOS

4.1- OBJETIVOS GENERALES:

1- Evaluar la calidad metodológica de los artículos que estudian pruebas diagnósticas de laboratorio clínico publicados en revistas científicas, su evolución temporal, sus características específicas y su variación de acuerdo a las características de los autores y sus centros de procedencia.

2- Valorar la consistencia entre observadores en el análisis de cristales en muestras de líquido sinovial identificando los problemas metodológicos, los de factibilidad de este tipo de investigación y la pertinencia de las recomendaciones disponibles sobre su diseño y ejecución.

4.2- OBJETIVOS ESPECÍFICOS:

1.1- Evaluar aquellos aspectos de la investigación sobre pruebas de laboratorio más deficientes y que por tanto más comprometen su validez y aplicabilidad.

1.2- Evaluar si en la investigación sobre diagnóstico publicada en revistas nacionales e internacionales se puede apreciar una tendencia temporal a la mejora en los aspectos metodológicos clave.

1.3- Analizar la concordancia de los investigadores en la evaluación crítica de los artículos y descripción de sus principales discrepancias.

1.4- Estudiar la influencia de determinadas variables como son el número de centros participantes en el estudio, el país al que pertenecen dicho centros, el carácter clínico o de laboratorio de la revista en la que aparece publicado el artículo, y el ámbito en el que se realiza la prueba diagnóstica sobre la calidad del artículo.

1.5- Valorar cualitativamente la aplicabilidad de los criterios utilizados para analizar la calidad de los artículos.

2.1- Determinar la consistencia en el análisis de cristales de urato monosódico y pirofosfato cálcico en muestras de líquido sinovial, mediante el examen al microscopio, para el diagnóstico de artropatías relacionadas con cristales.

2.2- Estudiar las posibles causas de las inconsistencias que aumentan la imprecisión en el estudio de cristales en muestras de líquido sinovial.

2.3- Valorar las recomendaciones a seguir cuando se realiza un estudio de variabilidad entre observadores que analizan una prueba diagnóstica de laboratorio.

2.4- Contribuir al desarrollo de guías validadas en el análisis de cristales de urato monosódico y pirofosfato cálcico en muestras de líquido sinovial.

5- MATERIAL Y MÉTODOS

Como he ido comentando a lo largo del trabajo, este estudio de investigación diagnóstica sobre pruebas de laboratorio en los Servicios de Análisis Clínicos se compone de tres partes: las dos primeras comprenden la valoración de los trabajos que evalúan pruebas diagnósticas de laboratorio y la tercera constituye un estudio de variabilidad entre observadores para la determinación de cristales de urato monosódico y de pirofosfato cálcico en muestras de líquido sinovial.

Las dos primeras partes del estudio contienen un análisis de los artículos aparecidos en dos revistas científicas clínicas nacionales, *Medicina Clínica (Barc)* y *Revista Clínica Española,* (trabajo que se ha publicado en *Revista Clínica Española*, anexo V) y una valoración de los estudios aparecidos en dos revistas internacionales, pero de ámbito exclusivo de laboratorio, *Clinical Chemistry* y *Clinical Chemistry and Laboratory Medicine* (de este otro estudio, publicado en *Clinical Chemistry* anexo VII, se publica una referencia en el editorial del mismo número, anexo VII). Por tener ambos estudios la misma metodología, esta se comenta de manera conjunta.

La tercera sección del presente trabajo, el estudio de variabilidad de una prueba diagnóstica de laboratorio, por tener una metodología diferente se comenta de manera separada (esta última parte del trabajo, está en prensa y próximo a su publicación en *Annals of the Rheumatic Diseases*, anexo VIII).

5.1- REVISIÓN CRÍTICA DE LOS TRABAJOS QUE ESTUDIAN PRUEBAS DIAGNÓSTICAS DE LABORATORIO, Y PUBLICADOS EN REVISTAS CIENTÍFICAS NACIONALES E INTERNACIONALES

En esta sección del estudio correspondiente a la valoración de la calidad metodológica de artículos que estudian pruebas de laboratorio, se revisan los trabajos publicados en dos revistas clínicas de ámbito nacional desde 1997 al año 2000 (*Medicina Clínica (Barc)* y *Revista Clínica Española*) [68-84] y en dos revistas de laboratorio internacionales (*Clinical Chemistry* y *Clinical Chemistry and Laboratory Medicine*) durante los años 1996, 2001 y 2002 [85-163]. La elección de las revistas estuvo guiada por su calidad (las cuatro publicaciones aparecen en el *Index Citation Report*) y por la frecuencia con la que publicaban valoraciones diagnósticas.

La selección de estas dos revistas nacionales, se realizó con el fin de comprobar si se había producido una mejora en España en la calidad metodológica de la investigación diagnóstica, respecto a las dos revisiones nacionales previas realizadas entre los años 1990 y 1996 [7, 8], y a la internacional publicada en el año 1995 y si sus recomendaciones habían influido en la mejora de nuestros artículos de exactitud diagnóstica. Elegimos el período de tiempo comprendido entre 1997 y 2000, por pensar que había dado tiempo suficiente para la incorporación de las recomendaciones en la investigación de pruebas de laboratorio en España. Seleccionamos publicaciones de medicina general, al igual que en la revisión internacional [5] y en una de las nacionales previa [7]. Las dos revisiones nacionales anteriores [7, 8], además de los criterios de Reid y colaboradores [5], también han utilizado los criterios propios de los autores que aplicamos en este trabajo, lo que facilita la comparación a través del tiempo.

La aparición en 1996 de los siete criterios de Reid y colaboradores [5] en las instrucciones para los autores de una revista internacional exclusiva de laboratorio como es *Clinical Chemistry*, y la publicación de una serie de recomendaciones en 1997 para los estudios de eficacia diagnóstica en la misma revista [24], que se completa en el año 2000 [25], nos permitirá comprobar si el esfuerzo realizado en el campo del laboratorio, se ve reflejado en una mejora de la calidad metodológica. Tomando como referencia el año anterior a la aparición de las recomendaciones, año 1996, estudiamos los cambios producidos en los trabajos publicados durante los años 2001 y 2002. En un principio, estudiamos solo los artículos publicados durante los años 1996 y 2001, pero por consejo del

editor de la revista *Clinical Chemistry,* ampliamos el estudio al año 2002, otorgando así un mayor plazo a la incorporación de los criterios metodológicos a la investigación diagnóstica de laboratorio.

5.1.1- ELECCIÓN DE LOS ARTÍCULOS SUSCEPTIBLES DE EVALUACIÓN.

Para la selección de los artículos a analizar, se realiza una búsqueda mediante el sistema informático MEDLINE.

Siguiendo las recomendaciones de otros autores [5-8], escogemos en primer lugar las bases de datos correspondientes a los años que se quieren analizar para cada revista en estudio: para *Revista Clínica Española* y *Medicina Clínica (Barc)* los años 1997, 1998, 1999 y 2000, y para *Clinical Chemistry* y *Clinical Chemistry and Laboratory Medicine*, los años 1996, 2001 y 2002. Mediante una búsqueda utilizando las palabras clave *"sensibilidad y especificidad"* (en inglés, *"sensitivity and specificity"*) se realiza un cribado preseleccionando los artículos que nos pueden interesar.

Es difícil seleccionar los estudios de exactitud diagnóstica porque no aparecen correctamente clasificados en la Librería Nacional de Medicina (*NLM*, *National Library of Medicine*). Usar *"sensitivity and specificity"* como palabras clave no es del todo una buena estrategia, por lo que realizamos otras búsquedas más sensibles: se realizan nuevos registros aplicando otras palabras clave indicadoras también de estudios de sensibilidad y especificidad diagnóstica como son: curva ROC y área bajo la curva (AUC), coeficiente de verosimilitud, y valor predictivo positivo y negativo. Los revisores sistemáticos recomiendan la búsqueda manual como la mejor estrategia [164], por lo que iniciamos esta exploración, que nos permitió constatar que algunos artículos a analizar no habían sido seleccionados con las búsquedas informáticas.

Finalmente, obtenemos todos los artículos originales que tratan sobre pruebas diagnósticas de laboratorio (microbiología, bioquímica, hematología, hormonas, inmunología y genética) con utilidad clínica publicados desde el año 1997 hasta el 2000, ambos inclusive, en *Medicina Clínica (Barc)* y en *Revista Clínica Española* y durante los años 1996, 2001 y 2002 en *Clinical Chemistry* y en *Clinical Chemistry and Laboratory Medicine*.

Conforme se va haciendo la selección de artículos, tanto de manera informatizada como manual, se procede a la lectura del resumen de los estudios y se van seleccionando aquellos que cumplen los criterios de inclusión, rechazando aquellos que cumplen los de exclusión:

Se incluyen todos aquellos artículos que comunican investigaciones sobre pruebas diagnósticas de laboratorio aplicadas a humanos y que en el título o en el resumen contienen las palabras clave "sensibilidad y especificidad" o un equivalente que denote que se ha evaluado la exactitud diagnóstica.

Se excluyen del estudio aquellos artículos que contienen estudios de solo sensibilidad o solo especificidad diagnóstica pero no de ambas; los trabajos que no versan sobre pruebas diagnósticas de laboratorio (las pruebas radiodiagnósticas por ejemplo, quedan excluidas), y las revisiones, cartas al editor o notas clínicas (solo analizamos artículos originales).

5.1.2- PROCEDIMIENTO DEL ESTUDIO.

En este estudio hemos participado tres examinadores: dos de ellos, I. Hernández-Aguado y J.M. Ramos, tenían ya experiencia en este tipo de trabajos, y han ayudado a marcar las directrices del trabajo del tercero, B. Lumbreras.

En el inicio contamos con dos fichas metodológicas: una siguiendo las normas de Reid y colaboradores [5] (Anexo I), y otra que incluye criterios propios, ficha de Hernández y García [7, 8] (Anexo II). Para unificar los estándares de estas dos plantillas y valorar cuáles son los más idóneos, cada uno de los evaluadores y de manera independiente aplica estas dos fichas a seis artículos seleccionados al azar de las dos publicaciones nacionales (tres de *Medicina Clínica (Barc)* y tres de *Revista Clínica Española*).

En una primera reunión se ponen los resultados en común y se examinan las concordancias y discrepancias obtenidas. Tras constatar el solapamiento de las guías en algunos criterios y la poca precisión de algunos de los criterios de la guía de Hernández y García [7, 8], se decide elaborar una ficha conjunta. Esta nueva plantilla (Anexo III), cuenta con los siete criterios metodológicos de Reid y colaboradores [5] (señalados con un asterisco*) a los que se añaden otros de la guía propia como son: los referentes a los objetivos del estudio y al patrón de referencia utilizado, la descripción del método a evaluar y la definición de término normal, la descripción del origen de la población a estudio, y con referencia al apartado de resultados, la expresión de los mismos de forma continua, el índice para pruebas conjuntas si se valora más de un test con el mismo objetivo diagnóstico, y el cálculo del valor predictivo. Eliminamos algunos de los criterios de la guía de Hernández y García [7, 8], tales como algunos aspectos relativos al patrón de referencia, como el indicar si es un estándar aceptado o si es necesario su realización de forma estandarizada; la cuestión sobre en qué fase se encuentra el estudio diagnóstico y los criterios referentes a las conclusiones y aspectos formales del estudio, todo ello por considerarse menos relevante en un estudio de exactitud diagnóstica. Además, unificamos algunos parámetros como los descriptores del criterio referente al diseño del estudio, ya que la secuencia de aplicación de la prueba diagnóstica y la de referencia es un concepto incluido en la prevención del sesgo de secuencia.

En conclusión, se utilizó este cuestionario propio para la evaluación de los estudios de las pruebas diagnósticas con la finalidad de ser más exhaustivos, flexibilizar contenidos para hacerlo aplicable a cualquier tipo de estudio e incluir cuestiones no recogidas en cuestionarios convencionales. Su explicación es sencilla por lo claro de algunas cuestiones, por seguir recomendaciones de cuestionarios previos y porque en cuestiones comunes con el cuestionario de Reid y colaboradores [5] se siguen las definiciones de éste, ya que se han aplicado conjuntamente.

Para validar esta nueva plantilla metodológica, los tres investigadores aplicamos estas nuevas recomendaciones a los seis trabajos que previamente se han analizado. Volvemos a comparar los resultados obtenidos, lo que nos ayuda a definir más claramente cada uno de los estándares propuestos y a minimizar las posibles dudas que hayan surgido. Una vez descritos los

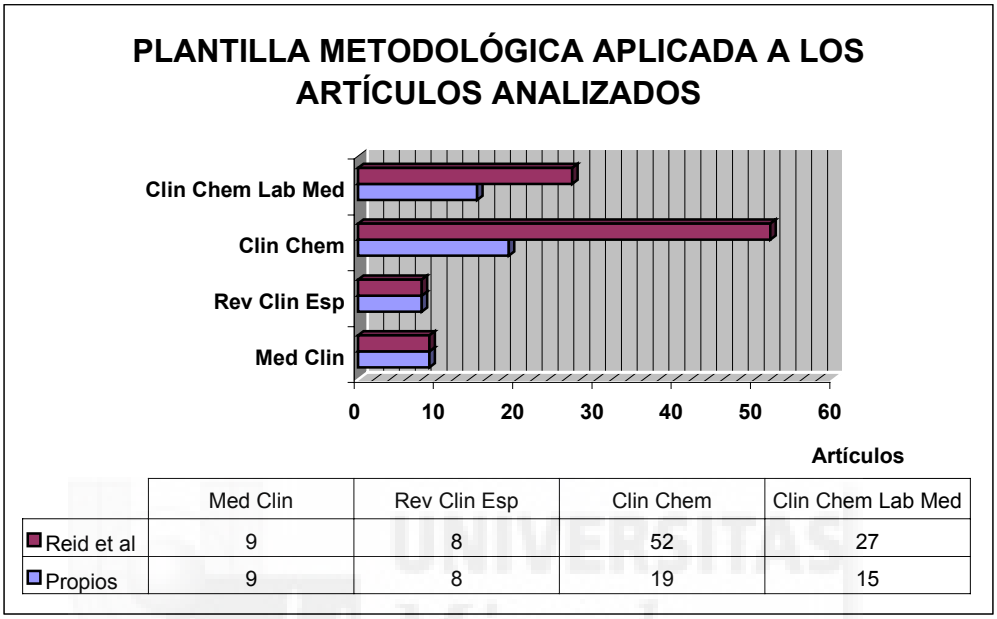
criterios de calidad metodológica de esta nueva guía y redactado su forma de aplicación, seguimos analizando el resto de los artículos.

Cada artículo se analiza por un mínimo de dos evaluadores. De esta manera podremos verificar los resultados obtenidos y describir aquellos criterios cuya aplicación conlleva más dificultad. Para ello, uno de los evaluadores analiza todos los artículos, mientras que los distribuye de manera aleatoria entre los otros dos examinadores expertos para su evaluación.

Cada cierto número de artículos analizados, los tres evaluadores nos reunimos. En cada artículo comparamos los resultados obtenidos para cada criterio metodológico. En caso de discordancia en la respuesta entre dos de los evaluadores, el tercero, que no ha analizado ese artículo, estudia el trabajo en cuestión y lo evalúa con la plantilla metodológica. Entre los tres se llega a un consenso que sirve de base para futuras decisiones del cumplimiento metodológico.

No aplicamos a todos los trabajos evaluados esta plantilla metodológica clave: se utiliza en el análisis de los 17 artículos de las revistas de ámbito nacional y en una muestra de 34 de los 79 trabajos estudiados de publicaciones de ámbito internacional; a los otros 45 estudios pertenecientes a las dos revistas de ámbito internacional, les aplicamos únicamente los criterios de Reid y colaboradores [5]. Esto es debido, a que una vez analizamos una serie de artículos con ambas guías metodológicas, el editor de la revista *Clinical Chemistry*, nos sugirió que aumentáramos el número de trabajos a evaluar, ampliación a la que solo aplicamos esta guía de recomendaciones. Es decir, todos los estudios reciben una evaluación metodológica con los criterios de la guía de Reid y colaboradores [5] y 51 trabajos se evaluaron con la guía elaborada para la presente tesis doctoral (de esta manera, podremos determinar la aplicabilidad de estos criterios). (Figura 2)

Figura 2: Representación gráfica del tipo de plantilla metodológica aplicada, (la compuesta por criterios propuestos por los autores [7, 8] *y la utilizada por Reid y colaboradores* [5]*) a cada uno de los 96 artículos analizados de las cuatro revistas en estudio.*



	Med Clin	Rev Clin Esp	Clin Chem	Clin Chem Lab Med
Reid et al	9	8	52	27
Propios	9	8	19	15

5.1.3- GUÍA PARA LA REVISIÓN METODOLÓGICA DE LOS ARTÍCULOS.

Los artículos previamente seleccionados para el estudio mediante el cumplimiento de los criterios de inclusión: estudios de sensibilidad y especificidad o cociente de verosimilitud sobre pruebas diagnósticas de laboratorio aplicadas al hombre y con utilidad clínica, son analizados a través de la aplicación de la ficha de calidad metodológica previamente elaborada.

La plantilla metodológica considera cinco apartados: objetivos, patrón de referencia, método de realización de la prueba a valorar, diseño del estudio y resultados. Los apartados que tienen un asterisco han sido extraídos de los criterios de Reid y colaboradores [5].

- Se debe valorar si se han establecido previamente los objetivos que se quieren cumplir en ese estudio. El investigador debe definir claramente qué pretende estudiar. Entre otros aspectos, debe especificarse si la prueba a valorar es nueva, si puede aportar ventajas sobre

las que se emplean habitualmente o si es una prueba conocida pero se propone su uso para otra función o actividad clínica.

- Se especifican los objetivos del estudio cuando los autores de forma explícita indican qué problema diagnóstico tienen y qué prueba pretenden investigar para solucionarlo.
- La valoración de una prueba diagnóstica está justificada si los autores señalan explícitamente alguna de estas situaciones: se trata de una prueba nueva, de una prueba de uso anterior con aplicación novedosa o de una antigua con controversias en su aplicación. Si no es nueva, los autores deben apoyarse en referencias bibliográficas indicando explícitamente qué aspectos no resueltos van a estudiarse, y si lo es, debe decirse claramente que no hay nada escrito sobre este tema.

➢ Debe especificarse también el patrón de referencia, que será la prueba estándar con la que vamos a comparar la prueba diagnóstica a valorar. Se trata de la prueba diagnóstica empleada habitualmente en la clínica para hacer el diagnóstico definitivo.

- Aparece especificado solo si se explica claramente en qué consiste la aplicación de esta prueba estándar, tanto en los sujetos como en los controles (en caso de que los haya). Si la prevalencia de la enfermedad es muy baja y así lo indiquen los autores, se eligen voluntarios sanos como grupo de comparación, a los que no se les aplica el patrón de referencia.
- Se previene el sesgo de incorporación cuando en la definición de la prueba de referencia no está incluido alguno de los resultados de la prueba diagnóstica que estamos valorando.
- Está realizado en toda la serie cuando se especifica claramente y de forma explícita, que la prueba de referencia se realiza de manera sistemática en todos los pacientes que forman parte del estudio a no ser que, como antes se ha mencionado, de uno de los grupos a estudio se tenga con un grado de seguridad razonable la certeza de ausencia de enfermedad y así se indique.

- El siguiente punto a describir en el estudio, es el método de realización de la prueba diagnóstica que se está pretendiendo evaluar.
  - Cuando la prueba diagnóstica a valorar se trata de una prueba nueva o nueva con alguna modificación, el método está suficientemente descrito si los autores dan suficiente detalle como para que pueda reproducirse. Si se trata de una prueba de uso habitual, basta con una referencia bibliográfica o comercial.
  - (*) Reproducibilidad: En trabajos cuya prueba diagnóstica es valorada por un observador, se debe incluir alguna prueba que evalúe la variabilidad del sujeto para disminuir al mínimo la subjetividad. En trabajos sin interpretación del observador, se debe medir la variabilidad de la técnica (no son suficientes los valores que presentan las casas comerciales).
  - Definición de los términos normalidad / anormalidad: los autores especifican qué resultado de la prueba diagnóstica van a considerar como positivo o negativo.

- En el apartado del diseño del estudio se recogen el ámbito de realización (el nivel asistencial, tipo de laboratorio, etc.), la descripción del origen de la población a estudio, el espectro de sujetos y la prevención de los sesgos de secuencia o verificación diagnóstica y de revisión o comparación ciega.
  - Descripción del origen de la población a estudio: está explicado cuando se cumplen por lo menos dos de los siguientes tres criterios: descripción de los filtros asistenciales o las formas de acceso que han tenido los pacientes para llegar al centro asistencial donde se está realizando el estudio; la muestra del estudio es un reflejo de la población clínica (cuando la población a estudio es una muestra consecutiva aleatoria de aquella población clínica real en la que se pretende aplicar la prueba diagnóstica) y aparece un criterio de selección final de la muestra (los autores cuantifican la población candidata el estudio y describen cuántos de ellos fueron finalmente elegidos para el estudio).
  - (*) El espectro de los sujetos. Para que se cumpla este criterio deben aparecer por lo menos tres de los siguientes cuatro supuestos: deben estar definidas las características demográficas (edad y sexo), síntomas clínicos o estadios de la enfermedad (severidad de

la enfermedad, duración y morbilidad) y los criterios que se han seguido para la selección de los sujetos (se incluyen tanto criterios de inclusión como de exclusión).

- (*) Prevención del sesgo de secuencia o verificación diagnóstica: el artículo se considera libre del sesgo si todos los sujetos del estudio (independientemente del resultado de la prueba diagnóstica) reciben un diagnóstico definitivo mediante la prueba de referencia. Se considera que está presente si los sujetos con resultados de la prueba positivos o negativos no tienen igual oportunidad de someterse a la prueba de referencia.
- (*) Prevención del sesgo de revisión o comparación ciega: para estudios de cohortes prospectivos en los que el paciente recibe primero la prueba diagnóstica, el estándar es aceptado si la prueba de referencia se evalúa independientemente, es decir, sin conocer el resultado de la prueba diagnóstica. En estudios prospectivos si la prueba de referencia precede a la diagnóstica y en las series de casos y controles, se cumple el estándar si aparece una afirmación de la independencia de la interpretación de los resultados.

- Es importante, tener especial cuidado tanto en la expresión como en la interpretación de los resultados. Hay que elegir la medida de frecuencia en la que se expresan los resultados de acuerdo con el objetivo de medición del estudio.
  - Si la prueba diagnóstica es susceptible de expresar sus resultados de forma continua o discreta en más de dos categorías, se valora si los autores presentan los resultados en forma de curvas ROC, cociente de verosimilitud o mediante el cálculo de la sensibilidad y especificidad en distintos puntos de corte.
  - Cuando el estudio evalúa más de una prueba con el mismo objetivo diagnóstico, se valora si los autores presentan índices de exactitud para expresar los resultados de estas pruebas conjuntamente.
  - El cálculo del valor predictivo de una prueba es necesario cuando la muestra a estudio representa a la población clínica real, y por tanto, debe calcularse.
  - (*) Los resultados deben aparecer distribuidos por estratos. Se cumple este criterio, cuando los índices de exactitud se expresan según distintos subgrupos clínicos o demográficos de la población estudiada.

- (*) El criterio de la precisión de los resultados se cumple cuando se presenta la precisión estadística de las estimaciones de los índices de exactitud mediante el cálculo del intervalo de confianza o error estándar.
- (*) Presentación o no de los resultados indeterminados obtenidos. Se cumple cuando se presentan todos los criterios de la prueba a valorar, tanto los positivos y negativos como los indeterminados, y además recoge la inclusión o exclusión de los resultados indeterminados en el cálculo de los índices de exactitud.

5.1.4- VARIABLES SECUNDARIAS QUE PUEDEN INFLUIR EN LA CALIDAD METODOLÓGICA DE LOS ARTÍCULOS EVALUADOS.

Después de aplicar a los artículos las plantillas metodológicas, analizamos la influencia de determinadas características de los trabajos evaluados, en la consecución de una mejor calidad metodológica.

En primer lugar, estudiamos el número de centros que han participado en la investigación y el país de procedencia de cada uno de estos centros. Para esta última característica establecimos un total de cinco categorías: EEUU y Europa; EEUU; Europa; Asia y otros(Canadá; Méjico y Europa; Sudamérica; África y Australia).

También, analizamos el carácter de las publicaciones que hemos seleccionado para la evaluación metodológica. Estas pueden ser de carácter clínico general (*Medicina Clínica* y *Revista Clínica Española*) o exclusivo de laboratorio (*Clinical Chemistry* y *Clinical Chemistry and Laboratory Medicine*).

Por último, estudiamos el ámbito en el que se realiza el estudio de la prueba diagnóstica. Este puede ser clínico o universitario, haberse realizado en un laboratorio, o contar con la participación tanto de centros clínicos como de laboratorio.

5.1.5- ANÁLISIS ESTADÍSTICOS.

La creación, gestión de la base de datos y los análisis siguientes de los resultados se efectúan con el programa SPSS (versión 11.5 para Windows Inc., Chicago, Illinois).

El análisis para la comparación de los resultados obtenidos en los distintos periodos de estudio u otras características, se realizará utilizando las técnicas habituales para estos casos: pruebas ji-cuadrado, t-student, anova, U de Mann-Whitney, exacta de Fischer o Kruskall-Wallis, según proceda.

El cálculo de la variabilidad entre los observadores, se estimó a partir de las discrepancias para cada uno de los ítems analizados. Se determinó el porcentaje de concordancia simple, y el obtenido para los criterios de Reid y colaboradores [5] y para los criterios propios [7, 8].

5.2- ESTUDIO DE VARIABILIDAD ENTRE OBSERVADORES: ANÁLISIS DE CRISTALES DE URATO MONOSÓDICO Y PIROFOSFATO CÁLCICO DIHIDRATADO EN MUESTRAS DE LÍQUIDO SINOVIAL

Esta segunda parte del estudio corresponde a la realización del estudio de variabilidad observacional de una prueba diagnóstica de laboratorio. Como ya he referido, estudiamos la concordancia en la detección e identificación de cristales de urato monosódico y pirofosfato cálcico en muestras de líquido sinovial entre varios investigadores. Se trata de un estudio eminentemente de concordancia, pero debido a la disponibilidad de un experto en este tipo de análisis, evaluamos también la exactitud de la prueba tomándolo como patrón de referencia.

5.2.1- TIPO DE ESTUDIO:

Estudio prospectivo de concordancia, con enmascaramiento posterior a una intervención formativa específica.

5.2.2- ORIGEN DE LAS MUESTRAS.

Se incluyen de manera consecutiva las muestras de líquido sinovial de todos los pacientes que acuden a Consultas Externas o al Hospital de Día de la Sección de Reumatología del Hospital General Universitario de Alicante, por sospecha de artropatía por cristales y otras artropatías inflamatorias, desde septiembre del año 2001 hasta junio del año 2003.

Se excluyen aquellas muestras que no tienen un volumen suficiente para su evaluación por todos los observadores participantes.

5.2.3- PARTICIPANTES EN EL ESTUDIO.

- Observadores: Las muestras son estudiadas por cuatro analistas, residentes del Servicio de Análisis Clínicos del citado hospital, de manera ciega e independiente. Ninguno de ellos tiene experiencia previa en el análisis de cristales en líquido sinovial.

- Reumatólogo experto en la detección e identificación de cristales (E. Pascual): sirve de patrón de referencia o estándar con el que se comparan los resultados obtenidos por los cuatro observadores para el cálculo de la sensibilidad y especificidad diagnóstica: analiza la presencia de cristales en las muestras recibidas y posteriormente las envía al laboratorio para su evaluación por los cuatro analistas.

5.2.4- CURSO DE FORMACIÓN PARA INSTRUIR A LOS CUATRO OBSERVADORES PARTICIPANTES EN EL ESTUDIO.

El curso es impartido por E. Pascual, reumatólogo con reconocida experiencia en análisis de cristales en muestras de líquido sinovial, mediante el empleo sucesivo del microscopio óptico y de luz polarizada.

Consta de una sesión teórica en la que el reumatólogo experto mediante diapositivas, explica las características morfológicas y de birrefringencia de los cristales de urato monosódico y de pirofosfato cálcico dihidratado en muestras de líquido sinovial y el procedimiento que se debe emplear para la detección e identificación de cristales.

Más adelante, los analistas observan alícuotas de líquido sinovial sin cristales para familiarizarse con el aspecto de las células que se pueden encontrar en una muestra de líquido sinovial. De esta forma se detectará fácilmente la presencia de cristales como material extraño.

En un paso siguiente, los observadores evalúan muestras de líquido sinovial con y sin cristales de manera ciega e independiente, hasta que el reumatólogo experto considera que los analistas detectan e identifican ambos tipos de cristales de manera apropiada:

Analizan un total de 30 muestras de líquido sinovial bajo supervisión del experto: 5 muestras de líquidos con cristales de urato monosódico (20%) en las que se obtiene un 100% de aciertos; 5 muestras de líquidos con cristales de urato monosódico (20%) en las que se obtiene un 80% de

aciertos y 15 muestras de líquidos sinoviales sin cristales (60%) que se analizan correctamente en su totalidad.

Se realiza una sesión final para aclarar las posibles dudas que hayan podido surgir y repasar los conceptos fundamentales del procedimiento.

A continuación se detallan las guías seguidas en el curso de formación y que se van a seguir durante la evaluación de la técnica:

5.2.5- GUÍAS BÁSICAS DEL CURSO DE ENTRENAMIENTO.

Para el análisis de cristales de muestras de líquido sinovial se utilizan una serie de herramientas:

- Microscopio de 400 aumentos, *Olympus BH*. A 1000 aumentos probablemente se detecten mejor los cristales pequeños, sobre todo de pirofosfato cálcico dihidratado. Pero en esta medida, la birrefringencia puede ser menos llamativa.

- Filtros polarizados: uno en la posición normal debajo del condensador del microscopio, y el otro en cualquier posición entre el microscopio y el ocular. A falta de un microscopio polarizado, un observador habituado puede apreciar mediante un microscopio ordinario la presencia de cristales e incluso identificarlos como urato monosódico o pirofosfato cálcico con razonable certeza [165].

- Compensador rojo de primer orden: útil para clasificar la birrefringencia como positiva o negativa. Está en cualquier posición entre los filtros polarizados. La detección de cristales es precisa sin esta pieza y la identificación también es posible, pero la identificación estándar requiere el compensador.

Al menos para fines formativos, consideramos que el análisis de cristales ha de realizarse en dos pasos consecutivos:

1) Detección de los cristales: Para asegurar que los cristales no van a pasar desapercibidos.
2) Identificación de cristales: Determinación del tipo de cristal detectado en el paso previo.

1) Detección de cristales

a) Detección de cristales de urato monosódico: Se coloca una gota de la muestra de líquido sinovial en un portaobjetos limpio (si está sucio puede tener artefactos birrefringentes) y se enfoca la preparación con el objetivo de 400X. Bajo la luz ordinaria y sin filtros polarizados, la mayoría de los cristales de urato monosódico aparecen como agujas largas y finas [166], intra o extracelulares (los cristales intracelulares aparecen generalmente en las muestras de líquido sinovial de articulaciones asintomáticas [49]).

Si cruzamos los filtros polarizados el campo del microscopio aparece oscuro: la luz de la fuente polarizada por el filtro del analizador y desviada a su paso a través de los cristales de urato monosódico, al atravesar el segundo filtro polarizado hará que los cristales aparezcan como barras delgadas muy brillantes (birrefringencia). Todos los cristales de urato monosódico muestran la misma birrefringencia. (Figura 3).

b) Detección de cristales de pirofosfato cálcico dihidratado: Se obtiene una muestra de líquido sinovial y se coloca una gota en un portaobjetos limpio, observándose también con el objetivo de 400X del microscopio de luz ordinaria (con 1000X y aceite de inmersión se distinguen mejor y puede servir para entrenamiento).

Con luz ordinaria se buscan los cristales por su morfología: la forma varía desde barras muy finas a cristales romboidales o paralelepípedos de tamaños diversos y son con frecuencia intracelulares [167] (Figura 4). La mayoría de los cristales de cálcico dihidratado apenas muestran birrefringencia, por lo que pueden no ser detectados bajo luz polarizada [66].

Figura 3: Cristales de urato monosódico en una muestra de líquido sinovial, vistos al microscopio de luz ordinaria (izquierda) y de luz polarizada (derecha). Se puede observar su fuerte birrefringencia bajo luz polarizada.

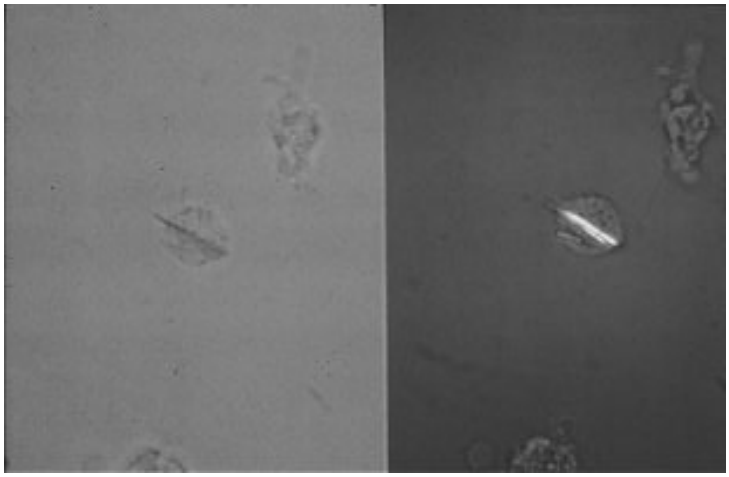
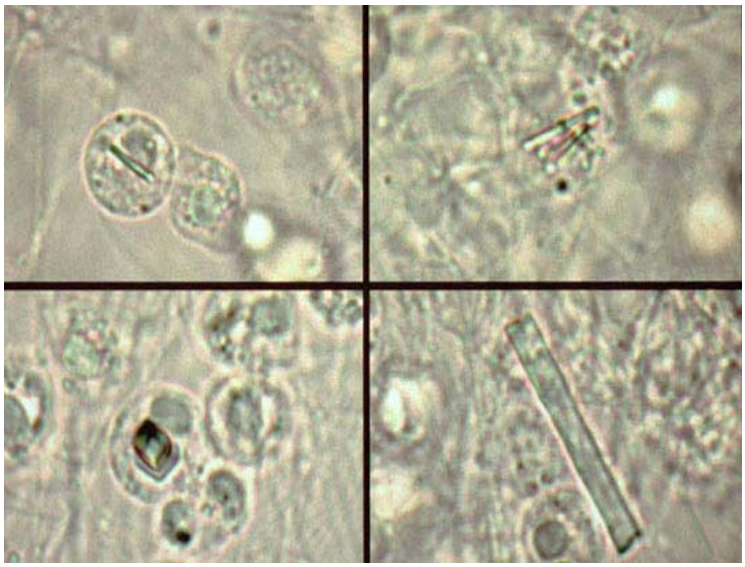


Figura 4: Cristales de pirofosfato cálcico dihidratado en muestras de líquido sinovial, vistos al microscopio de luz ordinaria. Se pueden observar las distintas formas que puede adoptar un cristal de pirofosfato cálcico dihidratado.



2) Identificación de cristales

a) Identificación de cristales de urato monosódico: Todos los cristales de urato monosódico tienen forma de aguja y son fuertemente birrefringentes, cuando los observamos con los filtros polarizados sin el compensador rojo de primer orden. La forma y las características ópticas de los cristales son altamente orientativas y pueden ser suficientes para la identificación cuando hay solamente un tipo de cristal. Pero en una misma muestra de líquido sinovial pueden estar presentes ambos tipos de cristales, urato monosódico y pirofosfato cálcico dihidratado [61], por lo que es necesario usar los filtros polarizados y un compensador rojo de primer orden: el cristal cuyo eje longitudinal es paralelo al eje del compensador (marcado como z) aparece amarillo brillante; si es perpendicular, aparece azul brillante. Este fenómeno se conoce como birrefringencia negativa (Figura 5).

b) Identificación de cristales de pirofosfato cálcico dihidratado: Los cristales de pirofosfato cálcico dihidratado tienen diversas formas, pero solo los cristales aciculares se pueden confundir con los de urato monosódico. Para distinguirlos usamos los filtros polarizados y un compensador rojo de primer orden: el cristal cuyo eje longitudinal es paralelo al eje del compensador (marcado como z) aparece azul pálido y si es perpendicular aparece amarillo pálido. Este fenómeno se conoce como birrefringencia positiva (Figura 6).

La identificación definitiva de los cristales requiere el examen de una muestra mediante difracción de rayos X, pero este procedimiento no está generalmente disponible y solo se usa en proyectos de investigación.

Figura 5: Cristales de urato monosódico observados con el compensador rojo de primer orden y los filtros polarizados. Se puede comprobar cómo los cristales paralelos al eje del compensador aparecen amarillo brillante y los cristales perpendiculares, aparecen azul brillante.

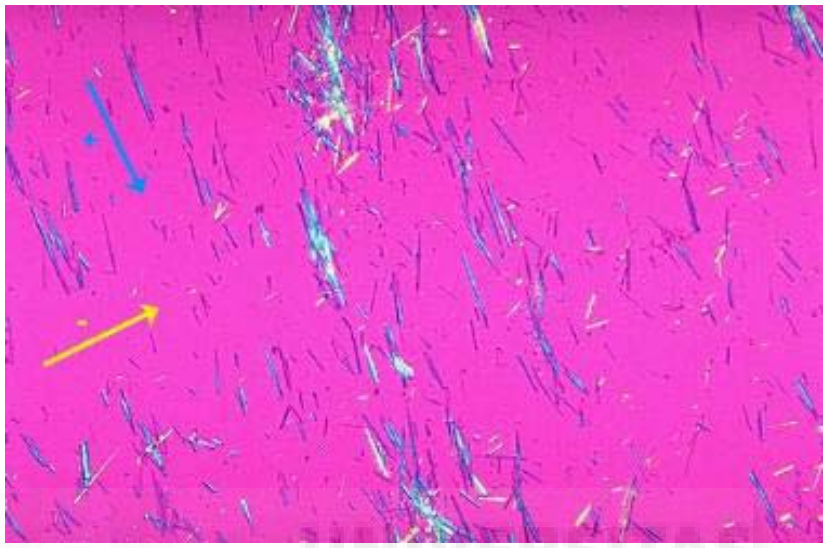
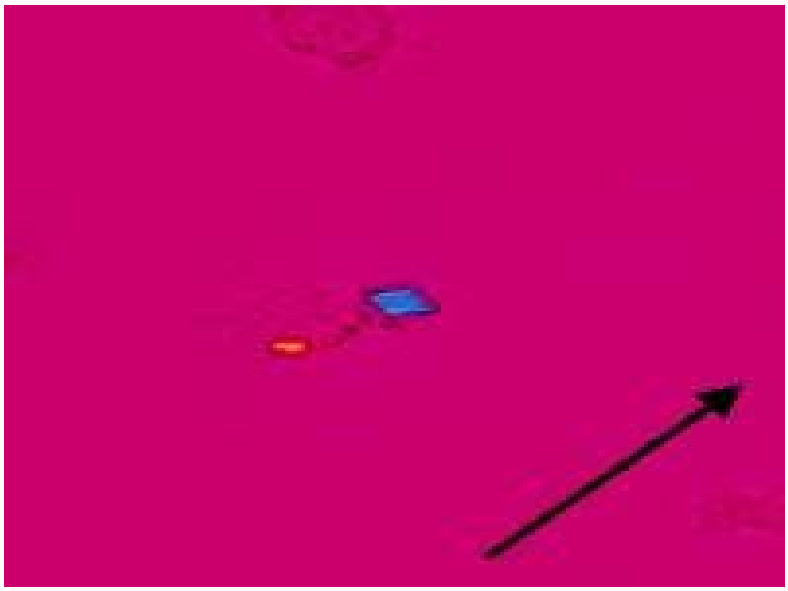


Figura 6: Cristales de pirofosfato cálcico dihidratado observados con el compensador rojo de primer orden y los filtros polarizados. Se puede comprobar cómo el cristal paralelo al eje del compensador aparece azul pálido y el cristal perpendicular, aparece amarillo pálido.



5.2.6- EXTRACCIÓN DE LAS MUESTRAS PARA SU POSTERIOR ANÁLISIS.

La aspiración articular se lleva a cabo por el reumatólogo en condiciones estériles. La muestra debe ser tomada antes de la aplicación de un tratamiento con esteroides, ya que los cristales de esteroides pueden ser confundidos con los cristales a estudio. El líquido sinovial obtenido se recoge mediante agujas estériles y jeringa de plástico, ambas desechables, para evitar contaminación a partir de material birrefringente estéril. Puede humedecerse la jeringa aproximadamente con 25U de heparina/ml de líquido sinovial como anticoagulante, pero deben evitarse los anticoagulantes en polvo tales como oxalato y EDTA, puesto que pueden presentar artefactos que induzcan a confusión en el examen microscópico [168].

5.2.7- PROCEDIMIENTO DEL ESTUDIO

Todas las muestras de líquido sinovial son examinadas y clasificadas por el reumatólogo experto que va a ser el patrón de referencia y que conoce el origen de la muestra. Recoge en un cuaderno las distintas variables a estudiar como son la hora de extracción de la muestra o el nombre y diagnóstico clínico del paciente. Detecta si hay o no cristales y en su caso, los identifica como de urato monosódico o de pirofosfato cálcico (Tabla 3).

Tabla 3: Cuaderno de anotaciones del reumatólogo experto, patrón de referencia en la detección e identificación de cristales en muestras de líquido sinovial:

Fecha
Número de líquido*
Hora de extracción de la muestra
Nombre del paciente
Diagnóstico clínico
Infiltración en los últimos seis meses
Tipo de cristal

** Dos numeraciones: una para los líquidos obtenidos en Consultas Externas y otra para los extraídos en el Hospital de Día del Servicio de Reumatología*.

La muestra examinada se divide en alícuotas (una por cada observador que va a analizar la muestra) y se remite al laboratorio de Análisis Clínicos en el menor tiempo posible (todas las observaciones se realizan en las primeras dos horas desde la extracción de la muestra). Dichas muestras únicamente tienen una identificación del número de líquido de que se trata, pero no una indicación del diagnóstico del paciente al que pertenece.

Los cuatro observadores del estudio de manera independiente y ciega, evalúan la muestra con el mismo microscopio empleado por el experto. Los analistas mantienen el enmascaramiento durante todo el periodo que dura el estudio y anotan en registros independientes su juicio diagnóstico y una explicación detallada de lo que observan.

5.2.8- EXPRESIÓN DE LOS RESULTADOS OBTENIDOS POR LOS ANALISTAS

Antes de empezar el estudio se han elegido y consensuado las escalas de expresión de los resultados: si se detectan cristales en la muestra a estudio se debe indicar cuántos campos han sido necesarios analizar hasta observar su presencia y en el caso de que haya sido en el primer campo, si había menos o más de cinco cristales en dicho campo. (Tabla 4)

Tabla 4: Cuaderno de anotaciones de los cuatro observadores que participan en el estudio, para la detección e identificación de cristales en muestras de líquido sinovial:

Fecha:	Hora de observación:	Número:
DETECCIÓN		
Luz ordinaria	Cristales en el primer campo	Ninguno; <5; >5.
	Nº campo del primer cristal	
	Tiempo empleado	
Luz polarizada	Cristales en el primer campo	Ninguno; <5; >5.
	Nº campo del primer cristal	
	Tiempo empleado	
IDENTIFICACIÓN		
	Tipo de cristal	Urato monosódico; pirofosfato cálcico; ambos; otros
	Presencia de birrefringencia	
	Tiempo empleado	

Para calcular los valores de sensibilidad y especificidad diagnóstica y valores predictivos positivo y negativo en la detección e identificación de cristales en muestras de líquidos sinovial, el diagnóstico definitivo de los análisis realizados por los cuatro evaluadores lo aporta el reumatólogo que previamente ha examinado las muestras. Tomamos cada observación realizada por un observador como un único resultado, teniendo en cuenta que no se trata de muestras independientes, ya que todos analizan alícuotas de la misma muestra a estudio.

Para el cálculo de la exactitud diagnóstica en el primer paso, la detección de cristales, tomamos como resultado verdadero positivo si tanto el reumatólogo como el observador, detectan la presencia de un cristal en la muestra; los resultados verdaderos negativos son aquellos en los que ni el patrón de referencia ni el observador detectan la presencia de un cristal en la alícuota de líquido sinovial; los resultados falsos positivos aparecen cuando el patrón de referencia no detecta ningún cristal pero el observador sí, y los resultados falsos negativos son aquellos en los que el patrón de referencia no detecta la presencia de cristales en las muestras analizadas pero sí lo hace el observador.

Respecto a la identificación de cristales, y para explicar más claramente los resultados obtenidos, separaremos las determinaciones en dos grupos en función de si el cristal identificado es de urato monosódico o pirofosfato cálcico. En relación con las muestras que contienen cristales de urato monosódico, definimos resultados verdaderos positivos cuando tanto el patrón de referencia como el observador han identificado el cristal detectado como urato; resultados verdaderos negativos son aquellos en los que el observador y el reumatólogo experto coinciden en determinar que el cristal detectado no es de urato; los resultados falsos positivos son aquellos en los que el observador identifica el cristal detectado como de urato cuando es de pirofosfato, y los resultados falsos negativos el caso contrario, el patrón de referencia identifica el cristal como urato pero no así el observador. Con los cristales de pirofosfato cálcico dihidratado, procedemos de igual manera.

5.2.9- GRADO DE ACUERDO ALCANZADO ENTRE LOS OBSERVADORES DURANTE EL ANÁLISIS DE LAS MUESTRAS DE LÍQUIDO SINOVIAL.

Consideramos que se alcanza el acuerdo entre la expresión de los resultados de un analista y el reumatólogo, patrón de referencia, si ambos detectan la presencia de cristales independientemente del número de campo que ha sido necesario observar, del tiempo empleado para ello y del tipo de luz utilizada.

Así mismo, el resultado para la identificación de cristales es concordante entre ambos cuando ambos describen finalmente el mismo tipo de cristal, independientemente del tiempo empleado para la identificación.

5.2.10- OTRAS DETERMINACIONES REALIZADAS EN LAS MUESTRAS DE LÍQUIDOS SINOVIAL ANALIZADAS Y QUE PUEDEN INTERFERIR EN LOS RESULTADOS DEL ESTUDIO DE CONCORDANCIA.

1- En cada muestra se realiza un recuento celular: número de leucocitos y hematíes. Se lleva a cabo de forma manual utilizando una cámara hemocitométrica (*Neubauer Improved*) y diluyendo la muestra 1/10 con suero salino en recuentos altos. También se podía haber hecho de forma automática por tener una alta precisión, pero la forma manual es la recomendada [169].

Se estudia también el diferencial leucocitario. Se emplea ácido acético que lisa los hematíes y pone en relieve los núcleos de los leucocitos, permitiendo su clasificación en polimorfonucleares y células mononucleadas.

2- Por último, determinamos en cada muestra los parámetros bioquímicos: glucosa y proteínas. Se utiliza el analizador automático *Hitachi 917* (*Roche Diagnostic*). Las muestras con alta celularidad se centrifugan para determinar los parámetros en el sobrenadante.

Por lo tanto, las principales variables a estudio son la concordancia de los observadores participantes en la determinación de la presencia o ausencia de cristales en las muestras de líquido sinovial a estudio y en el tipo de cristal identificado, mediante el uso sucesivo del microscopio óptico y el de luz polarizada.

Otras variables a estudio son el recuento y diferencial leucocitario e información de la propia observación (tiempo transcurrido entre la artrocentesis y el análisis, tiempo empleado en el análisis y número de campo en el que se detecta el primer cristal)

5.2.11- ANÁLISIS ESTADÍSTICOS.

La creación, gestión de la base de datos y los análisis descriptivos de los resultados se efectuaron con el programa SPSS (versión 11.5 para Windows Inc., Chicago, Illinois).

Las determinaciones de sensibilidad y especificidad diagnóstica, así como su intervalo de confianza al 95%, se calculan mediante el paquete estadístico Epiinfo (versión 6.0).

El cálculo de la variabilidad entre los observadores, se estimó a partir del índice de concordancia kappa, mediante el paquete estadístico STATA 8.0 (un valor por encima de 0.61 indica que la concordancia es buena, siempre que la prevalencia no sea muy extrema, es decir, mayor del 90% o menor del 10%).

6- RESULTADOS

Al igual que en el apartado referente a la metodología empleada en el estudio, dividimos la sección de resultados en dos partes:

1- En el primer apartado se muestran los resultados obtenidos tras la revisión crítica de la metodología de los artículos que tratan la exactitud diagnóstica de diversas pruebas de laboratorio, tanto en revistas nacionales como internacionales.

2- En el segundo apartado se analiza la determinación de la variabilidad entre los observadores que analizan muestras de líquido sinovial para la detección de cristales de urato monosódico y de pirofosfato cálcico.

6.1- REVISIÓN CRÍTICA DE LOS TRABAJOS QUE ESTUDIAN PRUEBAS DIAGNÓSTICAS DE LABORATORIO, Y PUBLICADOS EN REVISTAS CIENTÍFICAS NACIONALES E INTERNACIONALES

Respecto al estudio de la calidad metodológica de los trabajos publicados en revistas científicas, exponemos en primer lugar los resultados obtenidos tras la valoración de los artículos que estudian pruebas diagnósticas de laboratorio de las dos publicaciones nacionales, *Medicina Clínica (Barc)* y *Revista Clínica Española*. Después, analizamos los datos resultantes de la evaluación de las dos publicaciones internacionales, *Clinical Chemistry* y *Clinical Chemistry and Laboratory Medicine*.

Comparamos los resultados obtenidos en el presente trabajo, tanto de las publicaciones nacionales como internacionales, con los resultados del estudio previo de Reid y colaboradores [5], además de realizar una correlación entre los trabajos nacionales y las dos revisiones nacionales previas [7, 8]. Analizamos también la influencia en la calidad metodológica de los artículos, de determinadas características relativas con la autoría de los estudios y los centros donde se realizan.

Por último, describimos de manera cualitativa la aplicación de las guías metodológicas empleadas (la de Reid y colaboradores [5] y la propuesta por los autores [7, 8]) y las principales dificultades encontradas.

6.1.1- REVISTAS NACIONALES: *Medicina Clínica (Barcelona) y Revista Clínica Española*.

Dentro de este apartado referente a las publicaciones nacionales describimos en primer lugar, el procedimiento llevado a cabo para seleccionar los artículos que van a participar en la evaluación de la calidad de su metodología. En esta selección, tanto mediante la búsqueda manual como informática, obtuvimos los mismos artículos, por lo que solo describimos el procedimiento llevado a cabo mediante la estrategia informatizada.

Medicina Clínica (Barcelona)

- Durante el período de tiempo a estudio 1.997-2.000, se publican en esta revista un total de 2.095 artículos.
- De estos, al seleccionar con las palabras clave "sensibilidad y especificidad" y con otros términos que indican que se ha determinado la exactitud diagnóstica, se obtienen 54 artículos: 10 corresponden al año 1997, 10 a 1998, 18 a 1999 y 16 al año 2000.
- Al aplicar los criterios de inclusión y exclusión, el número de artículos se reduce a 9: 1 perteneciente al año 1997, 2 al año 1998, 2 a 1999 y 4 que corresponden al año 2000; excluyendo el resto de estudios por no tratar pruebas de laboratorio, tener formato de carta dirigida al editor, revisión o nota clínica, o estudiar solo sensibilidad o solo especificidad de la técnica diagnóstica.

Revista Clínica Española

- En los cuatro años estudiados, se publicaron en esta revista un total de 1.085 artículos.
- Tras preseleccionar acotando esta búsqueda con las palabras clave "sensibilidad y especificidad", además de con otros términos definitorios de exactitud diagnóstica,

nos quedamos con 22 artículos: 10 pertenecientes al año 1997, 9 a 1998, uno a 1999 y 2 al año 2000.

- Después de aplicar los criterios de inclusión y de exclusión, la búsqueda se reduce a 8 artículos: 4 publicados durante el año 1997, 3 del año 1998 y 1 del 2000. Al igual que en la publicación anterior, tuvimos que rechazar artículos por no ser artículos originales, no tratar pruebas de laboratorio o no calcular sensibilidad, especificidad o ambas de la prueba diagnóstica, ni sustitutos oportunos.

Los procedimientos diagnósticos evaluados en estos trabajos pertenecen a las áreas de laboratorio de bioquímica, inmunología, hormonas, microbiología y hematología. En la tabla 5 se muestra la frecuencia de aparición de estas secciones en los artículos según la publicación analizada.

Tabla 5: Áreas de laboratorio a las que pertenecen los 17 artículos revisados de las revistas de ámbito nacional Medicina Clínica (Barcelona) y Revista Clínica Española.

	Total (n/%)	***Med Clin* (n/%)**	***Rev Clin Esp* (n/%)**
Microbiología	7 (41%)	3 (33%)	4 (50%)
Bioquímica básica	5 (30%)	3 (33%)	2 (25%)
Inmunología	3 (18%)	2 (22%)	1 (13%)
Hematología	1 (6%)	1 (12%)	-
Hormonas	1 (6%)	-	1 (13%)
TOTAL	**17 (100%)**	**9 (100%)**	**8 (100%)**

Aunque en el análisis de los artículos, aplicamos conjuntamente los criterios de Reid y colaboradores [5] y los propuestos por nosotros [7, 8], al comentar los resultados distingo ambas evaluaciones.

6.1.1.1- RESULTADOS OBTENIDOS TRAS LA APLICACIÓN DE LOS CRITERIOS PROPIOS [7, 8].

Los resultados obtenidos después de la aplicación de la guía metodológica que contiene criterios propios de los autores en ambas revistas se reflejan en la tabla 6.

Hay que destacar que apenas en la mitad de los trabajos evaluados se justifica adecuadamente la necesidad de evaluar la prueba diagnóstica. En muchos trabajos tratan pruebas que son nuevas y no lo indican, o sin ser novedosas no incluyen ninguna referencia bibliográfica o trabajo anterior que las haya estudiado.

Respecto a la definición y características del patrón de referencia, éste se describe de manera idónea en casi todos los artículos, aunque en seis trabajos el patrón de referencia incorpora en su definición pruebas diagnósticas que luego va a evaluar, lo que provoca una sobreestimación de los índices de exactitud.

En referencia al método de realización de la prueba diagnóstica que se está valorando, el procedimiento aparece explicado de manera apropiada en la mayoría de los trabajos, bien con una cita bibliográfica o en su totalidad, incluida la definición de qué entienden los autores por valor normal.

En cambio, el aspecto del diseño del estudio relativo a la descripción del origen de la población, apenas aparece reflejado en solo uno de los artículos estudiados. Sobre todo, no está descrito el criterio que ha seguido el autor para seleccionar finalmente la muestra en estudio, ni ninguna especificación que nos aclare si la muestra refleja la población real.

Respecto a la comunicación de los resultados, aunque destaca que en ocho de los once artículos susceptibles de ambas revistas se expresan sus resultados de forma continua o discreta en más de dos categorías, de los catorce estudios que valoran más de una prueba con el mismo objetivo diagnóstico, solamente tres expresan sus resultados de manera conjunta.

Tabla 6: Descripción del cumplimiento de los estándares propuestos por nosotros [7, 8] en los diecisiete artículos analizados de las dos revistas nacionales, Medicina Clínica (Barcelona) y Revista Clínica Española.

	TOTAL (17) (n/%)	***Med Clin* (9) (n/%)**	***Rev Clín Esp* (8) (n/%)**
- Objetivos			
Especificados	**17 (100%)**	9 (100%)	8 (100%)
Justificación de la prueba	**8 (47%)**	5 (55%)	3 (37%)
- Patrón de referencia			
Especificado	**15 (88%)**	7 (77%)	8 (100%)
No incorpora la prueba a valorar	**11 (65%)**	6 (66%)	5 (63%)
Realizado en toda la serie	**12 (71%)**	6 (66%)	6 (75%)
- Prueba diagnóstica			
Descripción suficiente	**15 (88%)**	9 (100%)	6 (75%)
Definición normalidad	**14 (82%)**	8 (88%)	6 (75%)
- Diseño del estudio			
Descripción del origen de la población	**1 (6%)**	1 (11%)	0
- Resultados			
Expresión continua*	**15 (88 %)**	7 (77%)	8 (100%)
Índice para pruebas conjuntas**	**6 (35%)**	5 (55%)	1 (13%)
Cálculo valores predictivos***	**10 (59%)**	9 (100%)	1 (13%)

* Los resultados se expresan de forma continua en forma de curvas ROC, coeficiente de verosimilitud o calculaba la sensibilidad y especificidad en distintos puntos de corte cuando son susceptibles. También es correcto que no se expresen así, si no son susceptibles a ello.
** Para estudios que analizan más de una prueba con el mismo objetivo. También se cumple si solo se analiza una prueba, y por lo tanto no se calcula dicho índice.
*** Cuando es correcto realizarlo. Cuando no es pertinente su cálculo, también se valora el que no se determine.

En la tabla 7 se puede observar el número medio de criterios seguidos por artículo, para cada una de las dos revistas analizadas y para las dos en conjunto: la media de cumplimiento se sitúa entre seis y siete de los 11 ítems propuestos.

Tabla 7: Media de los criterios metodológicos propios [7, 8] cumplidos, para cada uno de los 17 artículos analizados de las revistas Medicina Clínica (Barc) y Revista Clínica Española.

	Artículos (n/%)	**Criterios/artículo (media/varianza)**
Med Clín (Barc)	9 (53%)	6,4 (1,1)
Rev Clin Esp	8 (47%)	6,9 (1,6)
TOTAL	**17 (100%)**	**6,7 (1,4)**

6.1.1.2- RESULTADOS OBTENIDOS TRAS LA APLICACIÓN DE LOS CRITERIOS DE REID Y COLABORADORES [5].

(Tabla 8 y figura 7).

El primer apartado de estos criterios metodológicos corresponde a la descripción de la composición del espectro. Este estándar se cumple ampliamente en la mayoría de los artículos evaluados: la edad y el sexo son los descriptores que más frecuentemente aparecen en los estudios, seguidos de los síntomas clínicos o estados de la enfermedad y de los criterios de exclusión e inclusión.

Aunque en la mayoría de los trabajos, se tiene cuidado de prevenir el sesgo de secuencia, no ocurre lo mismo con la prevención del sesgo de revisión o comparación ciega, que apenas se ha tenido en cuenta.

El criterio que recomienda la expresión de los valores de sensibilidad y especificidad en los distintos subgrupos, aparece reflejado en menos de la mitad de los artículos analizados.

Lo mismo sucede con el empleo del error estándar o intervalos de confianza para determinar la precisión de los resultados, que apenas aparece en la mitad de los trabajos evaluados. El número medio de pacientes incluidos en los trabajos fue de 224, con un rango comprendido entre 33 y 1.022 pacientes.

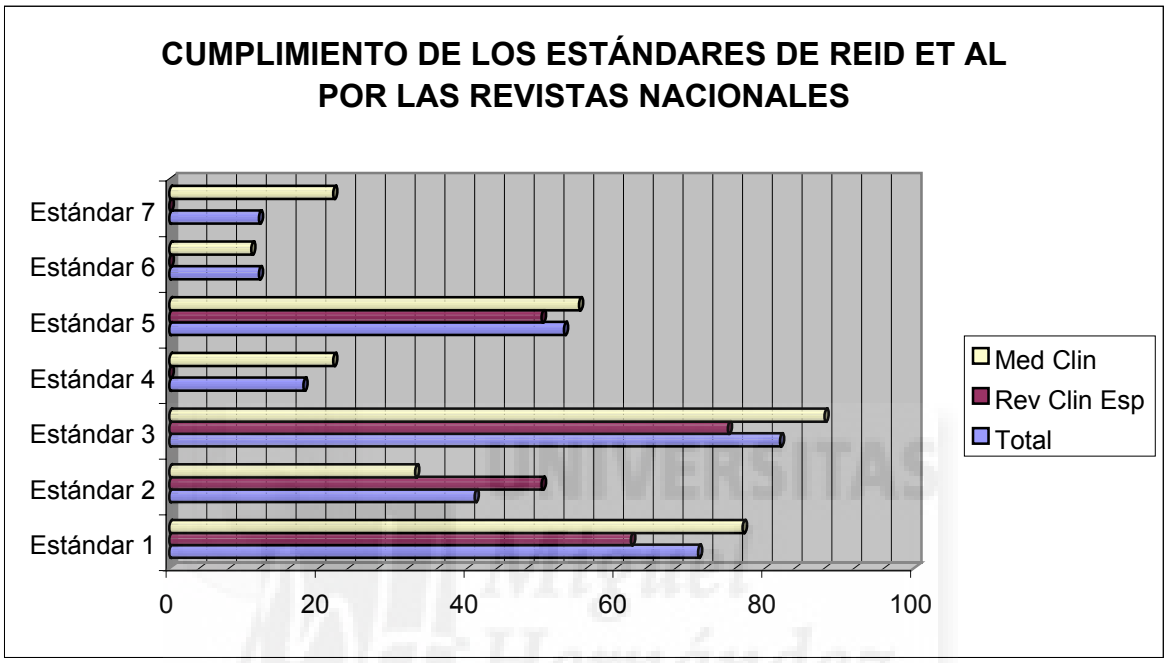
El estudio debe presentar los resultados indeterminados que ha obtenido, ya que la prueba puede tener una baja efectividad clínica si sus resultados no pueden ser interpretados. Pero este dato apenas se refleja en los trabajos analizados y en casi ninguno explican el por qué de su exclusión o inclusión.

De los diez trabajos en los que habría que haber medido la reproducibilidad de la prueba diagnóstica que se evalúa (tanto del instrumento como del observador), se ha estudiado en dos.

Tabla 8: Descripción del cumplimiento de los estándares propuestos por Reid y colaboradores [5] en los diecisiete artículos analizados de las revistas nacionales, Medicina Clínica (Barcelona) y Revista Clínica Española.

	TOTAL (17)	***Med Clin* (9)**	***Rev Clin Esp* (8)**
	(n/%)	**(n/%)**	**(n/%)**
1- Composición del espectro	**12 (71%)**	7 (77%)	5 (62%)
Distribución por edad	**12 (71%)**	7 (77%)	5 (62%)
Distribución por sexo	**13 (76%)**	7 (77%)	6 (75%)
Síntomas clínicos/estadios de la enfermedad	**10 (60%)**	6 (66%)	4 (50%)
Criterios selección y exclusión	**8 (47%)**	5 (55%)	3 (37%)
2- Análisis por subgrupos	**7 (41%)**	3 (33%)	4 (50%)
3- Prevención sesgo de secuencia	**14 (82%)**	8 (88%)	6 (75%)
4- Prevención sesgo de revisión	**3 (18%)**	2 (22%)	1 (13%)
5- Precisión de los resultados	**9 (53%)**	5 (55%)	4 (50%)
6- Resultados indeterminados	**2 (12%)**	1 (11%)	1 (13%)
7- Reproducibilidad	**2 (12%)**	2 (22%)	0

Figura 7: Representación gráfica del seguimiento de los estándares de Reid y colaboradores [5] por los 17 trabajos analizados de las revistas nacionales, Medicina Clínica (Barcelona) y Revista Clínica Española.



Nota:

Estándar 1 = Composición del espectro.

Estándar 2 = Análisis por subgrupos.

Estándar 3 = Prevención del sesgo de secuencia.

Estándar 4 = Prevención del sesgo de revisión.

Estándar 5 = Precisión de los resultados.

Estándar 6 = Resultados indeterminados.

Estándar 7 = Reproducibilidad.

En resumen y como se aprecia en la tabla 9, la media de criterios cumplidos para los 17 estudios evaluados en las dos revistas nacionales *Medicina Clínica (Barc) y Revista Clínica Española* es de 2,9 ítems.

Tabla 9: Media de cumplimiento por artículo de los criterios metodológicos de Reid y colaboradores [5] evaluados, para cada uno de los 17 artículos analizados de las revistas Medicina Clínica (Barc) y Revista Clínica Española.

	Artículos (n/%)	**Criterios/artículo (media/varianza)**
Med Clín (Barc)	9 (53%)	2,6 (2,6)
Rev Clin Esp	8 (47%)	3,2 (0,9)
TOTAL	**17 (100%)**	**2,9 (1,7)**

6.1.1.3- COMPARACIÓN ENTRE LOS RESULTADOS OBTENIDOS EN LAS DOS REVISTAS NACIONALES Y LOS DATOS DE LAS DOS REVISIONES NACIONALES PREVIAS *[7, 8]*.

Comparamos los resultados obtenidos tras analizar los artículos publicados en las dos revistas nacionales de medicina general *Medicina Clínica* y *Revista Clínica Española* durante los años 1997-2000, con los observados en la revisión de *Medicina Clínica*, años 1992-1995 [7] y de *Enfermedades Infecciosas y Microbiología Clínica,* años 1990-1996 [8]. Aunque ambas revistas aplicaron a los trabajos analizados los criterios establecidos por los autores, solo la primera aplicó también la guía metodológica de Reid y colaboradores [5].

Respecto al cumplimiento de los criterios propios, si observamos los resultados obtenidos por las dos revistas nacionales analizadas en el presente estudio durante los años 1997-2000 con los resultados obtenidos por las dos revisiones previas durante los años 1990-1996 [7, 8], aparecen

escasas mejoras estadísticamente significativas: tan solo en el aspecto referente a si el patrón de referencia incorpora o no en su definición pruebas a valorar, y la expresión de forma continua de los resultados susceptibles a ello. Otros parámetros, como el que define el origen de la población a estudio, empeora de manera significativa. (Tabla 10)

Tabla 10: Comparación del cumplimiento de los criterios metodológicos de los autores en los artículos publicados entre los años 1997-2000 en Medicina Clínica y Revista Clínica Española con los artículos publicados en Medicina Clínica durante los años 1992-1995 [7] y Enfermedades Infecciosas y Microbiología Clínica, años 1990-1996 [8].

	***Med Clin* (42) (1992-5) (n/%)**	**Enf Inf Mic Clín (45) (1990-6) (n/%)**	***Med Clin y Rev Clin Esp* (17) (1997-2000) (n/%)**
- Objetivos			
Especificados	38 (90%)	42 (93%)	17 (100%)
Justificación de la prueba	32 (76%)	21 (47%)	8 (47%)
- Patrón de referencia			
Especificado	38 (90%)	44 (98%)	15 (88%)
No incorpora la prueba a valorar	12 (32%)	4 (9%)	11 (65%)[a]
Realizado en toda la serie	30 (79%)	29 (66%)	12 (71%)
- Prueba diagnóstica			
Descripción suficiente	33 (79%)	38 (84%)	15 (88%)
Definición normalidad	19 (56%)	30 (71%)	14 (82%)
- Diseño del estudio			
Descripción del origen de la población	13 (31%)	4 (26%)	1 (6%)
- Resultados			
Expresión continua*	5 (33%)	1 (7%)	15 (88 %)[a]
Índice para pruebas conjuntas**	5 (33%)	4 (18%)	6 (35%)
Cálculo valores predictivos***	19 (59%)	21 (75%)[a]	10 (59%)

[a] Diferencias estadísticamente significativas (p<0,05) calculadas por Chi[2] entre la revisión actual y las anteriores.

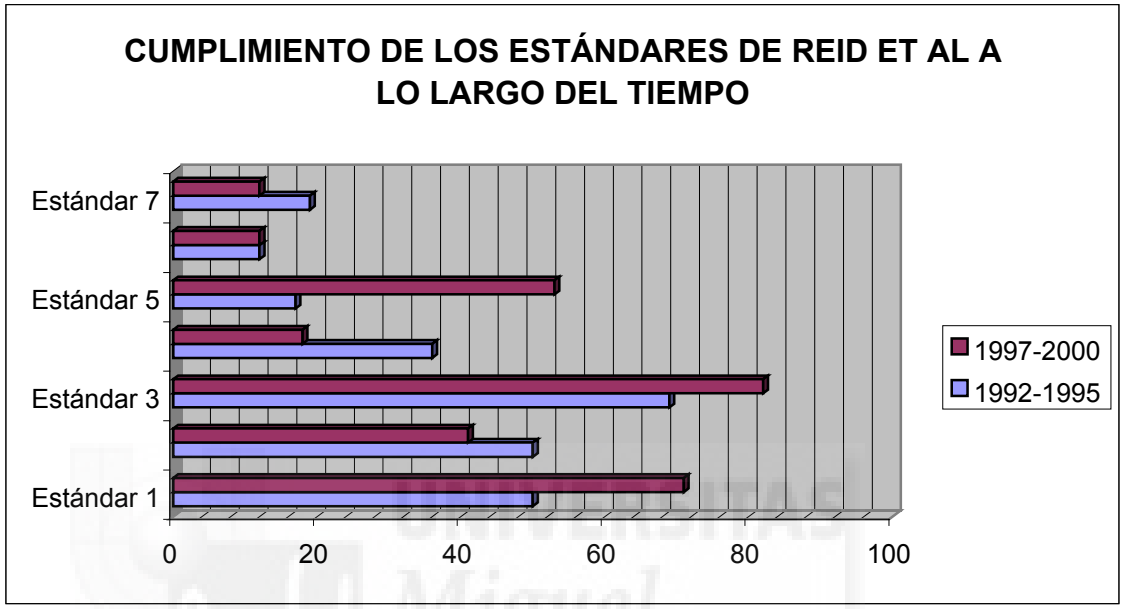
Si comparamos los resultados obtenidos por nuestra revisión tras aplicar los criterios metodológicos de Reid y colaboradores [5], con los obtenidos por la revisión efectuada en *Medicina Clínica* durante los años 1992-1995, ya que es la única que también utiliza estos criterios, observamos una ligera mejoría en la mayoría de los apartados, siendo solo significativa en el apartado referente al cálculo de la precisión de los resultados. Otros apartados como es el cálculo de la reproducibilidad de la técnica o la prevención del sesgo de revisión empeoran, aunque no de manera significativa. (Tabla 11 y figura 8).

Tabla 11: Comparación del cumplimiento de los criterios metodológicos de Reid y colaboradores [5] en los artículos publicados entre los años 1997-2000 en Medicina Clínica y en Revista Clínica Española con los artículos publicados en Medicina Clínica durante los años 1992-1995 [7].

	***Med Clin* (42) (1992-5) (n/%)**	***Med Clin* y *Rev Clin Esp*) (17) (1997-2000) (n/%)**
1- Composición del espectro	21 (50%)	12 (71%)
Distribución por edad	22 (52%)	12 (71%)
Distribución por sexo	28 (67%)	13 (76%)
Síntomas clínicos/estadios de la enfermedad	26 (62%)	10 (60%)
Criterios selección y exclusión	15 (36%)	8 (47%)
2- Análisis por subgrupos	21 (50%)	7 (41%)
3- Prevención sesgo de secuencia	29 (69%)	14 (82%)
4- Prevención sesgo de revisión	15 (36%)	3 (18%)
5- Precisión de los resultados	7 (17%)	9 (53%)[a]
6- Resultados indeterminados	5 (12%)	2 (12%)
7- Reproducibilidad	8 (19%)	2 (12%)

[a] *Diferencias estadísticamente significativas (p<0,05) calculadas por Chi^2 entre la revisión actual y la anterior.*

Figura 8: Comparación del cumplimiento de los criterios metodológicos de Reid y colaboradores [5] en los artículos publicados entre los años 1997-2000 en Medicina Clínica y Revista Clínica Española con los artículos publicados en Medicina Clínica [7] durante los años 1992-1995.



Nota:

Estándar 1 = Composición del espectro.

Estándar 2 = Análisis por subgrupos.

Estándar 3 = Prevención del sesgo de secuencia.

Estándar 4 = Prevención del sesgo de revisión.

Estándar 5 = Precisión de los resultados.

Estándar 6 = Resultados indeterminados.

Estándar 7 = Reproducibilidad.

Si comparamos el número medio de estándares metodológicos de Reid y colaboradores [5] por artículo que cumple la revisión actual, con el cumplimiento mostrado en la revista *Medicina Clínica* entre los años 1992-1995, vemos que no hay diferencias estadísticamente significativas: en la revisión actual obtenemos una media de criterios cumplidos de 2,9 frente al análisis previo 2,3. (Tabla 12)

Tabla 12: Relación de la media de los criterios metodológicos de Reid et al [5] cumplidos para cada artículo analizado de las revistas Medicina Clínica y Revista Clínica Española (años 1997-2000) en comparación con lo obtenido en la revisión previa de Medicina Clínica durante los años 1992-1995.

	Artículos (n/%)	**Criterios/artículo (media/varianza)**
Revisión actual (1997-2000)	17 (100%)	2,9 (1,7)
Med Clin (1992-5)	42 (100%)	2,3 (2,6)

Diferencias estadísticamente significativas ($p<0,05$) calculadas por Chi^2 entre la revisión actual y la anterior.

6.1.1.4- COMPARACIÓN ENTRE LOS RESULTADOS OBTENIDOS EN LAS DOS REVISTAS NACIONALES Y LOS DATOS DE LA REVISIÓN INTERNACIONAL PREVIA *[5].*

Comparamos los resultados obtenidos en la revisión de los artículos publicados durante los años 1997-2000 en *Medicina Clínica* y *Revista Clínica Española* después de aplicar los criterios metodológicos de Reid y colaboradores [5], con los datos obtenidos por la revisión internacional previa [5], en la que se instauraron dichos criterios. Como he comentado previamente, en esta revisión anterior se analizaron 34 artículos entre los años 1990 y 1993 de las revistas: *Journal of the American Medical Association*, *Lancet*, *British Medical Journal* y *New England Journal of Medicine*.

Como se observa en la tabla 13, se produce una mejora estadísticamente significativa de los artículos valorados actualmente frente a la revisión previa, en los criterios referentes a la composición del espectro, el análisis por subgrupos y el cálculo de la precisión de los resultados. En cambio, se observa una peor calidad en el criterio referente a la previsión del sesgo de revisión.

Tabla 13: Comparación entre los resultados obtenidos en la revisión actual de los artículos publicados en Medicina Clínica y Revista Clínica Española (años 1997-2000) tras aplicar los criterios de Reid y colaboradores, y los datos obtenidos en la revisión internacional previa [5] (años 1990-93).

	Reid y colaboradores (años 1990-3) (34*	**TOTAL (1997-2000) (17)**
1- Composición del espectro	11 (32%)	12 (71%)[a]
2- Análisis por subgrupos	4 (12%)	7 (41%)[a]
3- Prevención sesgo de secuencia	21 (62%)	14 (82%)
4- Prevención sesgo de revisión	16 (47%)[a]	3 (18%)
5- Precisión de los resultados	8 (24%)	9 (53%)[a]
6- Resultados indeterminados	13 (38%)	2 (12%)
7- Reproducibilidad	11 (32%)	2 (12%)

[a] *Diferencias estadísticamente significativas ($p<0,05$) calculadas por Chi^2 entre la revisión actual y la previa.*

* Incluye las revistas: *New England Journal of Medicine, Journal of the American Medical Association, British Medical Journal* y *Lancet*.

Si comparamos el número medio de ítems cumplidos entre los artículos revisados en el trabajo actual a escala nacional, y lo obtenido por Reid y colaboradores [5], comprobamos que no se ha producido una mejora entre ambos resultados. (Tabla 14).

Tabla 14: Relación de la media de los criterios metodológicos de Reid y colaboradores [5] cumplidos para cada artículo analizado de las revistas Medicina Clínica y Revista Clínica Española (años 1997-2000) en comparación con la revisión previa de Reid et al (años 1990-3).

	Artículos (n/%)	**Criterios/artículo (media/varianza)**
Revisión actual (1997-2000)	17 (100%)	2,9 (1,7)
Reid y colaboradores (1990-3)	34 (100%)	3,0 (0,9)

Diferencias estadísticamente significativas ($p<0,05$) calculadas por Chi^2 entre la revisión actual y la previa.

6.1.2- REVISTAS INTERNACIONALES: Clinical Chemistry y Clinical Chemistry and Laboratory Medicine.

En esta segunda parte de la valoración de revistas científicas nos centramos en dos publicaciones internacionales de ámbito exclusivo de laboratorio. Al igual que se ha hecho con las revistas nacionales, describimos el proceso llevado a cabo para la selección de los artículos finalmente a analizar; primero la búsqueda informática y luego manual.

Clinical Chemistry

- En primer lugar y mediante el sistema informático MEDLINE, delimitamos los años de estudio: en los años 1996, 2001 y 2002 se publicaron en esta revista un total de 440, 436 y 417 artículos respectivamente.
- Tras preseleccionar delimitando con las palabras clave "sensibilidad y especificidad", además de con otros términos que indiquen se ha evaluado la exactitud diagnóstica, obtenemos 139 estudios en el año 1996, 73 artículos en 2001 y 85 trabajos en 2002.
- Después de aplicar los criterios de inclusión y de exclusión, la búsqueda se restringe a 45 artículos: 10 publicados en 1996, 15 en 2001 y 20 en 2002. No analizamos los

trabajos que no son artículos originales, no tratan pruebas de laboratorio o no calculan sensibilidad, especificidad o ambas de la prueba diagnóstica.

- Realizamos después una búsqueda manual en los años a estudio, ya que como hemos reseñado con anterioridad, es la mejor estrategia de búsqueda. Además de los 45 artículos encontrados previamente, pudimos seleccionar 7 trabajos más: 1 correspondiente al año 1996, 2 al año 2001 y 4 al año 2002.
- En total, vamos a evaluar un total de 52 trabajos: 11 publicados en 1996, 17 en 2001 y 24 en 2002.

Clinical Chemistry and Laboratory Medicine (*European Journal of Clinical Chemistry and Clinical Biochemistry*).

- Durante ese período de tiempo, y mediante el sistema MEDLINE, preseleccionamos un total de 604 artículos: 162 en el año 1996, 210 en 2001 y 232 trabajos en 2002.
- De estos, al aplicar las palabras clave "sensibilidad y especificidad", junto con otros términos que denoten que se ha determinado la exactitud diagnóstica de la prueba, obtenemos 20 artículos correspondientes al año 1996, 39 al año 2001 y 40 al 2002.
- Al aplicar los criterios de inclusión y exclusión, el número de artículos se reduce a 22: 6 perteneciente al año 1996, 8 al 2001 y 8 al 2002. Excluimos el resto de estudios por no tratar pruebas de laboratorio, tener formato de carta dirigida al editor, revisión o nota clínica, o estudiar solo la sensibilidad o solo la especificidad de la técnica.
- Al igual que hemos hecho con la otra revista, realizamos una búsqueda manual. Además de los 20 trabajos seleccionados con la búsqueda informática, obtenemos 7 trabajos más: 1 perteneciente al año 1996, 2 al 2001 y 2 al año 2002.
- Es decir, cuando finalizamos ambos tipos de búsqueda, obtenemos un total de 27 trabajos: 7 correspondientes al año 1996, 10 al 2001 y 10 al 2002.

Comprobamos como de los procedimientos diagnósticos evaluados, es el área de bioquímica la más frecuentemente estudiada (39%), seguida de inmunología (32%) y microbiología (11%) (Tabla 15).

Tabla 15: Áreas de laboratorio a las que pertenecen los 79 artículos revisados de las revistas de ámbito internacional Clinical Chemistry y Clinical Chemistry and Laboratory Medicine.

	Total (n/%)	**Clin Chem (n/%)**	***Clin Chem and Lab Med* (n/%)**
Bioquímica	**31 (39%)**	23 (44%)	8 (30%)
Inmunología	**25 (32%)**	15 (29%)	10 (36%)
Microbiología	**9 (11%)**	4 (8%)	5 (19%)
Genética	**6 (8%)**	5 (9%)	1 (4%)
Hormonas	**5 (6%)**	3 (6%)	2 (7%)
Hematología	**2 (3%)**	1 (2%)	1 (4%)
Farmacocinética	**1 (1%)**	1 (2%)	-
TOTAL	**79 (100%)**	**52 (100%)**	**27 (100%)**

Aplicamos a una muestra de 34 artículos de estas revistas internacionales, la guía metodológica que combina los criterios de Reid y colaboradores [5]y los criterios propios [7, 8]. Al resto de los 45 trabajos evaluados, únicamente aplicamos los criterios metodológicos de Reid y colaboradores [5]. Pero, al igual que en la revisión previa de las revistas nacionales, los resultados obtenidos para cada guía de calidad los comento de manera separada.

6.1.2.1- RESULTADOS OBTENIDOS TRAS LA APLICACIÓN DE LOS CRITERIOS PROPIOS [7, 8].

(Tabla 16)

En una muestra de 34 artículos, del total de los 79 estudiados en el ámbito internacional, aplicamos la plantilla de criterios propios de Hernández y García [7, 8]. En concreto a 19 de la revista *Clinical Chemistry* (9 del año 1996 y 10 publicados durante el año 2001) y 15 de *Clinical Chemistry Laboratory Medicine* (5 publicados en el año 1996, entonces *European Journal of Clinical Chemistry Clinical Biochemistry* y 10 durante el año 2001).

En esta muestra analizada de los trabajos internacionales, se puede ver que tanto la especificación de los objetivos del estudio como la justificación de la investigación, sobre todo a partir del año 2001, aparece ampliamente reflejada en los artículos. Debido al incremento de nuevas pruebas diagnósticas de laboratorio y con ello la necesidad de su estudio, se deben realizar únicamente aquellas valoraciones que estén justificadas.

Es importante especificar claramente el patrón de referencia ya que, en los estudios sobre pruebas diagnósticas, se precisa una prueba estándar que decida finalmente si la enfermedad está presente o no. En estos trabajos, aunque aparece definido en un porcentaje apropiado, en solo cerca de la mitad de los trabajos evaluados se ha realizado a todos los participantes del estudio.

En lo referente al método de realización de la prueba diagnóstica que se está valorando, el procedimiento aparece explicado en todos los estudios analizados durante ambos períodos de tiempo, así como definido qué entienden los autores por un valor normal.

El aspecto del diseño del estudio relativo a la descripción del origen de la población a estudio apenas está descrito en estos estudios internacionales. Sobre todo llama la atención la ausencia de información relativa al criterio que se ha seguido para seleccionar finalmente la muestra.

Respecto a la comunicación de los resultados de la valoración de la prueba diagnóstica, destaca que la mayoría de los artículos que son susceptibles de expresar sus resultados de forma continua o discreta en más de dos categorías, los presentan en forma de curva ROC, cociente de verosimilitud o calculaba la sensibilidad y especificidad en distintos puntos de corte.

Apenas aparece reflejado algún tipo de índice para los resultados en los trabajos que evalúan varias pruebas con mismo objetivo diagnóstico, ni se calcula el valor predictivo de la muestra. Hay que decir en este punto, que en alguno de los trabajos no era correcto este último cálculo, y sin embargo aparecía, lo que puede dar lugar a confusiones, ya que solo se debe calcular en aquellos estudios en los que la muestra represente a la población clínica real.

Encontramos diferencias estadísticamente significativas ($p<0,05$) entre el año 1996 y el año 2001 para el criterio referente a si se ha justificado o no correctamente la valoración de la prueba (un mayor cumplimiento en el año 2001), y para el criterio que valora si se ha calculado un índice para pruebas conjuntas en el caso de que se valoren dos o más tests diagnósticos (este parámetro se cumple peor en el año 2001).

Tabla 16: Descripción del cumplimiento de los estándares propuestos por los autores en los treinta y cuatro artículos analizados durante los años 1996 y 2001 en las revistas Clinical Chemistry y Clinical Chemistry Laboratory Medicine (European Journal of Clinical Chemistry Clinical Biochemistry):

	AÑO 1996			AÑO 2001			
	TOTAL (n/%) (14)	**Clin Chem (n/%) (9)**	**Clin Chem Lab Med (n/%) (5)**	**TOTAL (n/%) (20)**	**Clin Chem (n/%) (10)**	**Clin Chem Lab Med (n/%) (10)**	**p+**
- Objetivos							
Especificados	**11 (79%)**	7 (78%)	4 (80%)	**18 (90%)**	8 (80%)	10 (100%)	0,24
Justificación de la prueba	**8 (57%)**	5 (56%)	3 (60%)	**17 (85%)**	7 (70%)	10 (100%)	<0,05
- Patrón de referencia							
Especificado	**11 (79%)**	8 (89%)	3 (60%)	**17 (85%)**	8 (80%)	9 (90%)	0,93
No incorpora la prueba a valorar	**13 (93%)**	8 (89%)	5 (100%)	**19 (95%)**	10 (100%)	9 (90%)	0,99
Realizado en toda la serie	**6 (43%)**	4 (44%)	2 (40%)	**10 (50%)**	4 (40%)	6 (60%)	0,32
- Prueba diagnóstica							
Descripción suficiente	**14 (100%)**	9 (100%)	5 (100%)	**20 (100%)**	10 (100%)	10 (100%)	0,99
Definición normalidad	**14 (100%)**	9 (100%)	5 (100%)	**18 (90%)**	10 (100%)	8 (80%)	0,77
- Diseño del estudio							
Descripción del origen de la población	**2 (14%)**	1 (11%)	1 (20%)	**4 (20%)**	1(10%)	3 (30%)	0,77
- Resultados							
Susceptibles de expresión continua *	**9 (64%)**	7 (78%)	2 (40%)	**14 (70%)**	7 (70%)	7 (70%)	0,99
Indice para pruebas conjuntas **	**8 (57%)**	6 (67%)	2 (40%)	**7 (35%)**	7(70%)	0	<0,05
Cálculo valores predictivos***	**9 (64%)**	5 (56%)	4 (80%)	**18 (90%)**	10 (100%)	8 (80%)	0,24

* Los resultados se expresan de forma continua en forma de curvas ROC, coeficiente de verosimilitud o calculaba la sensibilidad y especificidad en distintos puntos de corte cuando son susceptibles. También es correcto que no se expresen así, si no son susceptibles a ello.
** Para estudios que analizan más de una prueba con el mismo objetivo. También se cumple si solo se analiza una prueba, y por lo tanto no se calcula dicho índice.
*** Cuando es correcto realizarlo. Cuando no es pertinente su cálculo, también se valora el que no se determine.
+Valor de p calculado por Chi^2 comparando proporción total de cumplimiento entre 1996 y 2001.

Si calculamos el número medio de criterios metodológicos propuestos por los autores cumplidos por los 34 artículos evaluados de las dos revistas internacionales, no obtenemos ninguna diferencia estadísticamente significativa entre los dos años de estudio 1996 y 2001 (p=0,46), ni para cada revista a lo largo del tiempo: *Clinical Chemistry* entre 1996 y 2001 p=0,44; *Clinical Chemistry and Laboratory Medicine* entre 1996 y 2001 p=0,51. (Tabla 17).

Tabla 17: Promedio de la suma de los criterios metodológicos propios [7, 8] cumplidos, para los 34 artículos analizados de las revistas Clinical Chemistry y Clinical Chemistry and Laboratory Medicine durante los años 1996 y 2001.

	Nº artículos (n/%)	**Criterios/artículo (media/varianza)**
AÑO 1996	**14 (42%)**	**5,7 (2,6)**
Clin Chem	9 (26%)	6,0 (1,3)
Clin Chem Lab Med	5 (16%)	5,5 (1,4)
AÑO 2001	**20 (58%)**	**6,1 (2,2)**
Clin Chem	10 (29%)	6,5 (2,5)
Clin Chem Lab Med	10 (29%)	4,8 (4,7)
TOTAL	**34 (100%)**	**6,0 (2,1)**

Diferencias estadísticamente significativas (p<0,05) calculadas por Chi^2 entre los dos años de estudio y las dos revistas internacionales

6.1.2.2- RESULTADOS OBTENIDOS TRAS LA APLICACIÓN DE LOS CRITERIOS DE REID Y COLABORADORES [5].

Por último, aplicamos los criterios de Reid y colaboradores [5] a los 79 artículos analizados de las dos publicaciones internacionales; 52 pertenecientes a *Clinical Chemistry* y 27 a *Clinical Chemistry Laboratory Medicine*.

El rango de cumplimiento de los criterios metodológicos está comprendido entre 0% y 83% para los estudios publicados en 1996, del 15% al 81% para los estudios publicados en 2001, y entre el 6% y el 82% para los estudios publicados en 2002 (Tabla 18 y Figura 9).

Los artículos publicados en el 2002 muestran un cumplimiento mayor de los criterios aplicados que los publicados en 1996 o 2001, excepto en dos de ellos: la presentación del cálculo de la reproducibilidad de la prueba, que aparece en la mayoría de los trabajos en el año 1996 y en el año 2001 aunque disminuye en el 2002, y la valoración de la exactitud en subgrupos pertinentes, que aunque mejora en el año 2002 respecto a lo obtenido en el 2001, no supera el cumplimiento obtenido en el año 1996.

Entre los criterios que han mejorado a partir del año 2002 se encuentra el cálculo de la precisión estadística de los índices y la composición del espectro. La presencia o no de resultados indeterminados apenas aparece reflejada en los estudios analizados. El número medio de pacientes / muestras de los artículos es de 331 (rango: 10-2.971 pacientes / muestras).

Se encontraron diferencias estadísticamente significativas entre el año 1996 y el año 2001, para los criterios metodológicos referentes a la composición del espectro, al análisis por subgrupos, a la prevención de los sesgos de secuencia y revisión, al cálculo de la precisión de los resultados y a la presentación de resultados indeterminados. La mejora de la metodología entre los años 2001 y 2002 fue estadísticamente significativa en el criterio que trata de la composición del espectro, en la prevención del sesgo de secuencia y en la determinación de la precisión de los resultados; la

presentación o no de los resultados indeterminados, empeoró en el año 2002 con respecto al año 2001, de forma significativa.

Tabla 18: Descripción del cumplimiento de los estándares propuestos por Reid y colaboradores [5] en los setenta y nueve artículos analizados de las dos revistas internacionales, Clinical Chemistry y Clinical Chemistry Laboratory Medicine durante los años 1996, 2001 y 2002.

	Año 1996 (18) (n/%)	**Año 2001 (27) (n/%)**	**Año 2002 (34) (n/%)**
1- Composición del espectro	4 (22%)[ab]	10 (37%)[c]	24 (71%)
Distribución por edad	9 (50%)	18 (67%)	28 (82%)
Distribución por sexo	9 (50%)	19 (70%)	26 (76%)
Síntomas clínicos/estadios de la enfermedad	6 (33%)	11 (41%)	12 (35%)
Criterios selección y exclusión	4 (22%)	10 (37%)	18 (53%)
2- Análisis por subgrupos	8 (44%)[a]	7 (26%)	13 (38%)
3- Prevención sesgo de secuencia	6(33%)[ab]	13 (48%)[c]	24 (71%)
4- Prevención sesgo de revisión	4 (22%)[ab]	10 (37%)	15 (44%)
5- Precisión de los resultados	4 (22%)[ab]	12 (44%)[c]	22 (65%)
6- Resultados indeterminados	0 [ab]	4 (15%)[c]	2 (6%)
7- Reproducibilidad	15 (83%)	22 (81%)	21 (68%)

Cálculos efectuados mediante Chi^2:
[a] *Diferencias estadísticamente significativas (p<0,05) entre 1996 y 2001.*
[b] *Diferencias estadísticamente significativas (p<0,05) entre 1996 y 2002.*
[c] *Diferencias estadísticamente significativas (p<0,05) entre 2001 y 2002.*

Figura 9: Representación gráfica del seguimiento de los estándares de Reid y colaboradores [5] por los 79 trabajos analizados de las revistas internacionales, Clinical Chemistry y Clinical Chemistry and Laboratory Medicine.

Nota:

Estándar 1 = Composición del espectro.

Estándar 2 = Análisis por subgrupos.

Estándar 3 = Prevención del sesgo de secuencia.

Estándar 4 = Prevención del sesgo de revisión.

Estándar 5 = Precisión de los resultados.

Estándar 6 = Resultados indeterminados.

Estándar 7 = Reproducibilidad.

Si calculamos el número medio de criterios metodológicos de Reid y colaboradores [5] cumplidos por los 79 artículos evaluados de las dos revistas internacionales, obtenemos un valor de 2,6 en 1996, de 2,5 en el año 2001 y 3,5 en 2002 (p <0,05). Para *Clinical Chemistry*, los valores son 2,5 en 1996, 2,8 en 2001 y 4,1 en 2002 (p <0,05); para *Clinical Chemistry and Laboratory Medicine* 2,7 en 1996, 2,4 en 2001 y 1,9 en 2002 (p = 0,31).

Tabla 19: Relación de la suma de los criterios metodológicos de Reid y colaboradores [5] cumplidos, para los 79 artículos analizados de las revistas Clinical Chemistry y Clinical Chemistry and Laboratory Medicine durante los años 1996, 2001 y 2002.

	Nº artículos (n/%)	**Criterios/artículo (media/varianza)**
AÑO 1996	**18 (23%)**	**2,6 (2,4)**
Clin Chem	11 (14%)	2,5 (2,5)
Clin Chem Lab Med	7 (9%)	2,9 (2,5)
AÑO 2001	**27 (34%)**	**2,5 (3,1)**
Clin Chem	17 (22%)	2,8 (3,5)
Clin Chem Lab Med	10 (12%)	2,1 (2,3)
AÑO 2002	**34 (43%)**	**3,5 (2,4)[a]**
Clin Chem	24 (31%)	4,1 (1,6)[a]
Clin Chem Lab Med	10 (12%)	1,9 (0,9)
TOTAL	**79 (100%)**	**2,9 (2,6)**

[a]Diferencias estadísticamente significativas (p<0,05) calculadas por Chi^2 entre la revisión actual y la previa.

6.1.2.3- COMPARACIÓN ENTRE LOS RESULTADOS OBTENIDOS EN LA EVALUACIÓN DE LOS ARTÍCULOS PUBLICADOS EN LAS DOS REVISTAS INTERNACIONALES, Y LOS PRESENTADOS POR LA REVISIÓN PREVIA DE REID Y COLABORADORES [5].

Comparamos los resultados obtenidos tras analizar los artículos publicados en las dos revistas internacionales de laboratorio *Clinical Chemistry* y *Clinical Chemistry and Laboratory Medicine* durante los años 1996, 2001 y 2002 con los observados por Reid y colaboradores [5] en una muestra de los estudios publicados entre 1990 y 1993 por un grupo de publicaciones clínicas relevantes: *New England Journal of Medicine, Journal of the American Medical Association, British Medical Journal* y *Lancet* (tabla 20 y figura 10).

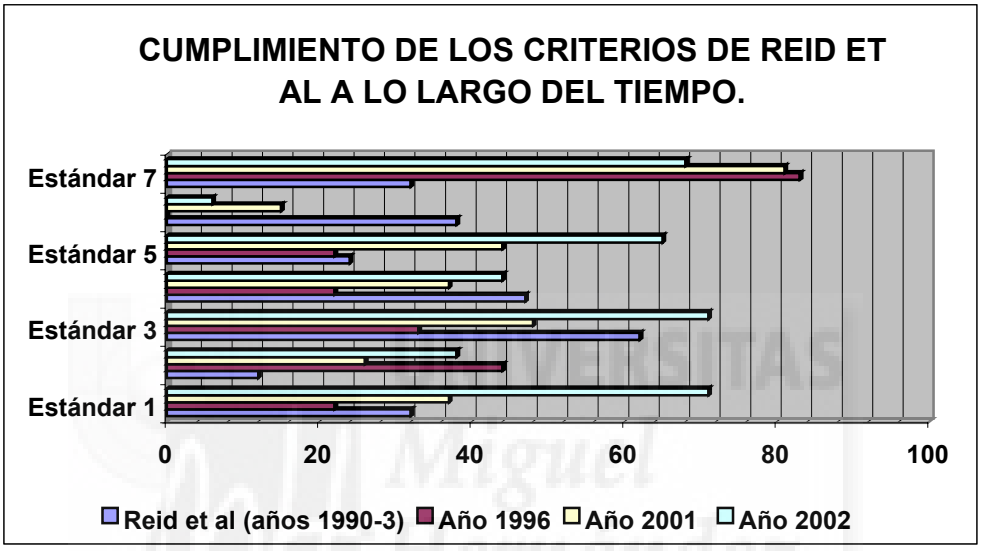
Podemos observar algunas mejoras estadísticamente significativas: los criterios referentes al análisis de los parámetros de exactitud diagnóstica y cálculo de la reproducibilidad, ya mejoran a partir del año 1996; el criterio que define el cálculo de la precisión de los resultados, aumenta su cumplimiento a partir del año 2001, mientras que no es hasta el año 2002 cuando aumenta de manera significativa el parámetro que define la composición del espectro participante en el estudio.

Tabla 20: Comparación del cumplimiento de los criterios metodológicos de Reid y colaboradores [5] en los artículos publicados en 1996, 2001 y 2002 en Clinical Chemistry y Clinical Chemistry and Laboratory Medicine con los artículos publicados en revistas de medicina general durante los años 1990-1993.

	Reid (años 1990-3) (34)* (n/%)	**Año 1996 (18) (n/%)**	**Año 2001 (27) (n/%)**	**Año 2002 (34) (n/%)**
1- Composición del espectro	11 (32%)	4 (22%)	10 (37%)	24 (71%)[a]
2- Análisis por subgrupos	4 (12%)	8 (44%)[a]	7 (26%)[a]	13 (38%)[a]
3- Prevención sesgo de secuencia	21 (62%)	6 (33%)[a]	13 (48%)	24 (71%)
4- Prevención sesgo de revisión	16 (47%)	4 (22%)[a]	10 (37%)	15 (44%)
5- Precisión de los resultados	8 (24%)	4 (22%)	12 (44%)[a]	22 (65%)[a]
6- Resultados indeterminados	13 (38%)	0[a]	4 (15%)[a]	2 (6%)[a]
7- Reproducibilidad	11 (32%)	15 (83%)[a]	22 (81%)[a]	21 (68%)[a]

[a]*Diferencias estadísticamente significativas (p<0,05) calculadas por Chi² entre la revisión actual y la realizada por Reid y colaboradores [21] durante los años 1990-1993.*
* Incluye las revistas: *New England Journal of Medicine, Journal of the American Medical Association, British Medical Journal* y *Lancet*.

Figura 10: Representación gráfica de la comparación del cumplimiento de los criterios metodológicos de Reid y colaboradores [5] en los artículos publicados en 1996, 2001 y 2002 en Clinical Chemistry y Clinical Chemistry and Laboratory Medicine, con los artículos publicados en revistas de medicina general durante los años 1990-1993 en las revistas New England Journal of Medicine, Journal of the American Medical Association, British Medical Journal y *Lancet*.



Nota:

Estándar 1 = Composición del espectro.

Estándar 2 = Análisis por subgrupos.

Estándar 3 = Prevención del sesgo de secuencia.

Estándar 4 = Prevención del sesgo de revisión.

Estándar 5 = Precisión de los resultados.

Estándar 6 = Resultados indeterminados.

Estándar 7 = Reproducibilidad.

Por último, comparamos la media de cumplimiento de los criterios metodológicos de Reid et al entre los artículos evaluados de *Clinical Chemistry* y *Clinical Chemistry and Laboratory Medicine* durante los años 1996, 2001 y 2002, y los resultados obtenidos por Reid y colaboradores [5] en su revisión de 34 artículos de *New England Journal of Medicine, Journal of the American Medical Association, British Medical Journal* y *Lancet* durante los años 1990-3.

Como se observa en la tabla 21, se produce una mejora significativa del cumplimiento medio de los criterios de Reid y colaboradores, a partir del año 2002 ($p<0,05$), respecto a la revisión anterior de Reid et al [5].

Tabla 21: Comparación del número medio de criterios cumplidos entre los artículos evaluados de Clinical Chemistry y Clinical Chemistry and Laboratory Medicine durante los años 1996, 2001 y 2002, y lo obtenido por Reid et al [5] durante los años 1990-3.

	Nº artículos (n/%)	**Criterios/artículo (media/varianza)**
AÑO 1996	18 (100%)	2,6 (2,4)
AÑO 2001	27 (100%)	2,5 (3,1)
AÑO 2002	34 (100%)	3,5 (2,4)[a]
Reid et al [5]	34 (100%)	3,0 (0,9)

[a] *Diferencias estadísticamente significativas ($p<0,05$) calculadas por Chi^2 respecto al trabajo realizado por Reid y colaboradores [5] durante los años 1990-1993.*

6.1.3- ESTUDIO DE LA INFLUENCIA DE OTRAS VARIABLES EN LA CALIDAD METODOLÓGICA DE LOS ARTÍCULOS ANALIZADOS.

En este análisis queremos comprobar si determinadas características de los 96 artículos analizados (tanto de las dos revistas nacionales, como de las dos internacionales) como son: la influencia del país donde se realiza el estudio, el número de centros que han participado en la investigación, si el ámbito de los centros participantes es clínico, de laboratorio o ambos, el año de ejecución y la revista en la que ha aparecido publicado, influyen en una mejor calidad metodológica. Analizamos de manera separada los artículos en los que se han aplicado las recomendaciones de Reid y colaboradores [5], de los que se les ha aplicado la guía propia con los 11 criterios metodológicos [7, 8].

Evaluamos en primer lugar, si influye el que los estudios se hayan realizado en un determinado país o el número de centros que han participado en la evaluación de la prueba diagnóstica. No observamos diferencias estadísticamente significativas en función del número de centros que participa en el estudio, ni con el país de origen del centro investigador. (Tabla 22).

Tabla 22: Análisis del cumplimiento de las recomendaciones de Reid y colaboradores [5] y las de los autores [7, 8] por los 96 artículos analizados, en función del número de centros participante en el estudio o del país al que pertenecen dichos centros.

	REID Y COLABORADORES [5]		**CRITERIOS PROPIOS [7, 8]**	
	Artículos (n/%)	**Criterios/artículo**	**Artículos (n/%)**	**Criterios/artículo**
Centros				
- 1	14 (15%)	2,9	6 (12%)	6,5
- 2	39 (40%)	2,6	25 (50%)	6,0
- 3	21 (22%)	3,4	10 (19%)	6,4
- >4	22 (23%)	3,0	10 (19%)	6,0
País				
- EEUU + Europa	2 (2%)	5,0	1 (2%)	9,0
- EEUU	9 (9%)	2,9	4 (8%)	6,8
- Europa	73 (77%)	2,9	41 (80%)	6,0
- Asia	7 (7%)	2,7	2 (4%)	6,0
- Otros*	5 (5%)	1,6	3 (6%)	6,0

Diferencias estadísticamente significativas ($p<0,05$) calculadas por Chi^2.
** Otros: Canadá, Méjico + Europa, Sudamérica, África y Australia.*

El siguiente paso consistió en verificar si el cumplimiento global de los criterios estaba relacionado o no con el carácter general o de laboratorio de las revistas estudiadas, o con el ámbito en el que se realizaba el estudio (clínico, de laboratorio o en colaboración de ambos).

Podemos comprobar como no se observan diferencias estadísticamente significativas a nivel global entre los artículos publicados en revistas clínicas generales y en aquellas que son exclusivas de laboratorio, tanto para los criterios de Reid y colaboradores (p=0,963), como para los criterios propios (p=0,054), aunque en este caso roza la significación. Aunque no se observan diferencias estadísticamente significativas entre aquellos estudios que se realizan en ámbitos clínicos con los realizados en laboratorio o en combinación de ambos para los criterios de Reid y colaboradores (p=0,207), si que se advierten para los criterios metodológicos propios (p<0,005) (Tabla 23).

Tabla 23: Seguimiento de los criterios de Reid y colaboradores [5] y de los ítems de la guía propia [7, 8], por los artículos de las revistas analizadas, y clasificadas en función de su carácter clínico general (Medicina Clínica (Barcelona) y Revista Clínica Española) o exclusivo de laboratorio (Clinical Chemistry y Clinical Chemistry Laboratory Medicine) y en función del ámbito del estudio (clínico, de laboratorio, o en asociación de ambos).

	REID Y COLABORADORES [5]		**CRITERIOS PROPIOS [7, 8]**	
	Artículos (n/%)	**Criterios/artículo (media/varianza)**	**Artículos (n/%)**	**Criterios/artículo (media/varianza)**
REVISTA				
LABORATORIO	79 (82%)	2,9 (2,8)	34 (67%)	5,9 (2,3)
GENERAL	17 (18%)	2,9 (1,7)	17 (33%)	6,7 (1,4)
ÁMBITO				
CLÍNICO	14 (15%)	3,5 (1,8)	7 (14%)	7,3 (1,6)[a]
LABORATORIO	32 (33%)	2,6 (3,1)	21 (41%)	5,8 (1,9)
AMBOS	50 (52%)	3,0 (2,5)	23 (45%)	6,1 (2,1)

[a] *Diferencias estadísticamente significativas (p<0,05) calculadas mediante Chi^2.*

Una vez comprobado el cumplimiento general de los criterios metodológicos en función del tipo de publicación donde aparecen los trabajos o del ámbito de realización del estudio, analizamos si se observan diferencias en función de estas características, pero para cada criterio metodológico analizado de forma individual.

Evaluamos primero los criterios de Reid y colaboradores [5] en función del ámbito del estudio y del tipo de revista en el que aparece publicado el trabajo (tabla 24). Destaca el mejor cumplimiento del criterio referente a la descripción de la composición del espectro en aquellos trabajos realizados en ámbitos meramente clínicos o en colaboración de estos con laboratorios. Este criterio, también muestra mejor calidad en las revistas de carácter clínico general, que en aquellas exclusivas de laboratorio. También llama la atención el mayor cálculo de la reproducibilidad de la técnica en aquellos trabajos realizados en laboratorios, y una mayor prevención del sesgo de secuencia en revistas de carácter clínico general.

Tabla 24: Relación entre el cumplimiento de cada criterio individual de Reid et al [5] con el ámbito en el que se realizó el estudio de la prueba diagnóstica (clínico o universitario, de laboratorio o ambos) y el carácter de cada revista, general o de laboratorio para los 96 artículos analizados.

	ÁMBITO			**REVISTA**	
	Clínico (14)	**Laboratorio (32)**	**Ambos (50)**	**General (17)**	**Laboratorio (79)**
- Composición del espectro	9 (64%)	9 (28%)[a]	30 (60%)	11 (79%)	37 (47%)[a]
- Análisis por subgrupos	5 (36%)	10 (31%)	18 (36%)	7 (41%)	26 (33%)
- Prevención sesgo de secuencia	8 (57%)	16 (50%)	31 (62%)	14 (82%)	41 (52%)[a]
- Prevención sesgo de revisión	4 (29%)	8 (25%)	20 (40%)	3 (18%)	29 (40%)
- Precisión de los resultados	12 (86%)	14 (44%)[a]	24 (48%)[a]	9 (53%)	40 (51%)
- Resultados indeterminados	1 (7%)	3 (9%)	2 (4%)	1 (6%)	5 (6%)
- Reproducibilidad	8 (57%)	24 (75%)[a]	26 (52%)	5 (85%)	54 (68%)

[a]Diferencias estadísticamente significativas ($p<0,05$) calculadas mediante Chi^2.

En relación con los criterios propuestos por nosotros (tabla 25), destaca un mayor cumplimiento en las revistas de carácter exclusivo de laboratorio del criterio que constata que el patrón de referencia no incluye la prueba a valorar en su definición, y un mayor cálculo de los valores predictivos en los artículos que se publican en revistas de carácter general. Respecto al ámbito de realización del estudio, llama la atención una mejor definición del origen de los sujetos a estudio en aquellos trabajos realizados en ámbito clínico.

Tabla 25: Relación entre el cumplimiento de cada criterio individual de la guía de criterios propia [7, 8] con el ámbito en el que se realizó el estudio de la prueba diagnóstica (clínico o universitario, de laboratorio o ambos) y el carácter de cada revista, general o de laboratorio.

	ÁMBITO			REVISTA	
	Clínico (7)	**Laboratorio (21)**	**Ambos (23)**	**General (17)**	**Laboratorio (34)**
- Objetivos especificados	6 (86%)	18 (86%)	22 (96%)	17 (100%)	29 (85%)
- Justificación de la prueba	5 (71%)	12 (57%)	16 (70%)	8 (47%)	25 (74%)
- PR especificado	7 (100%)	18 (86%)	18 (78%)	15 (88%)	28 (82%)
- Incorporación de la prueba a valorar	5 (71%)	17 (81%)	22 (96%)	5 (29%)	32 (94%)[a]
- Realizado en toda la serie	6 (86%)	12 (57%)	9 (39%)	12 (71%)	15 (44%)
- PD: descripción suficiente	7 (100%)	21 (100%)	20 (91%)	15 (88%)	33 (97%)
- Definición normalidad	6 (86%)	19 (90%)	21 (92%)	14 (82%)	32 (94%)
- Origen de la población	3 (43%)	3 (14%)[a]	3 (13%)[a]	3 (18%)	6 (18%)
- Susceptibles de expresión continua*	4 (57%)	13 (62%)	17 (74%)	11 (65%)	23 (68%)
- Índice pruebas conjuntas**	3 (43%)	9 (43%)	9(39%)	6 (35%)	15 (44%)
- Cálculo VP	5 (71%)	7 (33%)	12 (52%)	14 (82%)	10 (29%)[a]

* Susceptible de expresión continua en forma Curvas ROC, coeficiente de verosimilitud o calculaba la sensibilidad y especificidad en distintos puntos de corte.
** Para estudios que analizan más de una prueba con el mismo objetivo.
[a]Diferencias estadísticamente significativas ($p<0,05$) calculadas mediante Chi^2.

6.1.4- APLICABILIDAD DE LAS GUÍAS METODOLÓGICAS EMPLEADAS EN LA EVALUACIÓN CRÍTICA DE ARTÍCULOS QUE VALORAN PRUEBAS DIAGNÓSTICAS DE LABORATORIO.

Calculamos el porcentaje de concordancia entre los tres examinadores: el porcentaje total de concordancia simple fue del 85,9 %; en los criterios de Reid y colaboradores [5] se obtuvo un 82,9 % de acuerdo simple y en los criterios propios [7, 8] un 88,8%.

Del total de los 18 criterios aplicados de las dos guías metodológicas, hubo un mayor índice de discordancia en siete de ellos, tres de la guía de recomendaciones propia y cuatro del listado de Reid y colaboradores [5]. Son los ítems con más dificultad en su aplicación y por lo tanto los que deberían ser definidos explícitamente. Calculamos para cada uno de los 18 criterios estudiados, la discrepancia obtenida entre los investigadores y reflejada como número absoluto de discrepancias y porcentaje de desacuerdo obtenido en cada criterio. También se calcula el porcentaje que ocupa en el total de las disconformidades (tabla 26).

A continuación detallo los criterios en los que hemos encontrado las principales dificultades, con ejemplos que ilustran los problemas que obtuvimos:

- Justificación de la prueba: En este criterio obtenemos el porcentaje mayor de discordancias, un 33%. Recordemos que el criterio se cumplía si se señalaba alguna de estas situaciones: la prueba en estudio es nueva, es una prueba de uso anterior con aplicación novedosa o una antigua con controversias en su aplicación. Si no es nueva, debía apoyarse en referencias bibliográficas indicando explícitamente qué aspectos no resueltos van a estudiarse, y si lo es, debe decirse claramente que no hay nada escrito sobre ese tema. El problema lo encontramos fundamentalmente cuando en el trabajo que analizamos, encontramos una frase que se puede tomar como explicación de un estudio nuevo, cuando no es así.

  Esto se refleja por ejemplo, en un artículo que determina la exactitud de los anticuerpos anticitrulina en el diagnóstico de artritis reumatoide [96]. En el apartado de introducción del trabajo los autores exponen que, "*En este estudio evaluamos las*

características diagnósticas y analíticas de un nuevo Elisa comercial para la detección de anticuerpos anti.CCP", pero en el apartado de referencias bibliográficas, vemos que ya hay trabajos previos sobre este kit comercial. El principal problema que se plantea en este caso es, que uno de los observadores considera el criterio como válido, ya que los autores han descrito la prueba como nueva y por lo tanto susceptible de investigación, pero al haber datos de trabajos previos en el apartado de referencias bibliográficas la realización de esta investigación no está del todo justificada y por lo tanto, otros observadores no lo dan como válido. Finalmente, decidimos no darle validez al criterio ya que al tratarse de una prueba ya estudiada los autores tendrían que haber expuesto los trabajos previos realizados y las deficiencias que ahora querían analizar de dicha prueba.

- Especificación del patrón de referencia: En este criterio alcanzamos un 16% de discordancias. Como he comentado previamente el patrón de referencia aparece especificado solamente si se explica claramente en qué consiste su aplicación, tanto en los sujetos como en el grupo de comparación. Aunque por el contexto y las pruebas realizadas en el trabajo se puede deducir en qué consiste el patrón de referencia, es necesario para dar válido el criterio su explicación con detalle. La principal discrepancia que obtuvimos en este punto fue que alguno de los observadores era menos estricto en la aplicación del criterio dando por válida una explicación incompleta; otros observadores por el contrario, solo le daban credibilidad si se especificaba claramente la definición del patrón de referencia.

  Encontramos el siguiente caso en un artículo que analiza un ensayo inmunoradiométrico de alta sensibilidad para tiroglobulina en suero con mínima interferencia de anticuerpos [93]. En el apartado referente a material y métodos, en la sección que trata la obtención de muestras sanguíneas los autores exponen: "*Los valores de tiroglobulina fueron determinados para 209 sujetos considerados normales basándose en sus valores de tiroglobulina y T4 y en la investigación clínica*". Algunos observadores consideraron que con esta frase se especificaba el patrón de referencia en la población a estudio; otro observador consideró que no explicaba a partir de qué valores de T4 y tiroglobulina se consideraban los

sujetos sanos o no. Finalmente, decidimos ser estrictos en la aplicación del criterio tal como había sido el tercer observador y no dar como válido este ítem.

- Aplicación del patrón de referencia en toda la serie: Al analizar este criterio estuvimos en desacuerdo en un 20% de nuestras observaciones. Recordemos que este criterio se cumple si se especifica de forma explícita que la prueba de referencia se realiza de manera sistemática a todos los pacientes que forman parte del estudio a no ser que de uno de los grupos a estudio se tenga, con un grado de seguridad razonable, la certeza de ausencia de enfermedad y así se indique. La mayoría de las veces, alguno de nosotros da por válido el criterio aunque no aparece una frase que indique la aplicación del estándar a todos lo sujetos, por entender en función de los datos que así ha sido, pero otro de nosotros es más estricto exigiendo la aparición de la explicación.

  Este es el caso por ejemplo, del estudio del valor de la concentración de pseudouridina ascítica para discriminar entre cirrosis y cáncer hepático [92]. Los autores detallan: "*La pertenencia de los sujetos a uno de los dos grupos de pacientes (afectados por hepatocarcinoma o ascitis cirrótica) se estableció sobre la base de su histología (considerado el patrón diagnóstico de referencia) después de laparoscopia*". En este trabajo se especifica claramente en qué consiste el patrón de referencia, pero uno de los observadores defiende que hay que ser estricto en la definición del criterio que describe la aplicación del patrón de referencia a todos los participantes en el estudio y darlo como válido solo si aparece una frase que indique que la laparoscopia se realizó a todos los sujetos. Los otros observadores consideran que esta frase no es tan necesaria, ya que en el apartado de resultados queda patente que esta intervención sí se realizó a todos. Finalmente decidimos otorgar validez al criterio, considerando que el autor sí ha reflejado que la prueba estándar se ha realizado en todos los pacientes

- Análisis por subgrupos: La discrepancia en este ítem alcanzó un porcentaje de 22%. Este criterio se cumple si los índices de exactitud se expresan según distintos subgrupos clínicos o demográficos de la población estudiada.

Muchas veces la discrepancia se encuentra en la consideración de qué entendemos por grupo demográfico o clínico de la población. Un ejemplo lo encontramos en el artículo que estudia el valor de la Cistatina C en plasma como un marcador de filtración glomerular en cirrosis hepática descompensada [118]. El estudio divide a los sujetos en grupo control y grupo de casos, y estos a su vez, según tengan o no la función renal conservada. Pero la sensibilidad y especificidad diagnóstica se calcula únicamente diferenciando entre el grupo de casos y el de referencia. Si queremos estudiar el valor de la Cistatina C como marcador glomerular, lo idóneo hubiera sido que los valores de sensibilidad y especificidad se hubieran calculado para los individuos con función renal compensada y para aquellos que la tenían descompensada.

- Prevención del sesgo de secuencia: En este criterio, se alcanza el mayor grado de desacuerdos dentro de la aplicación de los criterios de Reid et al [5], un 32%. Se cumple si todos los sujetos del estudio (independientemente del resultado de la prueba diagnóstica) reciben un diagnóstico definitivo mediante la prueba de referencia. Se considera que está presente si los sujetos con resultados de la prueba positivos o negativos no tienen igual oportunidad de someterse a la prueba de referencia. Las discrepancias encontradas derivan de la interpretación de la secuencia de realización de la prueba de referencia y el patrón estándar a la población a estudio, cuando no aparece una frase que indique cómo se aplicaron.

  Por ejemplo, un estudio presentó la correlación entre la determinación de la Troponina T en un solo punto en la unidad de cuidados coronarios después de un infarto de miocardio, con el tamaño del infarto de miocardio[126]. El patrón de referencia consistió en el diagnóstico mediante la realización de un electrocardiograma. En el trabajo no se explica claramente que a todos los pacientes en estudio se les realizara un electrocardiograma y no solo a los que sufrieron el infarto, pero alguno de nosotros interpretó que independientemente del valor de la determinación de la troponina se les realizó a todos los sujetos el electrocardiograma. No dimos el criterio como válido ya que estrictamente debería aparecer

una frase que indicara que todos los sujetos independientemente del resultado de la prueba diagnóstica recibieron la confirmación por el patrón de referencia

- Precisión de los resultados: El desacuerdo alcanzado en este ítem fue del 17%. Este criterio se cumple cuando se presenta la precisión estadística de las estimaciones de los índices de exactitud mediante el cálculo del intervalo de confianza o error estándar. La discrepancia se obtuvo porque en la sección de análisis estadísticos sí refieren cómo se calculó el error estándar, pero luego en el apartado de resultados no aparece su valor. Se puede pensar que se cumple el criterio porque se calculó dicho parámetro pero ha de aparecer reflejado el valor obtenido.

  Esto ocurre por ejemplo, en el estudio de la determinación de la transferrina deficiente de carbohidrato mediante un nuevo kit por nefelometría [151]. En el apartado de análisis estadísticos, los autores relatan: " *Los datos fueron expresados como media (incluyendo el intervalo de confianza del 95%), mediana y rango para cada grupo*". Algunos observadores han tomado esta frase como el cálculo por parte de los autores del intervalo de confianza, pero no se puede tomar así ya que en el apartado de resultados no aparece el cálculo de dicho valor para los objetivos de medición relevantes, los índices de exactitud diagnóstica.

- Descripción del espectro del estudio: Este parámetro también alcanzó un 17% de disconformidad. Este criterio se cumple si aparecen por lo menos tres de los siguientes cuatro supuestos: deben estar definidas las características demográficas de sexo y edad, los síntomas clínicos o estadíos de la enfermedad (severidad de la enfermedad, duración y morbilidad) y los criterios que se han seguido para la selección de sujetos (criterios de inclusión y de exclusión). El mayor desacuerdo está en determinar lo estricto que hay que ser al definir sí el autor del artículo ha especificado o no los criterios de selección y exclusión de su muestra a estudio.

  Este problema lo tuvimos, por ejemplo, en el artículo que analiza el isoenzima 1 lactato dehidrogenasa sérico como predictor de muerte en pacientes con metástasis testicular de tumores de células germinales [145]. En el apartado de selección de pacientes para el

estudio, los autores exponen: " *Las series de evaluación comprenden 44 de 55 pacientes consecutivos con metástasis testicular de tumores de células germinales tratados en el Hospital Universitario de Odense, Dinamarca, en el período 1980-1984 y que tenían una determinación de S-LD-1. Excluimos 11 pacientes sin dicha determinación. Treinta y siete pacientes tratados en la Universidad de Tejas MD Anderson Cáncer Centro, Houston, Texas, en el período de 1994-1997, sirvieron como series de validación*". Alguno de los observadores defiende que los autores han explicado claramente los criterios de selección y exclusión de los pacientes participantes. Pero el criterio no se pudo dar finalmente como válido ya que no explicaron los criterios de elección de los 37 sujetos que participaban en las series de validación.

Tabla 26: Descripción de las discordancias encontradas al analizar los 96 artículos, mediante la aplicación de las guías de Reid y colaboradores [5] y los 51 artículos con la guía de recomendaciones propia [7, 8]. Se expresan como porcentaje de disconformidades obtenidas en la evaluación de cada criterio y como porcentaje del total de desacuerdos obtenido para cada guía.

	DISCREPANCIAS			Nº ARTÍCULOS ANALIZADOS
	Nº	% DEL CRITERIO	% DEL TOTAL	
1- Objetivos especificados	3	6%	5%	51
2- Justificación de la prueba	17	33%	25%	51
3- PR especificado	8	16%	12%	51
4- PR incorpora prueba a valorar	5	10%	8%	51
5- PR en toda la serie	10	20%	15%	51
6- PD: descripción suficiente	4	8%	6%	51
7- PD: definición normalidad	2	4%	3%	51
8- Origen de la población	7	14%	9%	51
9- Expresión continua de resultados	5	9%	8%	51
10- Índice para pruebas conjuntas	4	8%	6%	51
11- Cálculo valores predictivos	2	4%	3%	51
TOTAL CRITERIOS PROPIOS	**67**		**100%**	**51**
1- Composición del espectro.	16	17%	14%	96
2- Análisis por subgrupos	21	22%	22%	96
3- Prevención sesgo de secuencia	31	32%	27%	96
4- Prevención sesgo de revisión	14	15%	12%	96
5- Precisión de los resultados	16	17%	14%	96
6- Resultados indeterminados	9	9%	8%	96
7- Reproducibilidad	8	8%	7%	96
TOTAL CRITERIOS REID ET AL	**115**		**100%**	**96**

Al analizar los valores de discrepancias obtenidos en la aplicación de los criterios de Reid y colaboradores [5] y de forma individual para cada revista, observamos como el mayor grado de discordancia lo alcanzamos al analizar *Clinical Chemistry and Laboratory Medicine* (20%) y el menor grado en *Revista Clínica Española* (5%), aunque las diferencias no fueron estadísticamente significativas (p=0,592). En las cuatro revistas se mantiene el criterio que estudia la prevención del sesgo de secuencia como el que más desacuerdo ha producido (Tabla 27).

Tabla 27: Descripción de las discordancias obtenidas por cada una de las cuatro revistas analizadas (Clinical Chemistry, Clinical Chemistry and Laboratory Medicine, Medicina Clínica y Revista Clínica Española) al evaluar los 96 artículos mediante la guía de Reid y colaboradores [5].

	***Clin Chem* (52) (n/%)**	***Clin Chem Lab Med* (27) (n/%)**	***Med Clin* (9) (n/%)**	***Rev Clin Esp* (8) (n/%)**
1-Composición del espectro	6 (12%)	9 (33%)	1 (11%)	0
2- Análisis por subgrupos	10 (19%)	9 (33%)	1 (11%)	1 (13%)
3- Prevención sesgo de secuencia	15 (29%)	10 (37%)	4 (44%)	2 (25%)
4- Prevención sesgo de revisión	10 (19%)	2 (7%)	2 (22%)	0
5- Precisión de los resultados	10 (19%)	5 (19%)	1 (11%)	0
6- Resultados indeterminados	8 (15%)	0	1 (11%)	0
7- Reproducibilidad	5 (10%)	3 (11%)	0	0
TOTAL	**64 (18%)**	**38 (20%)**	**10 (16%)**	**3 (5%)**

Diferencias estadísticamente significativas (p<0,05) calculadas mediante Chi^2

Si estudiamos el desacuerdo obtenido en la aplicación de los criterios de la guía de elaboración propia [7, 8] en los artículos evaluados de las cuatro revistas, obtenemos que es en *Clinical Chemistry* donde alcanzamos un mayor desacuerdo y *Revista Clínica Española* sigue siendo la publicación con un menor porcentaje, aunque tampoco en esta caso las diferencias son estadísticamente significativas (p=0,662). Se mantiene el criterio referente a si la valoración de la prueba diagnóstica está justificada o no como el ítem con mayor desacuerdo. (tabla 28).

Tabla 28: Descripción de los desacuerdos obtenidos por cada una de las cuatro revistas (Clinical Chemistry, Clinical Chemistry and Laboratory Medicine, Medicina Clínica y Revista Clínica Española) analizadas al evaluar los 51 artículos mediante la guía de criterios propios [7, 8].

	Clin Chem **(19)** **(n/%)**	***Clin Chem Lab Med*** **(15)** **(n/%)**	***Med Clin*** **(9)** **(n/%)**	***Rev Clin Esp*** **(8)** **(n/%)**
1- Objetivos especificados	3 (16%)	0	0	0
2- Justificación de la prueba	9 (47%)	1 (7%)	3 (33%)	4 (50%)
3- PR especificado	4 (21%)	2 (13%)	2 (22%)	0
4- PR incorpora prueba a valorar	0	2 (13%)	2 (22%)	1 (13%)
5- PR en toda la serie	4 (16%)	3 (20%)	2 (22%)	1 (13%)
6- PD: descripción suficiente	1 (5%)	1 (7%)	1 (11%)	1 (13%)
7- PD: definición normalidad	1 (5%)	1 (7%)	0	0
8- Origen de la población	3 (16%)	3 (20%)	1 (11%)	0
9- Resultados susceptibles de expresión continua	1 (5%)	1 (7%)	2 (22%)	1 (13%)
10- Índice para pruebas conjuntas	1 (5%)	2 (13%)	0	1 (13%)
11- Cálculo valores predictivos	0	1 (7)	1 (11)	0
TOTAL	**54 (26%)**	**17 (10%)**	**14 (14%)**	**9 (10%)**

Diferencias estadísticamente significativas (p<0,05) calculadas mediante Chi^2

6.2- ESTUDIO DE VARIABILIDAD ENTRE LOS OBSERVADORES QUE ANALIZAN MUESTRAS DE LÍQUIDO SINOVIAL PARA LA DETECCIÓN DE CRISTALES DE URATO MONOSÓDICO Y DE PIROFOSFATO CÁLCICO.

En este estudio de concordancia observacional analizamos los principales resultados obtenidos por los analistas al examinar muestras de líquido sinovial, para la detección e identificación de cristales de urato mono sódico y pirofosfato cálcico.

Además estudiamos la aplicabilidad de la guía metodológica de Feinstein [11] para la realización de estudios de variabilidad observacional cuando evaluamos una prueba diagnóstica de laboratorio.

6.2.1- MUESTRAS SELECCIONADAS EN EL ESTUDIO DE CONCORDANCIA OBSERVACIONAL.

Durante el periodo de estudio (septiembre 2001- junio 2003) se analizan un total de 64 muestras de líquido sinovial de pacientes sospechosos de artropatía por cristales: 42 de ellas proceden de Consultas Externas del Servicio de Reumatología, y las otras 22 del Hospital de Día del citado Servicio, ambos pertenecientes al Hospital General Universitario de Alicante.

Se seleccionan las muestras de manera consecutiva, excluyendo aquellas que no tenían un volumen suficiente para su análisis por los cuatro analistas y para su determinación bioquímica y celular.

Los diagnósticos clínicos de los pacientes a estudio fueron:

- Gota, 12 muestras (19%);
- Artropatía relacionada con cristales de pirofosfato cálcico, 16 muestras (25%);
- Artritis reumatoide, 12 muestras (19%);
- Otras artritis inflamatorias incluyendo artritis juvenil idiopática, artritis psoriática, espondiloartropatías y poliartritis no clasificadas, 24 muestras (37%).

En la tabla 29 aparecen reflejadas las características demográficas (sexo, lugar de recogida de la muestra) y clínicas (haber recibido infiltración terapéutica en los últimos seis meses) en función de los diagnósticos de los 64 pacientes a estudio.

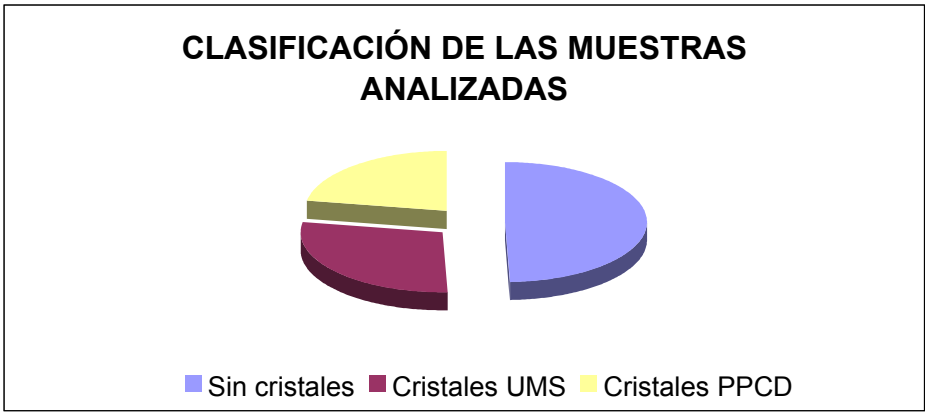
Tabla 29: Características demográficas y clínicas de los 64 pacientes a estudio.

DIAGNÓSTICO	SEXO		INFILTRACIÓN*		PROCEDENCIA	
	M	F	SÍ	NO	CONSULTA	HOSPITAL DE DÍA
Gota	8	4	1	7	6	6
Artritis Reumatoide	4	8	5	3	9	3
Artropatía por cristales de pirofosfato	7	9	13	0	7	9
Otras artritis	8	16	9	0	20	4
TOTAL	**27**	**37**	**28**	**10**	**42**	**22**

M= Masculino; F= Femenino.
** Solo contamos con datos de infiltración terapéutica previa de 38 pacientes.*

Los cuatro analistas, realizaron un total de 194 observaciones: 96 observaciones en muestras que no contienen cristales (49%); 55 observaciones en muestras con cristales de pirofosfato cálcico (28%) y 43 observaciones en muestras con cristales de urato monosódico (23%) (Figura 10).

Figura 11: Representación gráfica de la clasificación de las 194 observaciones realizadas en las muestras de líquido sinovial recibidas durante el periodo de estudio.



*UMS = Urato monosódico; PPCD = Pirofosfato cálcico dihidratado

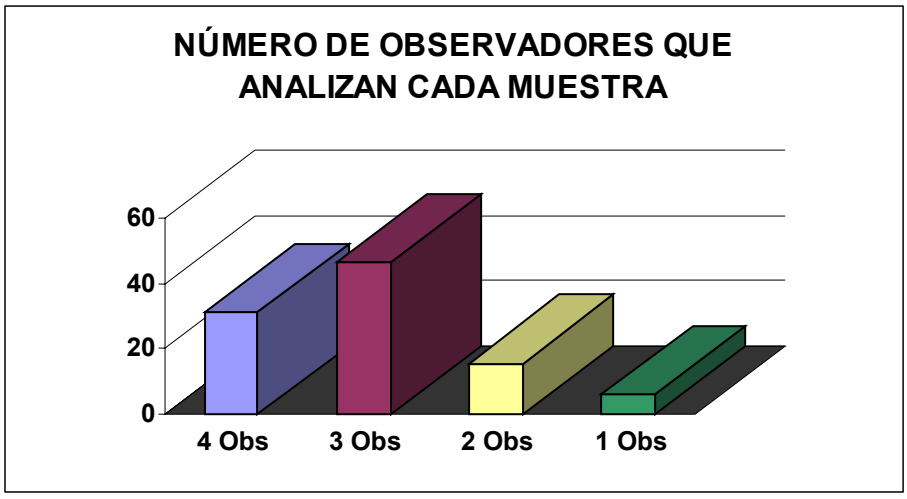
6.2.2- PARTICIPANTES EN EL ESTUDIO DE VARIABILIDAD OBSERVACIONAL.

Las muestras de líquido sinovial son analizadas por cuatro observadores, residentes del Servicio de Análisis Clínicos del citado hospital de manera ciega e independiente.

No todos examinan el total de las 64 muestras enviadas al laboratorio: observador 1 analiza 38 líquidos sinoviales (59%); observador 2, 56 líquidos sinoviales (88%); observador 3, 44 líquidos sinoviales (69%) y observador 4, 56 líquidos sinoviales (88%).

Por lo tanto, del total de las 64 muestras evaluadas, 20 líquidos sinoviales fueron examinados por los 4 observadores (31%), 30 líquidos por 3 observadores (47%), 10 líquidos por 2 observadores (16%) y 4 líquidos por solo 1 observador (6%). Como se observa en la figura 12, la mayoría de las muestras de líquido sinovial, son evaluadas por tres de los cuatro observadores, seguidas por las analizadas por los cuatro observadores, y en menor número, las examinadas por dos y un solo evaluador.

Figura 12: Representación gráfica del número de observadores que analizan cada muestra de líquido sinovial durante el estudio.



6.2.3- GRADO DE CONCORDANCIA ENTRE LOS OBSERVADORES.

En primer lugar, y a título informativo, hemos determinado la variabilidad observacional, mediante el cálculo del índice kappa, de cada uno de los cuatro analistas participantes, con el reumatólogo experto, patrón de referencia.

Tomamos el conjunto de las observaciones realizadas por los cuatro analistas para el total de las 64 observaciones (con cristales de urato monosódico, con cristales de pirofosfato y sin cristales) y las comparamos con los resultados obtenidos por el reumatólogo. Hay que tener en cuenta que como no todos los analistas analizaron el mismo número de muestras, no se pueden calcular determinados parámetros, como el intervalo de confianza del índice kappa. Obtenemos un valor de kappa de 0,85.

Si desglosamos dicho valor kappa para cada tipo de muestra estudiada tomando también los cuatro analistas en conjunto comparados con el experto, alcanzamos los siguientes valores: en las muestras con cristales de urato monosódico el valor de kappa es de 0,93, en muestras sin cristales el índice kappa toma el valor de 0,84 y para muestras con cristales de pirofosfato el valor de kappa es de 0,79.

Por último podemos calcular los valores obtenidos individualmente por cada observador en comparación con el patrón de referencia para el conjunto de todas las muestras analizadas (tabla 30).

Tabla 30: Índice kappa de acuerdo entre los cuatro observadores con el patrón de referencia en la detección e identificación de cristales.

Observador	Acuerdo (%)	Acuerdo esperado (%)	Kappa	Es*	IC 95%**
1	86,84	33,45	0,80	0,11	[0,58 – 1,00]
2	96,43	36,10	0,94	0,09	[0,76 – 1,00]
3	90,91	39,36	0,85	0,11	[0,63 – 1,00]
4	89,29	36,54	0,83	0,09	[0,64 – 1,00]

**Es = Error estándar.* ***IC 95% = Intervalo de confianza*

Centrándonos en la determinación de la variabilidad observacional entre los distintos observadores participantes en el estudio, calculamos los valores de concordancia obtenidos entre los cuatro analistas que evalúan las muestras de líquido sinovial.

En la tabla 31 aparecen reflejados los valores del índice de concordancia kappa para los cuatro observadores del estudio y su intervalo de confianza además del porcentaje de acuerdo y el acuerdo esperado por azar.

Expresamos los resultados de concordancia para los cuatro observadores de dos en dos. De tal forma que cada uno de los cuatro números representa a cada analista que participa en el estudio (analista 1, analista 2, analista 3 y analista 4).

Tabla 31: Índice kappa de concordancia entre los cuatro analistas que analizan las muestras de líquido sinovial a estudio.

Observadores	**Acuerdo (%)**	**Acuerdo esperado (%)**	**Kappa**	**Es***	**IC 95%****
1 – 2	94,12	33,56	0,91	0,12	[0,67 – 1,00]
1 – 3	84,62	36,09	0,76	0,14	[0,49 – 1,00]
1 – 4	87,10	33,40	0,81	0,13	[0,56 – 1,00]
2 – 3	92,68	38,25	0,.88	0,12	[0,66 – 1,00]
2 – 4	87,76	35,03	0,82	0,10	[0,61 – 1,00]
3 – 4	84,62	38,79	0,75	0,12	[0,52 – 1,00]

**Es = Error estándar.*
***IC 95% = Intervalo de confianza*

Debido a que para el estudio de concordancia hemos contado con un experto en la determinación e identificación de cristales en muestras de líquido sinovial, hemos querido calcular también los valores de exactitud diagnóstica de los cuatro analistas, tomando como patrón de referencia dicho experto.

Como he comentado anteriormente, el análisis de cristales en líquido sinovial se compone de dos pasos consecutivos, la detección y la identificación de los cristales, por lo que la determinación de estos índices de exactitud se va a hacer para cada paso individual del procedimiento.

6.2.4- CÁLCULO DE LA EXACTITUD DIAGNÓSTICA EN LA DETECCIÓN DE CRISTALES EN LAS MUESTRAS DE LÍQUIDO SINOVIAL ANALIZADAS.

El primer paso para el análisis de cristales en muestras de líquido sinovial, es la detección de los mismos. La detección de cristales en una muestra de líquido sinovial se define como la determinación de la presencia o ausencia de cristales en dicha muestra (en este caso, cristales de urato monosódico o de pirofosfato cálcico dihidratado).

Hemos calculado los valores de exactitud diagnóstica para cada uno de los observadores participantes en el estudio, en relación con el patrón de referencia, el reumatólogo experto. En la tabla 32 aparecen los resultados de dichos parámetros.

Debido a que no todos los observadores han evaluado el mismo número de muestras, no se puede calcular el valor global de la sensibilidad y especificidad diagnóstica.

Tabla 32: Parámetros de exactitud diagnóstica, (sensibilidad, especificidad, valor predictivo positivo y negativo) de los cuatro observadores participantes en el estudio para la detección de la presencia de cristales en las 64 muestras de líquido sinovial analizadas:

	OBSERVADOR 1	**OBSERVADOR 2**	**OBSERVADOR 3**	**OBSERVADOR 4**
VP	21	28	19	26
VN	13	25	21	24
FP	4	3	2	4
FN	0	0	2	2
% Sensibilidad (IC 95%)	100 (80–100)	100 (85–100)	91 (68–92)	93 (75–99)
% Especificidad (IC 95%)	77 (50–92)	89 (71 97)	91 (7199)	86 (86–95)
% VPP (IC 95%)	84 (63–95)	90 (73–98)	91 (68–92)	87 (68–96)
% VPN (IC 95%)	100 (72–00)	100 (83–100)	91 (71 99)	92 (73–99)

VP= Verdadero positivo; VN= Verdadero negativo; FP= Falso positivo; FN= Falso negativo; VPP= Valor predictivo positivo; VPN= Valor predictivo negativo; IC = Intervalo de confianza.

Con el empleo del microscopio de luz ordinaria obtenemos que de las 96 observaciones en muestras de líquido sinovial sin cristales nosotros definimos correctamente 85 de ellas (89%) y de las 98 observaciones en muestras que contienen cristales (de urato monosódico o de pirofosfato cálcico dihidratado) detectamos su presencia correctamente en 92 (94%). En cambio, si utilizamos el microscopio de luz polarizada, definimos correctamente como muestras sin cristales 86 de las 96 observaciones en muestras de líquido sinovial analizadas (90%) y detectamos la presencia de cristales en 58 de las 98 observaciones en muestras de líquido sinovial analizadas (51%).

6.2.5- CÁLCULO DE LA EXACTITUD DIAGNÓSTICA EN LA IDENTIFICACIÓN DE CRISTALES.

Calculamos los valores de sensibilidad y especificidad para el segundo paso del análisis de cristales, la identificación del tipo de cristal previamente detectado.

Al igual que en el caso anterior, tomamos cada observación como resultado único, aunque no se trate de muestras independientes.

Hay un total de 98 observaciones realizadas en muestras que contienen cristales, bien de urato monosódico, bien de pirofosfato cálcico dihidratado. No analizamos ninguna alícuota con ambas muestras de cristales. Del total de las observaciones realizadas, 43 pertenecen a muestras con cristales de urato monosódico y 55 a muestras que contienen cristales de pirofosfato cálcico dihidratado.

Analizamos en primer lugar los valores obtenidos al identificar muestras con cristales de urato monosódico (tabla 33) y después, los calores para las observaciones en muestras con cristales de pirofosfato cálcico (tabla 34).

En este caso tampoco se han podido calcular los valores globales de sensibilidad y especificidad, debido a que no todos los observadores participantes analizaron el mismo número de muestras de líquido sinovial durante su participación en el estudio.

Tabla 33: Valores de exactitud diagnóstica para la identificación de cristales de urato monosódico *en las 98 observaciones realizadas a muestras de líquido sinovial con cristales.*

	OBSERVADOR 1	**OBSERVADOR 2**	**OBSERVADOR 3**	**OBSERVADOR 4**
VP	11	12	7	11
VN	9	16	13	16
FP	1	0	0	0
FN	0	0	1	1
% Sensibilidad (IC 95%)	100 (68–100)	100 (70–100)	88 (47–99)	92 (60–99)
% Especificidad (IC 95%)	90 (54–100)	100 (76–100)	100 (72–100)	100 (76–100)
% VPP (IC 95%)	92 (60–100)	100 (70–100)	100 (56–100)	100 (68–100)
% VPN (IC 95%)	100 (63–100)	100 (76–100)	100 (64–99)	94 (69–99)

VP= Verdadero positivo; VN= Verdadero negativo; FP= Falso positivo; FN= Falso negativo; VPP= Valor predictivo positivo; VPN= Valor predictivo negativo; IC = Intervalo de confianza.

Tabla 34: Valores de exactitud diagnóstica para la identificación de cristales de pirofosfato cálcico dihidratado en las 98 observaciones realizadas a muestras de líquido sinovial con cristales.

	OBSERVADOR 1	**OBSERVADOR 2**	**OBSERVADOR 3**	**OBSERVADOR 4**
VP	9	16	12	14
VN	11	12	8	11
FP	0	0	0	1
FN	1	0	1	2
% Sensibilidad (IC 95%)	90 (54–99)	100 (75–100)	92 (62–99)	88 (60–98)
% Especificidad (IC 95%)	100 (68–100)	100 (70-100)	100 (60–100)	92 (60–99)
% VPP (IC 95%)	100 (63–100)	100 (76–100)	95 (70–100)	93 (66–99)
% VPN (IC 95%)	92 (60–99)	100 (70–100)	100 (51–99)	85 (54–97)

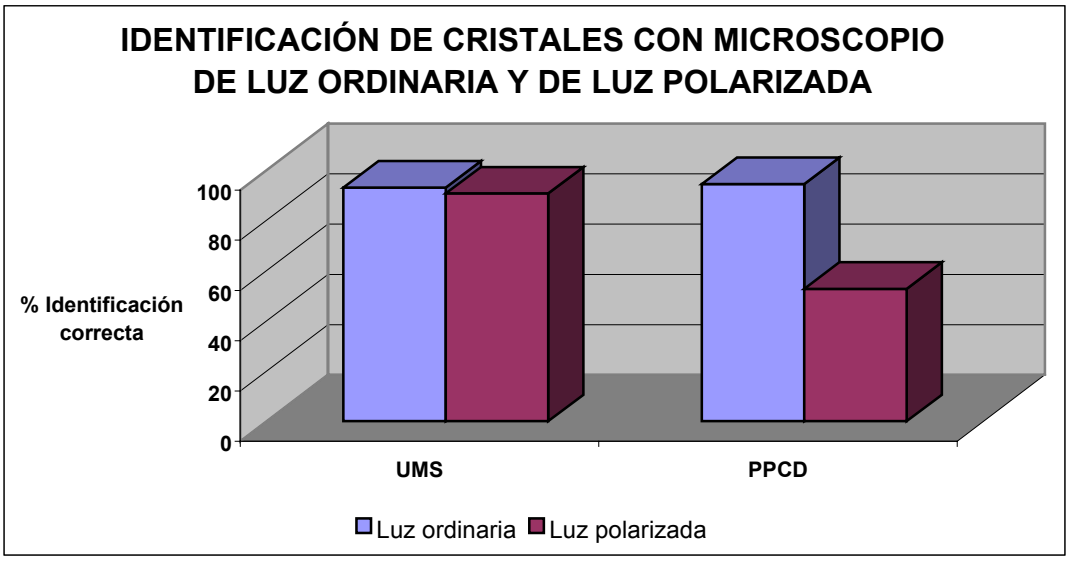
VP= Verdadero positivo; VN= Verdadero negativo; FP= Falso positivo; FN= Falso negativo; VPP= Valor predictivo positivo; VPN= Valor predictivo negativo; IC = Intervalo de confianza.

Al igual que en la detección de cristales, para la identificación de ambos tipos de cristales hemos comparado los resultados obtenidos cuando empleamos para los análisis el microscopio de luz ordinaria con el empleo del microscopio de luz polarizada para los cuatro observadores, en referencia con el estándar (Figura 13).

Cuando empleamos el microscopio de luz ordinaria, identificamos correctamente 52 de las 55 observaciones realizadas en muestras con cristales de pirofosfato cálcico dihidratado (95%) y 40 de las 43 observaciones en muestras con cristales de urato monosódico (93%). Mientras que con el uso del microscopio de luz polarizada, identificamos correctamente 29 de las 55 observaciones realizadas en muestras con cristales de pirofosfato cálcico dihidratado (53%) y 39 de las 43 observaciones en muestras con cristales de urato monosódico (91%).

Esto nos indica cómo los cristales de pirofosfato cálcico se identifican mucho más fácilmente con el microscopio de luz ordinaria, ya que como he referido previamente la mayoría de estos cristales no son birrefringentes.

Figura 13: Uso comparativo del microscopio de luz ordinaria y del microscopio de luz polarizada en la identificación de cristales (urato monosódico o pirofosfato cálcico dihidratado) de las 98 muestras analizadas en el estudio:



*UMS = Urato monosódico; PPCD = Pirofosfato cálcico dihidratado

6.2.6- DETERMINANTES DE LAS OBSERVACIONES REALIZADAS QUE PUEDEN EXPLICAR LAS CAUSAS DE LAS INCONSISTENCIAS.

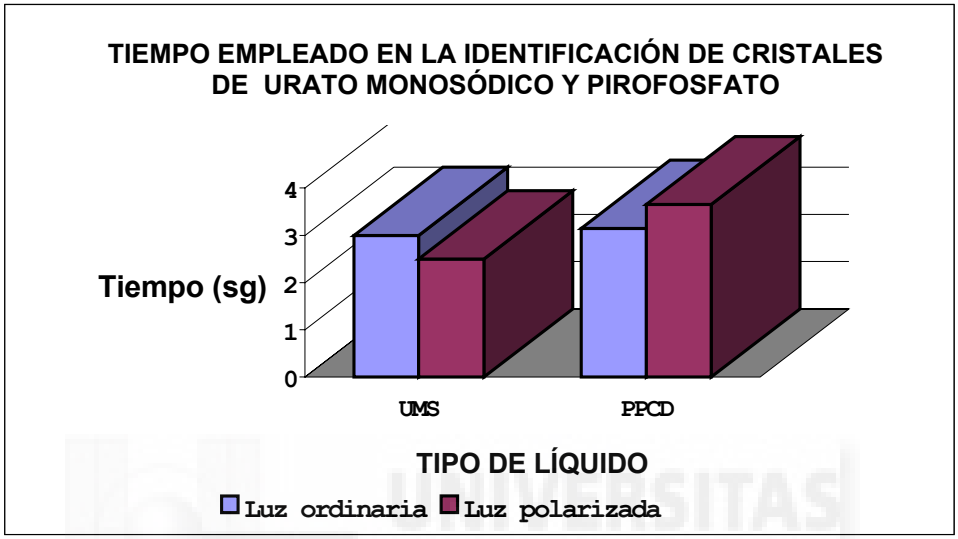
A continuación se exponen una serie de características, bien de las propias muestras de líquido sinovial analizadas, bien del propio proceso de la observación, que pueden explicar el por qué de algunas de las determinaciones erróneas que han realizado los analistas.

6.2.6.1- TIEMPO EMPLEADO EN LA DETECCIÓN E IDENTIFICACIÓN DE CRISTALES EN MUESTRAS DE LÍQUIDO SINOVIAL.

Hemos determinado el tiempo medio que han tardado los analistas en detectar un cristal en las muestras de líquido sinovial. Se ha medido tanto el tiempo empleado con el uso del microscopio de luz ordinaria como con el microscopio de luz polarizada. Para detectar la presencia o ausencia de cristales en las muestras de líquido sinovial analizadas se necesitan una media de 3,9 segundos cuando se emplea el microscopio de luz ordinaria y 4,1 segundos de media con el microscopio de luz polarizada.

Una vez detectada la presencia de cristales, el siguiente paso consiste en identificar el tipo de cristal, urato monosódico o pirofosfato cálcico dihidratado. Se ha medido también el tiempo que se tarda en identificar el tipo de cristal de que se trata tanto utilizando el microscopio de luz ordinaria como el de luz polarizada. Con el microscopio de luz ordinaria obtenemos un valor de 3,0 segundos para identificar un cristal de urato monosódico y de 3,2 segundos para la identificación de un cristal de pirofosfato cálcico dihidratado; si utilizamos el microscopio de luz polarizada tardamos 2,5 segundos en la detección de cristales de urato monosódico y 3,7 segundos para muestras con cristales de pirofosfato cálcico dihidratado (Figura 14). Nuevamente se muestra como los cristales de pirofosfato cálcico son más difícilmente identificables mediante el uso del microscopio de luz polarizada.

Figura 14: Representación del tiempo empleado en la identificación de cristales de urato monosódico y de pirofosfato cálcico dihidratado en las muestras de líquido sinovial, en función del tipo de luz empleada.



*UMS = Urato monosódico; PPCD = Pirofosfato cálcico dihidratado

6.2.6.2- NÚMERO DE CAMPOS EVALUADOS PARA LA DETECCIÓN E IDENTIFICACIÓN DE CRISTALES EN MUESTRAS DE LÍQUIDO SINOVIAL.

Estudiamos el número de campos ópticos que fueron necesarios analizar, tanto con el empleo del microscopio de luz ordinaria como de polarizada, para detectar e identificar cristales de urato monosódico y de pirofosfato cálcico. Comprobamos como en la identificación de cristales de urato monosódico analizamos un menor número de campos si utilizamos el microscopio de luz polarizada frente al uso del microscopio de luz ordinaria (1,5 frente a 2,0 número medio de campos ópticos analizados); para identificar cristales de pirofosfato cálcico es con el microscopio de luz ordinaria con el que tenemos que observar menor número de campos con respecto al microscopio de luz polarizada (2,6 frente a 3,9 número medio de campos ópticos analizados).

6.2.6.3- ANÁLISIS CELULAR Y DE PARÁMETROS BIOQUÍMICOS.

A) Recuento celular (hematíes y leucocitos) y diferencial leucocitario en las muestras analizadas en el estudio.

Hemos querido comprobar si el número de células presentes en una muestra de líquido sinovial puede influir en el resultado obtenido en su análisis. Para ello hemos tomado como observaciones correctas aquellas en las que los observadores coinciden en sus resultados con los del patrón de referencia y como observaciones incorrectas cuando no sucede así. En cada uno de los dos grupos de acuerdos o desacuerdos, medimos la media (y desviación estándar) del recuento celular de hematíes y leucocitos y el diferencial leucocitario expresado como porcentaje de polimorfonucleares (tabla 35).

Encontramos un mayor número de desacuerdos en aquellas muestras con elevado recuento de hematíes y un mayor grado de acuerdo en aquellas muestras con elevado porcentaje de leucocitos. La explicación puede estar está en que la mayoría de los cristales son intraleucocitarios y a mayor número de leucocitos más probabilidades hay de detectar e identificar un cristal correctamente, mientras que un mayor número de hematíes puede interferir en la visualización de los cristales. El diferencial leucocitario no influye ya que en la mayoría de los casos por tratarse de muestras inflamatorias el predominio es de polimorfonucleares.

Tabla 35: Valores medios (y desviación estándar) de recuento celular (hematíes y leucocitos) y diferencial leucocitario para las 64 muestras de líquido sinovial analizadas, en función de su correcta o incorrecta identificación.

	ACUERDOS (media/desv)	**DESACUERDOS (media/desv)**	**p**
HEMATÍES ($10^6/\mu l$)	3.843 (6.453)	9.437 (7.770)	<0,05
LEUCOCITOS ($10^3/\mu l$)	5.421 (6.330)	916 (958)	<0,05
% PMN*	74 (11)	73 (16)	0,68

**PMN= Polimorfonucleares.*

B)- Parámetros bioquímicos de glucosa y proteínas en las muestras analizadas en el estudio.

Al igual que hemos hecho en el análisis celular, hemos determinado la concentración media de los parámetros bioquímicos de glucosa y proteínas (media y desviación estándar) para comprobar su posible influencia en la consecución o no de acuerdos de los observadores comparando con el patrón de referencia. En este caso no se han encontrado diferencias estadísticamente significativas para los parámetros estudiados entre los dos tipos de observaciones (tabla 36).

Tabla 36: Valores medios (y desviación estándar) de glucosa y proteínas para las muestras de líquido sinovial analizadas, en función de su correcta o no identificación.

	ACUERDOS (media/desv)	**DESACUERDOS (media/desv)**	**p**
GLUCOSA (mg/dl)	114,6 (64,41)	107,8 (15,2)	0,41
PROTEINAS (g/dl)	4,02 (1,4)	3,8 (0,5)	0,24

6.2.6.4- TIEMPO TRANSCURRIDO DESDE LA ARTROCENTESIS POR EL REUMATÓLOGO HASTA EL ANÁLISIS DE LAS MUESTRAS POR LOS CUATRO ANALISTAS.

Otro de los posibles determinantes que ha podido influir en la consecución de resultados no concordantes ha podido ser el intervalo de tiempo diferente en unos análisis y otros entre la obtención de la muestra de líquido sinovial y su análisis.

Es recomendable realizar el estudio pocas horas después de la artrocentesis ya que se ha visto que se puede producir una reducción en el recuento de células y en el número de cristales de pirofosfato cálcico dihidratado, lo que puede llevar a una mala clasificación de la muestra. Aunque estudios previos sugieren que la muestra de líquido sinovial se puede almacenar hasta 72 horas después de su extracción [169].

Calculamos el tiempo transcurrido desde la extracción de la muestra hasta su análisis, tanto para las observaciones correctas como para las erróneas. En la tabla 37 se puede observar como no

influyó el tiempo transcurrido en nuestros resultados, ya que es muy similar en ambos casos y la media no supera las dos horas.

Tabla 37: Tiempo transcurrido entre la artrocentesis y el análisis de las muestras de líquido sinovial, para las observaciones erróneas y las correctas.

	ACIERTOS	FALLOS	p
MEDIA (DESVIACIÓN ESTÁNDAR) (min)	77(85)	88 (86)	0,46

6.2.6.5- INFILTRACIÓN DE LOS PACIENTES PREVIA A LA OBTENCIÓN DE LAS MUESTRAS PARA EL ESTUDIO.

El haber recibido una infiltración previa de tratamiento puede hacer que confundamos los cristales de cortocoesteroides con los cristales de urato monosódico o de pirofosfato cálcico dihidratado, por lo que queremos comprobar si el haber recibido o no una inyección de tratamiento previametne al análisis de la muestra ha podido influir en la obtención de resultados concordantes con los del reumatólogo experto.

Tenemos datos acerca de la infiltración previa o no del paciente en 115 de las 194 observaciones: de ellas hicimos un diagnóstico incorrecto en 11: en 6 no se había administrado infiltración previa y en 5 sí. No se observaron diferencias estadísticamente significativas entre ambas.

6.2.7.- APLICABILIDAD DE LA GUÍA METODOLÓGICA DE FEINSTEIN [11] EN UN ESTUDIO DE VARIABILIDAD DE UNA PRUEBA DIAGNÓSTICA DE LABORATORIO.

Como he mencionado al principio del estudio, se ha publicado únicamente una guía metodológica [11] con los criterios que debe cumplir un estudio de variabilidad observacional si queremos alcanzar resultados de calidad. Estas recomendaciones no se han puesto en práctica todavía en la determinación de ninguna prueba de laboratorio y por lo tanto no se ha podido constatar la dificultad en su aplicación o sus características más relevantes. Por este motivo, nosotros hemos realizado un estudio de concordancia de una prueba de laboratorio, con antecedentes controvertidos por su posible variabilidad observacional, mediante la aplicación de estas indicaciones de Feinstein [11].

De cada uno de los criterios más relevantes de esta guía metodológica, hemos estudiado su aplicación en nuestro estudio de análisis de cristales en muestras de líquido sinovial:

A) Idoneidad de las muestras estudiadas:

Después de establecer el objetivo específico del estudio, que consiste en evaluar la concordancia entre observadores en la detección e identificación de cristales en muestras de líquido sinovial, es necesario determinar si las muestras analizadas son adecuadas para la realización del estudio, es decir, si reflejan la realidad clínica. El problema surge cuando las muestras elegidas para el estudio, no coinciden con las que se pueden encontrar normalmente en la práctica clínica, ya que para el estudio se seleccionan muestras de pacientes con sospecha de artropatía por cristales para así aseguramos de tener líquidos sinoviales sin cristales y otros con cristales de urato monosódico y con cristales de pirofosfato cálcico dihidratado en cambio, en la práctica las muestras patológicas no suelen ser la mayoría.

B) Partes del análisis:

Antes de iniciar el estudio es necesario especificar que se trata de un estudio de variabilidad observacional y que la reproducibilidad de la técnica que estamos evaluando ya ha sido estudiada. En este caso, el uso del microscopio en la detección e identificación de cristales está estandarizado, por

lo que no es necesario analizar nuevamente su reproducibilidad y solo nos centraremos en el estudio de la concordancia entre los observadores.

C) Características de las observaciones realizadas.

Es importante que todas las observaciones se realicen de manera independiente y ciega ya que solo así se puede determinar la variabilidad de las observaciones efectuadas. En nuestro estudio se mantuvo este enmascaramiento hasta el final del análisis mediante la anotación del resultado de cada observador en registros independientes y la entrega de las muestras sin ninguna identificación aclaratoria. Esta información se considera un requisito imprescindible y aunque es sencillo incluirla en cualquier estudio, está ausente en muchos de los trabajos previos. En este caso, no fue difícil mantener el enmascaramiento, ya que los analistas reciben las muestras identificadas únicamente con un código que anota el reumatólogo y evalúan las muestras de manera individualizada y a distintos tiempos.

D) Características de los observadores participantes.

Debido a que se trata de un estudio de variabilidad observacional, es muy importante definir el grado de experiencia y formación de los observadores participantes en el estudio. Este es un punto del que adolecen los trabajos previos y quizás la causa de la inconsistencia de sus resultados.

En el presente trabajo los observadores no tienen experiencia previa en este tipo de análisis, pero un reumatólogo con amplia experiencia en este campo se ha encargado de su instrucción. Durante el curso formativo ha sido importante después de explicar de manera teórica los dos pasos en los que se compone el análisis de muestras de líquido sinovial, detección e identificación de cristales, hacer un seguimiento práctico de los observadores hasta que el patrón de referencia considera que han adquirido los conocimientos necesarios. De esta manera se ha podido comprobar como, cuando los observadores han sido entrenados en la detección e identificación de cristales en muestras de líquido sinovial, sus resultados son consistentes. Pensamos que es uno de los puntos donde más dificultad hemos encontrado, ya que no es fácil determinar que el grado de formación alcanzado por los observadores sea suficiente para realizar la prueba diagnóstica. En este trabajo,

una vez analizadas las primeras veinte muestras de líquido sinovial se revisan los resultados para confirmar que los analistas han sido capaces de detectar e identificar los cristales adecuadamente.

E) Escala de expresión de los resultados obtenidos y del desacuerdo alcanzado entre los observadores participantes.

Otro de los puntos clave en la realización de un estudio de variabilidad observacional es la elección mediante consenso de cómo expresar las observaciones realizadas. Es un punto que se debe aclarar previamente al comienzo del estudio y debe ser idéntico para todos los participantes. Se trata de un requisito que hay que tener en cuenta en la fase de preparación del estudio y que también es sencillo de llevar a cabo.

En nuestro trabajo se expresan las observaciones realizadas en un cuaderno de anotaciones para cada uno de los analistas, donde se formulan los resultados de la misma manera: es necesario examinar 30 campos del microscopio para confirmar ausencia de cristales. Además indican si el número de cristales encontrados en el primer campo es mayor o menor a cinco.

Antes del inicio del trabajo se debe llegar también al consenso de la definición de desacuerdo. En este caso definimos previamente que dos observadores van a tener discrepancias cuando sus resultados de detección de cristales (presencia o ausencia de cristales) y de identificación (tipo de cristal) sean distintos. Otras variables como es el número de campos necesarios observar para detectar o identificar un cristal no implicarán desacuerdo, pero ayudarán a describir la dificultad del análisis.

F) Índice de concordancia entre los observadores participantes.

En un estudio de variabilidad observacional es imprescindible el cálculo de un índice que nos cuantifique la concordancia entre los participantes, ya que solo así podremos saber si nuestro estudio es correcto. En este caso, hemos elegido para expresar la variabilidad entre los analistas el índice kappa: nuestros resultados muestran un alto nivel de concordancia entre los resultados obtenidos por los cuatro observadores después de su entrenamiento.

G) Procedimiento realizado en el estudio de la variabilidad de la detección de cristales en muestras de líquido sinovial.

Todo estudio de variabilidad se compone de dos fases: en un primer paso, seleccionamos y obtenemos las muestras a estudio y realizamos las observaciones, y en un segundo paso, transformamos las consideraciones obtenidas en resultados. Estas dos secciones deben estar definidas previamente a la realización del estudio.

En nuestro trabajo y en referencia a la obtención de las muestras, hay que resaltar que el análisis de las muestras se realizó casi inmediatamente después de su extracción (en un tiempo máximo de dos horas), por lo que no influyó en la obtención de resultados correctos o erróneos. En cuanto a la realización de las observaciones, el procedimiento de dos pasos seguido por nosotros en el análisis de cristales (primero detección de los cristales y después identificación de los cristales detectados) puede ser determinante de los buenos resultados obtenidos.

Respecto a la obtención de resultados, es importante que previamente se hayan definido los conceptos de acuerdo y desacuerdo en las observaciones realizadas para luego poder expresarlos mediante el índice kappa.

H) Mejoras y recomendaciones de la técnica evaluada.

Después de la realización de un estudio de variabilidad y analizar las posibles causas de sus inconsistencias, se deben establecer una serie de recomendaciones para su mejora. En nuestro trabajo, para aumentar la reproducibilidad es importante establecer las normas del procedimiento con que se realiza la prueba y sobre todo formar a los observadores para que adquieran los conocimientos suficientes.

7- DISCUSIÓN

Esta tesis doctoral ha tenido como objetivo principal el estudio de la metodología diagnóstica aplicada a las pruebas de laboratorio realizadas en los Servicios de Análisis Clínicos. Para ello, por un lado hemos analizado el estado de la calidad metodológica de los artículos que estudian pruebas de laboratorio, y por otro hemos estudiado la consistencia en el diagnóstico de una prueba determinada de laboratorio. Resaltar que este es el primer estudio que se centra en la evaluación de pruebas diagnósticas de laboratorio, lo que puede conducir a un mejor conocimiento de su metodología, y por lo tanto a una mejora de sus deficiencias.

Respecto a la primera sección del trabajo, destacar que el análisis de los resultados obtenidos indica que se ha producido una mejora apreciable en la calidad metodológica de la investigación, comparable con la encontrada en revisiones previas [5, 7, 8], aunque este avance es compatible con algunas deficiencias: los artículos analizados de las dos revistas nacionales, *Medicina Clínica* y *Revista Clínica Española*, muestran una escasa mejora en su metodología a nivel general, aunque sí se observa algún avance en algunos criterios de manera individual; en el ámbito internacional, se ha producido una mejora significativa de la metodología diagnóstica de los trabajos evaluados a partir del año 2002 y en concreto en la revista *Clinical Chemistry*, publicación desde la que más esfuerzos han surgido para aumentar la calidad diagnóstica. Tanto a nivel nacional como internacional, destaca la deficiencia que muestran dos criterios con gran relevancia en los estudios de exactitud diagnóstica; se trata de la descripción de la procedencia de los sujetos que participan en los estudios y la presentación o no de los resultados indeterminados obtenidos en la evaluación de una prueba de laboratorio.

Por otro lado, hay que destacar también la comprobación de un aspecto que esperábamos encontrar y que constituye uno de los principales problemas de las investigaciones efectuadas en los Servicios de Análisis Clínicos: la falta de comunicación entre el clínico y el laboratorio. Este se ha reflejado en el diferente cumplimiento de determinados criterios metodológicos en función del ámbito donde se ha realizado la evaluación de la prueba diagnóstica.

Por último, la utilización en la evaluación de artículos concretos de dos guías metodológicas, ha supuesto la constatación de la dificultad de aplicación de algunos de sus criterios, que se ha

reflejado en el mayor número de discordancias encontradas entre los investigadores participantes en el estudio.

A continuación, paso a describir con más detalle los resultados más relevantes obtenidos en el trabajo.

Como he comentado, no se ha producido un cambio significativo en la metodología diagnóstica de los trabajos publicados en las dos revistas nacionales de carácter médico general evaluadas, cuando comparamos con los resultados obtenidos en revisiones previas [5, 7, 8]. La explicación puede estar en que no ha transcurrido el tiempo suficiente para que se incorporen a la investigación las recomendaciones que surgieron de estas revisiones, o las indicaciones que grupos editores iniciaron en el año 1997 [24], o que estas no han tenido tanta repercusión en estas revistas. Sin embargo, si analizamos los criterios de manera individualizada, sí que se refleja un esfuerzo por mejorar en determinados ítems como son la determinación del error estándar de los parámetros de exactitud diagnóstica obtenidos y la descripción de las características clínicas y demográficas de los pacientes, que permite que se calculen los valores de sensibilidad y especificidad diagnóstica para cada subgrupo de la población en función de sus rasgos.

Centrándonos en el ámbito internacional, y con referencia a los resultados obtenidos, comprobamos como no se han producido mejoras estadísticamente significativas entre los años 1996 y 2001 en las dos publicaciones analizadas, *Clinical Chemistry* y *Clinical Chemistry and Laboratory Medicine*, así como tampoco para cada uno de los ítems metodológicos examinados de manera individual. Esto puede ser debido a que, aunque las recomendaciones de mejora de calidad metodológica por parte de los editores se hicieron en los años 1997 y 2000 [24, 25], no se han incorporado todavía a los trabajos de evaluación diagnóstica. En cambio, constatamos como la calidad de la metodología diagnóstica mejora sobre todo a partir del año 2002, donde este avance es estadísticamente significativo, pudiendo reflejar ya los efectos positivos del impacto de las intervenciones específicas de los editores de las revistas. Esta idea se ve reforzada por la diferente calidad encontrada en las dos revistas internacionales de laboratorio evaluadas: *Clinical Chemistry*

mejora a lo largo del periodo de estudio y no lo hace *Clinical Chemistry and Laboratory Medicine*. Hay que tener en cuenta que, *Clinical Chemistry* incluyó criterios similares a los de Reid y colaboradores [5] en sus instrucciones a los autores en 1996 y en los años 1997 y 2000 publicó versiones preliminares del listado de calidad de los 25 estándares de STARD [24, 25]. En cambio, *Clinical Chemistry and Laboratory Medicine* no comienza a abogar por una investigación de calidad hasta la publicación de los criterios de STARD a principios del año 2003 [170].

Resultados similares aunque en otra área de investigación, fueron observados por Moher y colaboradores [171] y Altman y colaboradores [172], que demostraron cómo la calidad de los trabajos publicados en revistas que promovieron los Consolidated Standards of Reporting Trials (CONSORT) (*BMJ, JAMA,* y *Lancet*) mostraban mayor mejora que aquellos que no abogaban su uso (*New England Journal of Medicine*).

Si analizamos, al igual que hemos hecho con las publicaciones nacionales, los criterios metodológicos aplicados de manera individual, observamos como también han mejorado de forma significativa los criterios referentes a la descripción de las características de los pacientes y al cálculo de la precisión de los parámetros de sensibilidad y especificidad diagnóstica. Respecto al primero de ellos, todavía hace falta incidir más en su cumplimiento, ya que a pesar de la importancia que tiene en un estudio de exactitud diagnóstica, apenas supera el 70% de los trabajos analizados. Este criterio se refiere a la generalización de los resultados y no a la posibilidad de que el estudio pueda producir resultados erróneos, ya que una descripción adecuada de los sujetos que participan lo que va a permitir es que los índices de sensibilidad y especificidad diagnóstica tengan mayor aplicabilidad clínica. En referencia a la determinación del error aleatorio, ha sido una de las cuestiones más olvidadas en los estudios sobre pruebas diagnósticas, sin embargo en estos trabajos se ha observado una mejora significativa, importante si tenemos en cuenta que el objetivo de medición de estos estudios es medir proporciones (sensibilidad y especificidad diagnóstica) y la forma más informativa de cuantificar el error aleatorio será estimando los intervalos de confianza alrededor de estas proporciones.

Reflexión aparte merece el criterio referente a la determinación de la reproducibilidad de una prueba de laboratorio. Aunque es un criterio que se cumple ampliamente en los trabajos estudiados, llama la atención la disminución, aunque no estadísticamente significativa, de su seguimiento a partir

del año 2002. Debido a la importancia que tiene esta característica, ya que el valor de una prueba de laboratorio puede depender de su capacidad de dar siempre el mismo resultado, sería necesario verificar si este alejamiento en su cumplimiento es un hecho que se ha producido de manera aislada en el año 2002, o es una tendencia que se irá observando en los siguientes años.

Tanto en los artículos publicados en revistas nacionales de medicina general, como en los que aparecen en las revistas internacionales de laboratorio, llama la atención el escaso cumplimiento de dos criterios metodológicos de gran importancia en los estudios de exactitud diagnóstica. El primero de ellos, es el ítem que describe la procedencia de los pacientes evaluados. Es decir, el criterio que ha seguido el autor para seleccionar finalmente a los sujetos, así como especificaciones que nos aclaren si esta muestra refleja o no la población real. El que no se especifiquen estos datos, tiene como consecuencia que el clínico no sepa si los resultados descritos en un artículo que evalúa una determinada prueba se pueden aplicar a su práctica diaria, cuando la situación ideal en la valoración de una prueba diagnóstica es realizarla en una población similar a la que luego pretende aplicarse la prueba. El segundo criterio, se refiere a la presentación o no de los resultados indeterminados obtenidos en un estudio. Es importante especificar todos los resultados que se han obtenido, incluidos aquellos que no quedan claros por ser equivocados o no diagnósticos, ya que la prueba puede tener una baja efectividad clínica si sus resultados no pueden ser interpretados. Realmente no hay causa que explique estas deficiencias, pero se ha observado como la definición de la procedencia de los pacientes se refleja más en aquellos artículos que se realizan en departamentos clínicos, por lo que hay que incidir nuevamente en la importancia que tiene la comunicación entre el clínico que trata a los sujetos y el analista que realiza la prueba en el laboratorio. En cuanto a la descripción de los resultados indeterminados obtenidos al evaluar una prueba diagnóstica, aunque es un concepto ya descrito por Reid y colaboradores [5] y recomendado tanto en STARD [32] como en otras guías metodológicas aparecidas recientemente, como es el caso de QUADAS (Quality Assesment of Diagnostic Accuracy Studies) [173], quizás no ha sido un parámetro especialmente bien definido y sobre el que se haya valorado su importancia real.

Como he referido, encontramos un diferente cumplimiento de determinados criterios entre aquellos estudios realizados en servicios analíticos, y otros que se llevan a cabo en departamentos exclusivamente clínicos. Esto ha reforzado la hipótesis de que la separación existente entre estos dos tipos de centros conduce a una deficiencia en la calidad de sus investigaciones. Hemos podido comprobar como aquellos criterios definitorios de las características demográficas de los sujetos y sus criterios de inclusión y exclusión o su procedencia, aparecen mejor especificados en aquellos trabajos que se han realizado en servicios clínicos hospitalarios o que se publican en revistas de carácter médico general. Otros, como son la determinación de la reproducibilidad de la técnica analítica o una mejor definición de la prueba con la que se compara el test diagnóstico que se evalúa, se especifican mejor en los trabajos realizados en laboratorios. Esto, se puede explicar por la procedencia de los individuos de consultas o unidades de ingreso hospitalarias, datos más cercanos a los clínicos. El laboratorio, en cambio, tiene un menor acceso a estas características clínicas y demográficas de los pacientes, pero sí cuida más los aspectos relativos a la realización de la técnica como son el cálculo de su reproducibilidad o la definición de dicha prueba, datos que muchas veces no llegan al clínico, o a este no le interesan tanto. Aunque hay que resaltar el interés que desde los servicios clínicos se está poniendo en la investigación clínica, y que se manifiesta en un mayor cálculo de la precisión de los resultados y de parámetros de exactitud diagnóstica como son los valores predictivos.

Con la realización de este trabajo, además de analizar la calidad diagnóstica de la evaluación de pruebas de laboratorio y constatar si ha habido alguna mejora en este campo, hemos querido describir la aplicabilidad de las guías metodológicas empleadas. Ésta, se ha podido valorar cualitativamente mediante el análisis de la concordancia de los investigadores en la evaluación crítica de los artículos y por lo tanto, con la descripción de sus principales discrepancias. Si conocemos con exactitud todos los criterios que debe cumplir un estudio de exactitud diagnóstica de una prueba de laboratorio, los investigadores podrán aplicarlos al diseño de sus estudios, mejorando el rigor metodológico de la investigación y aumentando la calidad en la comunicación de sus resultados.

Recordamos que para la realización del estudio aplicamos tanto la guía de estándares metodológicos utilizada por Reid y colaboradores [5], como una serie de recomendaciones propia de los autores [7, 8]. La guía de recomendaciones que actualmente está en vigencia para la publicación

de artículos de diagnóstico y que está respaldada por diversos grupos editoriales es el STARD (Standards for Reporting of Diagnostic Accuracy) [32]. Además, casi al término de esta tesis doctoral ha surgido una nueva iniciativa para el estudio de la calidad metodológica de los estudios de exactitud diagnóstica que se incluyen en revisiones sistemáticas, QUADAS (Quality Assesment of Diagnostic Accuracy Studies) [173]. Estas dos guías metodológicas comparten muchos de los criterios que hemos aplicado en este trabajo, pero todavía no se ha estudiado su aplicabilidad a estudios concretos de pruebas de laboratorio, ni los problemas que pueden surgir en el empleo de sus criterios, por lo que este análisis nos permite describir los beneficios y problemas metodológicos de éstas. A esto hay que añadir que, en la guía STARD [32] apenas aparecen definidos cada uno de los 25 ítems que la componen, ya que solo se muestra un ejemplo del cumplimiento de cada uno de ellos, lo que puede suponer una dificultad añadida en su empleo.

Entre todos los parámetros aplicados en este trabajo, destacan algunos en los que hemos encontrado más dificultad en su uso, y que por lo tanto, deberían definirse mejor para su aplicación en posteriores evaluaciones:

Uno de los principales problemas que ha surgido se refiere a la determinación de si la evaluación diagnóstica realizada está o no justificada. Como he comentado previamente, es necesario que cuando se presenta un análisis de una prueba diagnóstica, el autor especifique claramente si es la primera vez que se analiza o si hay estudios previos y qué aportan. Solo si esta especificación aparece en el trabajo, podremos considerar que cumple el criterio referente a la justificación de la prueba diagnóstica. Es importante tener en cuenta que muchas veces, aunque los autores indican que se trata de un test nuevo, en el apartado de referencias bibliográficas aparecen citas previas que también lo han estudiado. Pensamos que la consecución de este parámetro aunque no interfiere en la consecución de unos resultados de calidad, es importante por la valoración del papel clínico que puede jugar una nueva prueba de laboratorio, además de permitir conocer los aspectos más relevantes en la actual investigación diagnóstica. Aunque es un criterio que no se refleja en ninguna de las nuevas guías metodológicas que han ido apareciendo, pensamos que es una característica importante y que si debería incorporarse a este tipo de estudios.

Encontramos también dificultad en valorar si un artículo ha definido claramente o no qué prueba se toma como patrón de referencia para distinguir a los pacientes con enfermedad de los que no la tienen. Este criterio, forma parte de la guía establecida por los autores [7, 8], pero no de la de Reid y colaboradores [5]. Sin embargo, aunque se ha incorporado ya a la guía STARD [32], es en los nuevos criterios de QUADAS [173], donde ha cobrado más importancia, lo que consideramos un acierto. En esta, al igual que en la guía establecida por los autores [7, 8], se estudian varias características de esta prueba estándar además de su definición como son el estudiar si la prueba que se evalúa forma parte de su definición, lo que aumentaría el acuerdo entre esta prueba de referencia y la diagnóstica, sobreestimando por lo tanto los parámetros de exactitud diagnóstica, y si se ha aplicado el test o no a todos los sujetos que forman parte del estudio. Creemos que es un parámetro relevante en este tipo de investigaciones, ya que a veces estas pruebas estándares son muy complejas debido a que están formadas por un conjunto de pruebas complejas, las cuales muchas veces en la clínica no se aplican a todos los pacientes.

Otra discordancia entre los participantes en el estudio, la encontramos en referencia a si se han descrito o no los filtros que el investigador ha empleado para elegir a su muestra de estudio. En este apartado, hay que ser estrictos, ya que se deben especificar claramente, tanto en el grupo de estudio como en el grupo control (en caso de que lo haya). Este parámetro, corresponde a la guía establecida por Reid y colaboradores [5], y también forma parte de STARD [32] y de los ítems de QUADAS [173]. Nosotros pensamos que además de establecer estos criterios que han llevado a los autores a seleccionar a los pacientes, también es muy importante definir el ámbito del que proceden los sujetos, ya que para su aplicación en otros contextos, es importante conocer si el primer centro al que acudieron fue un centro de salud, de especialidades... entre otros. Esta última característica también forma parte de STARD [32], lo que consideramos un paso en la consecución de una investigación de calidad, pero falta quizás que la defina con más claridad.

Es primordial explicar a continuación las limitaciones que tiene nuestro estudio, y que aunque hemos intentado minimizarlas, han podido tener influencia en los resultados obtenidos.

Como se ha podido ver, los años de estudio de los artículos a nivel nacional e internacional no pertenecen al mismo período, aunque abarcan una etapa parecida y por lo tanto son comparables. Esto es debido a la realización de estos estudios en distintos momentos: el análisis en revistas nacionales, se hizo primero con el objetivo de evaluar el cambio producido en la investigación española, eligiendo para ello el periodo de tiempo de 1997 al año 2000. Al constatar la publicación de diversas recomendaciones en este campo, realizamos un segundo estudio analizando los cambios producidos en el ámbito internacional entre los años 1996 y 2001. Más adelante, y por consejo del editor de *Clinical Chemistry*, ampliamos la evaluación al año 2002.

Aunque las revisiones previas están realizadas en revistas de carácter médico general, para la segunda parte de la revisión metodológica hemos elegido dos revistas de ámbito exclusivo de laboratorio, *Clinical Chemistry* y *Clinical Chemistry and Laboratory* Medicine. La razón está en que el mayor esfuerzo para la mejora de este tipo de investigación parte de los profesionales de laboratorio y es en sus publicaciones donde observaríamos una mayor mejora. Además, esta elección nos ha permitido analizar qué aspectos de la guía metodológica que aplicamos aparecen reflejados en una mayor proporción en estas revistas de laboratorio y cuáles en publicaciones de carácter principalmente clínico (*Revista Clínica Española* y *Medicina Clínica (Barcelona)*).

Otra de las dificultades aparecidas en este estudio, se refiere a la variabilidad de los investigadores que hemos participado en la evaluación de los artículos. Pensamos que el trabajo realizado consistente en diversas reuniones para tratar las posibles discrepancias que surgían, así como el análisis de cada artículo por al menos dos investigadores independientes, y que en caso de discordancia era analizado por un tercero, ha reducido esta variabilidad, obteniendo una concordancia superior al 80%.

Una vez finalizado el trabajo, nos planteamos futuras líneas de investigación dentro del campo del diagnóstico de laboratorio.

La iniciativa STARD [32], que elabora una guía compuesta por 25 criterios metodológicos que deben seguir los estudios de diagnóstico, ha aparecido publicada en diversas revistas en enero del año 2003. Además, están surgiendo otras herramientas para la evaluación diagnóstica, como QUADAS [173], que establece 14 criterios necesarios para la obtención de una adecuada calidad diagnóstica en los artículos de exactitud diagnóstica que van a formar parte de revisiones sistemáticas. Debido a la amplia difusión de estos nuevos criterios, es necesario aportar más información para validar su aplicabilidad y utilidad en los artículos que evalúen pruebas de diagnóstico. Este proceso de validación debería incluir la aplicación de estos ítems metodológicos a una muestra de estudios de diagnóstico, tal y como se ha hecho en este trabajo, centrándose sobre todo en la medida de la consistencia y fiabilidad de los criterios.

Por otro lado, la práctica clínica se está viendo transformada por la rápida expansión de nuevas y sofisticadas tecnologías que ofrecen muchas opciones para las nuevas evaluaciones diagnósticas. Hay que destacar el gran avance en el conocimiento genético de la enfermedad, que se traduce en el desarrollo de nuevos tipos de pruebas diagnósticas, que necesitan estudiar su validez y utilidad antes de aplicarse a la clínica.

Aunque como hemos estado viendo en este trabajo, varios grupos han tratado las pautas para la evaluación de áreas del laboratorio clínico, no existen recomendaciones para cubrir el espectro de los estudios epidemiológicos del genoma humano. Por lo tanto, debido a la ausencia de una revisión de calidad, es necesario iniciar un estudio que permita la correcta aplicación de las pruebas genéticas y moleculares a la práctica clínica. Para ello, va a ser necesario determinar la idoneidad de los criterios existentes, y definir otros más adecuados para el análisis de pruebas genéticas y moleculares [174, 175].

En la segunda parte de este trabajo hemos realizado un estudio de variabilidad observacional cuando se analizan muestras de líquido sinovial para la detección e identificación de cristales de urato monosódico y pirofosfato cálcico dihidratado.

Hemos podido comprobar la elevada consistencia en las observaciones de los analistas participantes que han recibido una formación adecuada, cuando examinan las muestras de líquido sinovial mediante el uso sucesivo del microscopio óptico y polarizado. Las guías básicas de análisis de cristales en dos pasos, detección e identificación de los mismos, han demostrado ser válidas para la realización de la técnica.

La puesta en práctica de los criterios metodológicos de Feinstein [11] ha demostrado su idoneidad para la realización de un estudio de concordancia, aunque con algunas dificultades en su aplicación: los aspectos en los que más problemas se ha podido detectar, son fundamentalmente la elección de muestras adecuadas y el establecimiento del grado de formación apropiado de los observadores participantes.

Uno de los problemas principales que podemos encontrar en un estudio de variabilidad observacional, se centra en elegir una muestra idónea para la formación del observador, pero que a su vez refleje la realidad clínica. Aunque para la formación del analista ha sido necesario contar con un número similar de muestras de líquido sinovial sin cristales, con cristales de urato monosódico y con cristales de pirofosfato cálcico, para así obtener un parecido grado de experiencia en la determinación de todos ellos, estas muestras no suelen representar la práctica diaria, donde lo normal es obtener muestras no patológicas. Por lo tanto, surge la dificultad de determinar qué tipo de muestras son las idóneas para el periodo de instrucción. Las muestras aconsejables deberían ser aquellas de las “zonas grises” del diagnóstico; es decir, aquellas en las que el diagnóstico no puede ser claramente establecido. De hecho, resultaría útil para conocer el grado de experimentación de cada observador, cuantificar el número de aciertos que ha tenido en este tipo de muestras. En otro tipo de estudios de variabilidad observacional donde las muestras pueden ser guardadas durante tiempo, como pueden ser muestras histológicas, sería conveniente recopilar las procedentes de estas series representativas para utilizarlas con fines instructivos.

La otra dificultad se centra en la determinación del grado de formación de los cuatro analistas, ya que aunque todos reciben el mismo curso formativo, no todos alcanzan el mismo nivel de conocimientos. Cada uno adquiere distinta experiencia, sobre todo en función del mayor o menor número de muestras que analiza a lo largo del estudio, y puede ser complejo el precisar cuando un observador tiene el suficiente grado de práctica como para realizar una prueba adecuadamente. En este trabajo ha sido útil la comprobación, una vez analizada una serie de muestras, de los resultados obtenidos por los analistas. Pensamos que la principal fuente de la variabilidad obtenida en este estudio puede ser este distinto grado de instrucción de los observadores.

Respecto a los valores obtenidos en los cálculos de exactitud diagnóstica, hay que señalar que el concepto de detección de cristales (es decir si los cristales de pirofosfato cálcico dihidratado o de urato monosódico están presentes en la muestra de líquido sinovial que estamos analizando, su presencia no va a pasar desapercibida) no se ha utilizado previamente, puesto que hasta ahora la detección e identificación se hacían de manera simultánea considerándose un proceso de rutina. Por esto, no podemos comparar los resultados de nuestro trabajo en la detección de cristales con otros estudios previos, pero nuestros valores de sensibilidad y especificidad para la detección de cristales son elevados, lo que apoya la utilización de la búsqueda de cristales en dos pasos.

Nuestra sensibilidad y especificidad para la identificación de cristales han sido más elevadas que las encontradas en estudios anteriores [43][46]. Hay que añadir que en estos estudios previos no se menciona la experiencia y el nivel de entrenamiento en el análisis de cristales de los observadores participantes, uno de los requisitos imprescindibles para la consecución de un estudio de variabilidad observacional, según la guía metodológica de Feinstein [11].

Tenemos algunos resultados falsos positivos, pero son más bajos que aquellos divulgados previamente [45][47] y hay que destacar que es en los cristales de pirofosfato cálcico dihidratado donde se encuentra una mayor dificultad. Cuando utilizamos el microscopio de luz polarizada, se observan los cristales únicamente en el 53% de las muestras que los contienen. Por el contrario el microscopio de luz ordinaria detectó los cristales en el 95% de las muestras de líquido sinovial analizadas. Estos datos están en armonía con los datos difundidos previamente y confirman la

frecuente ausencia de birrefringencia en los cristales de pirofosfato cálcico dihidratado cuando los observamos bajo microscopio de luz polarizada [66].

Estos resultados se ven apoyados con el estudio de otras características de nuestro estudio, como son el mayor tiempo empleado en la detección e identificación de cristales en el empleo del microscopio de luz polarizada para el análisis de cristales de pirofosfato cálcico dihidratado, ya que la mayoría no son birrefringentes, y el menor tiempo en el análisis con el mismo tipo de luz de cristales de urato monosódico, ya que todos ellos muestran una birrefringencia intensa. Así mismo tendremos que observar un menor número de campos cuando empleamos el microscopio de luz polarizada para detectar un cristal de urato monosódico que un cristal de pirofosfato cálcico dihidratado.

Hemos querido contribuir con este estudio al desarrollo de guías para el análisis de cristales de urato monosódico y pirofosfato cálcico en muestras de líquido sinovial: el procedimiento de dos pasos seguido por nosotros en el análisis de cristales (primero detección de los cristales y después identificación de los cristales detectados) puede ser determinante de los buenos resultados obtenidos.

Los cristales de pirofosfato cálcico dihidratado, donde los problemas en la detección e identificación encontrados en estudios previos son más comunes, son a menudo no birrefringentes [66] y consideramos que su detección se tiene que basar en la identificación morfológica bajo microscopio de luz ordinaria. Los cristales de urato monosódico se detectan muy bien con microscopio de luz polarizada no compensado, donde brillan intensamente en el campo oscuro [64], por lo que el microscopio de luz polarizada compensada permite la identificación adecuada de estos cristales una vez detectados. Debe tenerse presente que con raras excepciones, los cristales responsables de artritis pertenecen solamente a estos dos tipos, urato monosódico y pirofosfato cálcico dihidratado [61].

Detectamos bien por lo tanto, las muestras de líquido sinovial que contienen cristales de urato monosódico según lo publicado también previamente y aunque en nuestros resultados la tasa de detección es ligeramente más alta con luz ordinaria, pensamos que la luz polarizada permite una detección más fácil y debe ser utilizada de manera rutinaria para este propósito.

Hay que tener en cuenta que, la identificación final de los cristales detectados fue anotada solamente después de la observación sucesiva con el microscopio de luz ordinaria y el de luz polarizada compensado. Sin embargo, en la mayoría de las ocasiones después de la observación con la luz ordinaria, ya se conoce el tipo de cristal presente en nuestra muestra: por su forma romboidal o paralelepípeda los cristales de pirofosfato cálcico dihidratado son fácilmente identificados y solamente los cristales de pirofosfato cálcico dihidratado aciculares se pueden confundir por cristales de urato monosódico. Es aquí, donde el microscopio de luz polarizada compensado juega un papel definitivo [59]. Ningunas de las muestras incluidas en este estudio tenían de manera simultánea cristales de urato monosódico y de pirofosfato cálcico dihidratado.

Los artefactos constituyen un elemento importante de confusión para aquellos observadores que son inexpertos o con poca experiencia en el análisis de líquido sinovial.

Podemos concluir según nuestros resultados, que si no contamos en la práctica diaria con un microscopio de luz polarizada, un observador entrenado puede detectar correctamente los cristales de pirofosfato cálcico dihidratado y de urato monosódico por medio del microscopio de luz ordinaria, que generalmente suele ser el disponible [165].

Una vez analizados los datos obtenidos, es imprescindible la determinación de las limitaciones que ha tenido nuestro trabajo. Solo siendo conscientes de las dificultades en la consecución de los resultados, podemos mejorar y alcanzar una investigación de calidad.

Al interpretar nuestros resultados debe tenerse presente que, los observadores no tienen ninguna experiencia previa en el análisis de cristales en muestras de líquido sinovial y que realizaron las observaciones que constituyen la base de este estudio solamente después de un período corto de entrenamiento. Pensamos que con experiencia adicional los resultados habrían sido mejores. De hecho, la mayoría de los errores ocurrieron durante las observaciones iniciales.

No todas las muestras fueron analizadas por los cuatro observadores debido a diversos problemas de disponibilidad de los mismos, ya que las determinaciones se llevaron a cabo durante el

horario laboral. Por lo tanto se puede pensar que alguno de los analistas adquirieron más experiencia que otros, aunque todos obtuvieron unos valores similares de concordancia en la detección e identificación de cristales.

Hay que tener en cuenta a la hora de realizar estudios de este tipo, la dificultad añadida que conllevan: el poner en marcha un análisis de concordancia entre distintos observadores, implica un esfuerzo para dichos participantes. Por esto, el investigador principal debe ser el encargado de proveer las muestras y facilitar el análisis de las mismas, de tal manera que resulte lo más cómodo posible a los observadores.

Por último señalar que, hubiéramos querido calcular la sensibilidad y especificidad obtenida con el uso del microscopio de luz polarizada y con el uso del microscopio de luz ordinaria, para saber qué tipo de herramienta es más exacta en la detección e identificación de cristales en muestras de líquido sinovial. Pero hay que tener en cuenta que aunque de manera teórica se explican como dos pasos separados, no son independientes: es decir, los observadores analizan la preparación con el microscopio de luz polarizada una vez que la han examinado con el microscopio de luz ordinaria. Por lo que pierde el sentido el cálculo de sensibilidad y especificidad en la identificación de los cristales.

Después de obtener esta alta concordancia en la determinación de cristales en muestras de líquido sinovial por los analistas participantes, podemos constatar la utilidad de realizar un estudio de variabilidad observacional. Esto nos lleva a señalar que en futuras investigaciones, antes de la utilización de una prueba diagnóstica en la clínica, se realicen junto con los análisis habituales, un estudio de reproducibilidad de la técnica. Además, si precisa de la interpretación por parte del analista, se hace imprescindible un estudio de variabilidad observacional. Sobre todo es importante incidir cuando se realiza un estudio de este tipo, en el grado de formación de los observadores participantes, para que adquieran los conocimientos suficientes.

La obtención de estos buenos resultados en una prueba con resultados previos controvertidos pero de gran utilidad diagnóstica, nos indica que la detección de cristales de urato monosódico y pirofosfato cálcico en muestras de líquido sinovial constituye una técnica apropiada para el

diagnóstico de gota y de artropatía relacionada con el depósito de cristales de pirofosfato. Hay que resaltar que, aunque actualmente los programas de control de calidad existentes supervisan la mayoría de las pruebas de laboratorio, el interés demostrado en esta prueba por los laboratorios clínicos ha sido bajo y no se ha establecido todavía como un procedimiento de rutina. Es necesario, por lo tanto, que se inicie un programa de control de calidad, que idealmente tendría que incorporar la formación de los observadores y el método de análisis.

Quedan algunos aspectos referentes al estudio del líquido sinovial que quedan por estudiar y pueden ser transcendentes para el diagnóstico de las enfermedades reumatológicas. Como he reflejado en este trabajo, la mayor dificultad en un análisis de cristales de muestras de líquido sinovial, se encuentra en la detección e identificación de cristales de pirofosfato cálcico debido a su escasa birrefringencia. Por esto, se debe valorar si la utilización de un microscopio de contraste de fases puede ser más apropiado para el análisis de estos cristales. Además, no se ha relacionado de manera sistemática el recuento celular leucocitario con el grado de inflamación de las muestras de líquido sinovial. Este dato también es útil a la hora de diferenciar las distintas patologías reumatológicas.

8- CONCLUSIONES

1- No se han producido mejoras evidentes en la calidad metodológica de los artículos publicados en dos revistas nacionales de carácter clínico general, *Medicina Clínica* y *Revista Clínica Española*, cuando los comparamos con los resultados obtenidos en revisiones previas, por lo que todavía no se han debido incorporar las recomendaciones existentes para alcanzar una metodología de calidad en la investigación diagnóstica.

2- Los resultados obtenidos con el análisis de los artículos publicados en dos revistas internacionales de laboratorio analizadas, *Clinical Chemistry* y *Clinical Chemistry and Laboratory Medicine*, muestran una mejora significativa en la investigación diagnóstica a partir del año 2002, fecha en la cual se habrían incorporado ya las indicaciones mostradas por los editores en la consecución de una metodología adecuada.

3- Tanto en las revistas nacionales como en las internacionales, destaca un escaso cumplimiento de dos parámetros importantes en los estudios de exactitud diagnóstica: la descripción de la procedencia de los sujetos que forman parte del estudio y la descripción de los resultados indeterminados que han podido surgir en la investigación.

4- Las estrategias introducidas por editores para mejorar la calidad de la metodología diagnóstica han mostrado ser efectivas, tal y como se refleja en los mejores resultados obtenidos por *Clinical Chemistry*, revista de la que parten las principales recomendaciones. Esto, augura que la generación del STARD, así como de otras herramientas metodológicas como QUADAS, tendrán efectos positivos en aquellas publicaciones que consigan implantar estas guías a sus artículos.

5- La metodología de la investigación de pruebas diagnósticas de laboratorio, muestra deficiencias en aquellos aspectos clave que reflejan la escasa relación entre los Servicios de Análisis Clínicos y los servicios clínicos hospitalarios, lo que hace imprescindible una estrecha colaboración entre ellos.

6- Se constata la dificultad en la aplicación de los guías metodológicos y por lo tanto, la necesidad de acompañar cada criterio de una clara especificación para su aplicación.

7- Una adecuada instrucción de los analistas por un reumatólogo experto en la determinación e identificación de cristales de urato monosódico y pirofosfato cálcico en muestras de líquido sinovial, se ha traducido en la obtención de unos resultados consistentes, así como una sensibilidad y especificidad diagnóstica para la identificación del tipo de cristal superior a la mostrada en estudios previos.

8- Es importante que en el entrenamiento de los observadores se expliquen los dos pasos consecutivos pero independientes en que consiste el estudio: primero detección de los posibles cristales, y una vez detectados, identificación del tipo de cristal qué es.

9- Los puntos en los que más dificultad hemos encontrado en este estudio de variabilidad observacional, se refieren a la idoneidad de la selección de las muestras y a la afirmación de cuándo el grado de instrucción de los observadores es el adecuado para la realización de la técnica.

9- BIBLIOGRAFÍA

1. McQueen MJ. Overview of evidence-based medicine: challenges for evidence-based laboratory medicine. Clin Chem 2001;47:1536-46.

2. Trenti T. Evidence-based laboratory medicine as a tool for continuous professional improvement. Clin Chim Acta 2003;333:155-67.

3. Sackett DI, Richardson WS, Rosemberg W, Haynes RB. Medicina basada en la evidencia. Madrid: Churchill Livinstong España 1997.

4. Pozo F. La eficacia de las pruebas diagnósticas (I). Med Clin (Barc) 1988;90:779-85.

5. Reid M C, Lachs M, Feinstein R. Use of methodological standards in diagnostic test research. JAMA 1995; 274:645-651.

6. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, Bossuyt PM. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282:1061-6.

7. Ramos Rincón JM, Hernández Aguado I. Investigación sobre pruebas diagnósticas en Medicina Clínica. Valoración de la metodología. Med Clin (Barc) 1998;111:129-34.

8. Ramos JM, Hernández I. Métodos de valoración de pruebas diagnósticas en Enfermedades infecciosas y microbiología clínica. Enferm Infecc Microbiol Clinic 1998; 16: 179-184.

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

10. Hernández Aguado I, Ramos Rincón JM, Porta Serra M, Vioque López J, Rebagliato Ruso M, Bolumar Montrull F. La investigación sobre pruebas diagnósticas. Rev Clin Esp 1999;199:748-52.

11. Feinstein AR. Clinical Epideiology. The architecture of clinical research. Philadelphia; Saunders: 1985.

12. Bruns DE. The STARD initiative and the reporting of studies of diagnostic accuracy. Clin Chem. 2003;49:19-20.

13. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Moher D, Rennie D, de Vet HC, Lijmer JG; Standards for Reporting of Diagnostic Accuracy. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Clin Chem 2003;49:7-18.

14. Gallagher EJ. Evidence-based emergency medicine/editorial. The problem with sensitivity and specificity. Ann Emerg Med 2003;42:298-303.

15. Brown MD, Reeves MJ. Evidence-based emergency medicine/skills for evidence-based emergency care. Interval likelihood ratios: another advantage for the evidence-based diagnostician. Ann Emerg Med 2003;42:292-7.

16. Reid MC, Lane DA, Feinstein AR. Academic calculations versus clinical judgements: practicing physicians' use of cuantitative measures of test accuracy. Am J Med 1988; 104:374-80.

17. Lachs MS, Nachamkin I, Edelstein PH, Goldman J, Feinstein AR, Schwartz JS. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. Ann Intern Med 1992;117:135-40.

18. Choi BCK. Future challenges for diagnostic research: striking a balance between simplicity and complexity. J Epidemiol Community Health 2002;56:334-5.

19. Brenner H, Stürmer T, Gefeller O. The need for expanding and re-focusing of statistical approaches in diagnostic research. J Epidemiol Community Health 2002;56:338-9.

20. Knottnerus. Challenges in dia-prognostic research. J Epidemiol Community Health 2002;56:340-1.

21. Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC, Mosteller F. Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med. 1994;120:667-76.

22. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. JAMA. 1994;271:703-7.

23. Sackett DL, Haynes RB. The architecture of diagnosis research. BMJ 2002; 324:539-541.

24. Bruns DE. The Clinical Chemist. Clin Chem 1997;43:2211-2.

25. Bruns E, Huth J, Magid E., Young S. Towards a checklist for reporting of studies of diagnostic accuranct of medical tests. Clin Chem 2000;467:893-5.

26. Clarke M, Chalmers I. Discussion sections in reports of controlled trials published in general medical journals: islands in search of continents? JAMA 1998;280:280-2.

27. Huwiler-Muntener K, Juni P, Junker C, Egger M. Quality of reporting of randomized trials as a measure of methodologic quality. JAMA 2002;287:2801-4.

28. Egger M, Juni P, Bartlett C; CONSORT Group (Consolidated Standards of Reporting of Trials). Value of flow diagrams in reports of randomized controlled trials. JAMA 2001;285:1996-9.

29. Moher D. CONSORT: an evolving tool to help improve the quality of reports of randomized controlled trials. Consolidated Standards of Reporting Trials. JAMA 1998;279:1489-91.

30. Scherer RW, Crawley B. Reporting of randomized clinical trial descriptors and use of structured abstracts. JAMA. 1998;280:269-72.

31. Altman DG. Better reporting of randomised controlled trials: the CONSORT statement. BMJ 1996;313:570-1.

32. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HC; Standards for Reporting of Diagnostic Accuracy. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Standards for Reporting of Diagnostic Accuracy. Clin Chem 2003;49:1-6.

33. Knottnerus JA. The evidence Base of Clinical Diagnoses. Londres: BMJ Books; 2002.

34. Feinstein AR. A bibliography of publications on observer variability. J Chronic Dis:1985;38:619-32.

35. Elmore JG, Feinstein AR. A bibliography of publications on observer variability (final installment). J Clin Epidemiol. 1992;45:567-80.

36. Flessland KA, Landicho HR, Borden KK, Prince HE. Performance characteristics of the PolyTiter Immunofluorescent Titration system for determination of antinuclear antibody endpoint dilution. Clin Diagn Lab Immunol. 2002;9:329-32.

37. Cline BL, Habib M, Gamil F, Abdel-Aziz F, Little MD. Quality control for parasitologic data. Am J Trop Med Hyg. 2000;62:14-6.

38. Jen P, Woo B, Rosenthal PE, Bunn HF, Loscalzo A, Goldman L. The value of the peripheral blood smear in anemic inpatients. The laboratory's reading v a physician's reading. Arch Intern Med. 1983;143:1120-5.

39. Kirk CR, Burke H, Savage DC, Hughes AO Accuracy of home blood glucose monitoring by children. Br Med J (Clin Res Ed) 1986;293:17.

40. Murtomaa H, Meurman JH, Rantama A, Levo S Interexaminer variability in common ratings in reading Streptococcus mutans dip-slides with or without a microscope. Scand J Dent Res. 1987;95:144-50.

41. Levine RJ, Mathew RM, Brown MH, Hurtt ME, Bentley KS, Mohr KL, Working PK. Computer-assisted semen analysis: results vary across technicians who prepare videotapes. Fertil Steril. 1989;52:673-7.

42. Sanchez-Carrillo CI, Ramirez-Sanchez TJ, Zambrana-Castaneda M, Selwyn BJ. Test of a noninvasive instrument for measuring hemoglobin concentration. Int J Technol Assess Health Care. 1989;5:659-67

43. Schumacher HR Jr, Sieck MS, Rothfuss S, Clayburne GM, Baumgarten DF, Mochan BS, Kant JA.Reproducibility of sinovial fluid analysis. Arthritis Rheum 1986;29:770-774.

44. Hasselbacher P. Variation in synovial fluid analysis by hospital laboratories. Arthritis Rheum 1987;30:637-42.

45. Von Essen R, Holtta AM. Quality control of the laboratory diagnosis of gout by synovial fluid microscopy. Scand J Rheumatol 1990;19:232-4.

46. Gordon C, Swan A, Dieppe P. Detection of crystals in synovial fluid crystal by light microscopy: sensitivity and reliability. Ann Rheum Dis. 1989; 48: 737-42.

47. Von Essen R, Hölttä AMH, Pikkarainen R. Quality control of synovial fluid crystal identification. Ann Rheum Dis 1998;57:107-9.

48. Pascual E and Jovani V. A quantitative study of the phagocytosis of urate crystals in the synovial fluid of asymptomatic joints of patients with gout. Br J Rheumatol.1995;34:724-6.

49. Pascual E, Batlle-Gualda E, Martinez A, Rosas J, Vela P. Synovial fluid analysis for diagnosis of intercritical gout. Ann Intern Med 1999;131:756-9.

50. Lopez Redondo MJ, Requena L, Macia M, Schoendorff C, Sanchez Yus E, Robledo A. Fingertip tophi without gouty arthritis. Dermatology. 1993;187:140-143.

51. Schlesinger N., Baker D. G., Schumacher H.R. How well have diagnostic tests and therarpies for gout been evaluated? Curr Opin Rheumatol. 1999;11:441-5.

52. McCarty DJ, Hollander JL: Identification of urate crystals in gouty synovial fluid. Ann Intern Med 1961;54:452-460.

53. Pascual E. Persistence of monosodium urate crystals, and low grade inflammation, in the synovial fluid of untreated gout. Arthritis Rheum 1991;34:141-5.

54. Sturrock RD. Gout. Easy to misdiagnose. BMJ 2000;320:132-3.

55. Rosenthal AK. Formation of calcium pyrophosphate crystals: biologic implications. Curr Opin Rheumatol 2000;12:219-22.

56. Snaith ML. ABC of Rheumatology: gout, hyperuricaemia and crystal arthrtis. BMJ. 1995 Feb 25;310(6978):521-4.

57. Rull M. Artropatías por cristales. La Revista de Investigación Clínica 2000;52:451-60.

58. Swan A, Amer H, Dieppe P. The value of synovial fluid assays in the diagnosis of joint disease: a literature survey. Ann Rheum Dis 2002;61:493-8.

59. Pascual E. The diagnosis of gout and CPPD crystal arthropathy. Br J Rheum 1996; 35:306-8.

60. Schumacher HR, Reginato AJ. Atlas of synovial fluid analysis and crystal identification. Lea & Febiger, Filadelfia 1991.

61. Jaccard YB, Gerster JC, Calame L. Mixed monosodium urate and calcium pyrophosphate crystal-induced arthropathy. Rev Rhum Engl Ed. 1996;63:331-5.

62. Schumacher H R. Synovial fluid analysis. In: Kelley W N, Harris E D, Ruddy S, Sledge C B, eds. *Texbook of rheumatology*. 2nd ed. Philadelphia: Saunders, 1985:561-7.

63. Gatter R A*. A practical handbook of joint fluid analysis*. Philadelphia: Lea and Febiger, 1984: 37-49.

64. McCarty DJ, Hollander JL: Identification of urate crystals in gouty synovial fluid. Ann Intern Med 1961;54:452-460.

65. Kohn NN, Hughes RE, Mc Carthy DJ Jr, Faires JS. The significance of calcium pyrophosphate crystals in the synovial fluid of arthritis patients: the “pseudogout syndrome”. II. Identification of crystals. Ann Intern Med. 1962;56:738-45.

66. Ivorra J, Rosas J, Pascual E. Most calcium pyrophosphate crystals appear as non-birefringent. Ann Rheum Dis 1999;58:582-84.

67. Buntinx F, Schouten HJ, Knottnerus JA, Crebolder HF, Essed GG. Interobserver variation in the assessment of the sampling quality of cervical smears. J Clin Epidemiol. 1993 Apr;46(4):367-70.

68. D. Orozco, V. F. Gil, V. Pedrera, F. Buigues, E. Medina, J. Merino. Validez de la determinación de la glucemia basal en el control de los pacientes diabéticos no dependientes de insulina. Med Clin (Barc) 1997; 108: 325-329.

69. M. Papo, J. C. Quer, R. Pastor, anticuerpos anticitoplasma del neutrófilo en la enfermedad inflamatoria del intestino. Med Clin (Barc) 1998; 110: 11-15.

70. M del Val Gómez Martínez, F. G. Gallardo, J. Marrch, F. Laguna. Rastreo con galio-67 en la tuberculosis y las infecciones por Mycobacterium avium-M intracellulare en pacientes con infección por el VIH. Med Clin (Barc) 1998;110:570-573.

71. J. Niubò. J. L. Pérez, N. Manito, A. García. La reacción en cadena de la polimerasa como marcador de la infección por citomegalovirus en los receptores de un transplante cardíaco. Med Clin (Barc) 1999; 112: 121-124.

72. Angels Costa, Ignacio Conget, Ricart Treserras, Ramón Gomis. Utilidad de la glucemia basal y de la hemoglobina glucosilada para la detección de la tolerancia anormal a la glucosa en familiares de pacientes con diabetes tipo 2. Med Clin (Barc) 1999; 112:241-244.

73. V. G. Gil, E. Peinado, E. Obrador, R. Pascual. Validez de las pruebas diagnósticas para confirmar o descartar un infarto agudo de miocardio. Med Clin (Barc) 2000; 114; 11-13.

74. V. G. Gil, E. Peinado, E. Obrador, R. Pascual. Validez de las pruebas diagnósticas para confirmar o descartar una apendicitis aguda. Med Clin (Barc) 2000; 114; 48-51.

75. M. Sánchez Carbayo, M. Urritia, M. L. Hernández, J. M. González de Buitrago. Citoqueratinas (UBC y CIFRA 21-1) y proteínas de la matriz nuclear (NMP "") como marcadores tumorales en la orina en el diagnóstico del cáncer vesical. Med Clin (Barc) 2000; 114: 361-366.

76. M. Romero Gómez, J. Vargas, L. Grande. Utilidad de la detección de antígenos de Helicobacter pylori en heces en el diagnóstico de infección y en el control de la erradicación tras el tratamiento. Med Clin (Barc) 2000;114: 571-573.

77. J. Muñoz Méndez, I. Alfajeme, J. Hernández Borge. Enzima conversora de la angiotensina en tromboembolia pulmonar como un marcador de lesión vascular. Rev Clin Esp 1997;84-91.

78. M. Casal, J. Gutierrez, M. Vaquero. Interés clínico de un nuevo sistema simple para el aislamiento de Mycobacterium tuberculosis. Rev Clin Esp 1997; 197: 148-151.

79. E. García Pachón e I. Padilla Navas. Utilidad diagnóstica de la colinesterasa en los exudados pleurales. Rev Clin Esp 1997;197: 402-405.

80. J. Santo- Domingo, G. Rubio, J. J. Marin. Transferrina deficiente en carbohidratos y otros marcadores de consumo de alcohol en el hospital general. Rev Clin Esp 1997;197:627-630.

81. O. López Bartolomé, A. Morán Vasallo, J. A. Ramírez Armengol. Diagnóstico microbiológicode Helicobacter pylori y su resistencia a los antimicrobianos. Rev Clin Esp 1998; 198: 420-423.

82. A. García Enguídanos, E. Crespo Azanza, J. Diez Recasens. Valoración de las pruebas diagnósticas no invasivas frente al second look en el cáncer epitelial de ovario. Rev Clin Esp 1998; 502-505.

83. M. Casal, J. Gutierrez, M. Vaquero. Evaluación clínica de un nuevoo sistema autoomático no radiométrico para el diagnóstico rápido de tuberculosis. Rev Clin Esp 1998;198: 651-654.

84. F. Bermejo San José, D. Boixeda de Miguel, J. P. Gisbert. Eficacia de cuatro técnicas de amplio uso para el diagnóstico de la infección por Helicobacter pylori en la enfermedad ulcerosa gástrica. Rev Clin Esp 2000; 200: 475-479.

85. Schellenberg F, Martin M, Caces E, Benard JY, Weill J. Nephelometric determination of carbohydrate deficient transferrin. Clin Chem 1996;42:551–7.

86. Woitge HW, Seibel MJ, Ziegler R. Comparison of total and bone-specific alkaline phosphatase inpatients with non skeletal disorders or metabolic bone diseases. Clin Chem 1996;42:1796–804.

87. Gerard B, Peponnet C, Brunie G, Cave H, Denamur E, d'Auriol L, et al. Fluorometric detection of HIV-I genome through use of an internal control, inosine-substituted primers, and microtiter plate format. Clin Chem 1996;42:696–703.

88. Laurino JP, Bender EW, Kessimian N, Chang J, Pelletier T, Usategui M. Comparative sensitivities and specificities of the mass measurements of CK-MB2, CK-MB, and myoglobin for diagnosing acute myocardial infarction. Clin Chem 1996;42:1454–9.

89. Villena V, Navarro-Gonzalvez JA, Garcia-Benayas C, Manzanos JA, Echave J, Lopez-Encuentra A, et al. Rapid automated determination of adenosine deaminase and lysozyme for differentiating tuberculous and non tuberculous pleural effusion. Clin Chem 1996;42:218–21.

90. Sacchetti L, Ferrajolo A, Salerno G, Esposito P, Lofrano MM, Oriani G, et al. Diagnostic value of various serum antibodies detected by diverse methods in childhood celiac disease. Clin Chem 1996;42:1838–42.

91. Rohlfs EM, Chaing SH, Chapman JF. Analytical and clinical evaluation of refractive index-matched anomalous diffraction (RIMAD) for assessment of fetal lung maturation. Clin Chem 1996;42:1861–8.

92. Castaldo G, Intrieri M, Calcagno G, Cimino L, Budillon G, Sacchetti L, et al. Ascitic pseudiuridine discriminates between hepatocarcinoma-derived ascites and cirrhotic ascites. Clin Chem 1996;42:1843–6.

93. Marquet PY, Daver A, Sapin R, Bridgi B, Muratet JP, Hartmann DJ, et al. Highly sensitive immunoradiometric assay for serum thyroglobulin with minimal interference from autoantibodies. Clin Chem 1996;42:258–62.

94. Helander A, Beck O, Jones AW. Laboratory testing for recent alcohol consumption: comparison of ethanol, methanol, and 5-hydroxytryptophol. Clin Chem 1996;42:618–24.

95. Gue´chot J, Laudat A, Loria A, Serfaty L, Poupon R, Giboudeau J. Diagnostic accuracy of hyaluronan and type III procollagen aminoterminal peptide serum assays as markers of liver fibrosis in chronic viral hepatitis C evaluated by ROC curve analysis. Clin Chem 1996;42:558–63.

96. Bizzaro N, Mazzanti G, Tonutti E, Villalta D, Tozzoli R. Diagnostic accuracy of the anti-citrulline antibody assay for rheumatoid arthritis. Clin Chem 2001;47:1089–93.

97. Hayashi N, Kawamoto T, Mukai M, Morinobu A, Koshiba M, Kondo S, et al. Detection of antinuclear antibodies by use of an enzyme immunoassay with nuclear Hep-2 cell extract and

recombinant antigens: comparison with immunofluorescence assay in 307 patients. Clin Chem 2001;47:1649–59.

98. Norberg T, Klaar S, Lindqvist L, Lindahl T, Ahlgren J, Bergh J. Enzymatic mutation detection method evaluated for detection of p53 mutations in cDNA from breast cancers. Clin Chem 2001;47:821–8.

99. Christenson RH, Duh SH, Sanhai WR, Wu AH, Holtman V, Painter P, et al. Characteristics of an albumin cobalt binding test for assessment of acute coronary syndrome patients: a multicenter study. Clin Chem 2001;47:464–70.

100. Turpeinen U, Methuen T, Alfthan H, Laitinen K, Salaspuro M, Stenman UH. Comparison of HPLC and small column (CDTect) methods for disialotransferrin. Clin Chem 2001;47:1782–7.

101. Umapathysivam K, Hopwood JJ, Meikle PJ. Determination of acid_-glucosidase activity in bloodspots as a diagnostic test for Pompe disease. Clin Chem 2001;47:1378–83.

102. Cassinat B, Darsin D, Guardiola P, Toubert ME, Rain JD, Gluckman E, et al. Intermethod discordance for _-fetoprotein measurements in Fanconi anemia. Clin Chem 2001;47:1405–9.

103. Bijwaard KE, Aguilera NS, Monczak Y, Trudel M, Taubenberger JK, Lichy JH. Quantitative real-time reverse transcription PCR assay for cyclin DI expression: utility in the diagnosis of mantle cell lymphoma. Clin Chem 2001;47:195–201.

104. Rickert M, Seissler J, Dangel W, Lorenz H, Richter W. Fusion protein for combined analysis of autoantibodies to the 65-kDa isoform of glutamic acid decarboxylase and islet antigen-2 in insulin-dependent diabetes mellitus. Clin Chem 2001;47:926–34.

105. Jensen UG, Brandt NJ, Christensen E, Skovby F, Norgaard-Pedersen B, Simonsen H. Neonatal screening for galactosemia by quantitative analysis of hexose monophosphates using tandem mass spectrometry: a retrospective study. Clin Chem 2001;47:1364–72.

106. Lempinen M, Kyla¨npa¨a¨-Ba¨ck ML, Stenman UH, Puolakkainen P, Haapiainen R, Finne P, et al. Predicting the severity of acute pancreatitis by rapid measurement of trypsinogen-2 in urine. Clin Chem 2001;47:2103–7.

107. Nurmikko P, Pettersson K, Piironen T, Hugosson J, Lilja H. Discrimination of prostate cancer from benign disease by plasma measurement of intact, free prostate-specific antigen lacking an internal cleavage site at Lys145–Lys146. Clin Chem 2001;47:1415–23.

108. Sillanaukee P, Olsson U. Improved diagnostic classification of alcohol abusers by combining carbohydrate-deficient transferrin and glutamyltransferase. Clin Chem 2001;47:681–5.

109. Mourad M, Malaise J, Chaib Eddour D, De Meyer M, Ko¨nig J, Schepers R, et al. Pharmacokinetic basis for the efficient and safe use of low-dose mycophenolate mofetil in combination with tacrolimus in kidney transplantation. Clin Chem 2001;47:1241–8.

110. Anton RF, Dominic KC, Bigelow M, Westby C, CDTect Research Group. Comparison of Bio-Rad %CDT TIA and CDTect as laboratory markers of heavy alcohol use and their relationships with _-glutamyltransferase. Clin Chem 2001;47:1769–75.

111. Jonsson M, Carlson J, Jeppsson JO, Simonsson P. Computersupported detection of M-components and evaluation of immunoglobulinsafter capillary electrophoresis. Clin Chem 2001;47:110–7.

112. Andersen JM, Hedstro¨m J, Kemppainen E, Finne P, Poulakkainen P, Stenman UH. The ratio of trypsin-2-_1-antitrypsin to trypsinogen-1 discriminates biliary and alcohol-induced acute pancreatitis. Clin Chem 2001;47:231–6.

113. Richard S, Miossec V, Moreau JF, Taupin JL. Detection of oligoclonal immunoglobulins in cerebrospinal fluid by an immunofixation-peroxidase method. Clin Chem 2002;48:167–73.

114. Ferre N, Camps J, Prats E, Vilella E, Paul A, Figuera L, et al. Serum paraoxonase activity: a new additional test for the improved evaluation of chronic liver damage. Clin Chem 2002;48:261–8.

115. Weber LT, Shipkova M, Armstrong VW, Wagner N, Schutz E, Mehls O, et al. Comparison of the Emit immunoassay with HPLC for therapeutic drug monitoring of mycophenolic acid in pediatric renal-transplant recipients on mycophenolate mofetil therapy. Clin Chem 2002;48:517–25.

116. Filler G, Priem F, Lepage N, Sinha P, Vollmer I, Clark H, et al. Trace protein, cystatin C, _2-microglobulin, and creatinine compared for detecting impaired glomerular filtration rates in children. Clin Chem 2002;48:729–36.

117. Fantz CR, Powell C, Karon B, Parvin CA, Hankins K, Dayal M, et al. Assessment of the diagnostic accuracy of the TDx-FLM II to predict fetal lung maturity. Clin Chem 2002;48:761–5.

118. Orlando R, Mussap M, Plebani M, Piccoli P, De Martin S, Floreani M, et al. Diagnostic value of plasma cystatin C as a glomerular filtration marker in decompensated liver cirrhosis. Clin Chem 2002;48:850–8.

119. Wasmuth JC, Oliver y Minarro D, Homrighausen A, Leifeld L, Rockstroh JK, Sauerbruch T, et al. Phospholipid autoantibodies and the antiphospholipid antibody syndrome: diagnostic accuracy of 23 methods studied by variation in ROC curves with number of clinical manifestations. Clin Chem 2002;48:1004–10.

120. Xu SQ, He M, Yu HP, Wang XY, Tan XL, Lu B, et al. Bioluminescent method for detecting telomerase activity. Clin Chem 2002;48:1016–20.

121. Poon TC, Mok TS, Chan AT, Chan CM, Leong V, Tsui SH, et al. Quantification and utility of monosialylated-fetoprotein in the diagnosis of hepatocellular carcinoma with nondiagnostic serum total-fetoprotein. Clin Chem 2002;48:1021–7.

122. Martinez M, Espana F, Royo M, Alapont JM, Navarro S, Estelles A, et al. The proportion of prostate-specific antigen (PSA) complexed to 1-antichymotrypsin improves the discrimination between prostate cancer and benign prostatic hyperplasia in men with a total PSA of 10 to 30 _g/L. Clin Chem 2002;48:1251–6.

123. Stephan C, Cammann H, Semjonow A, Diamandis EP, Wymenga LF, Lein M, et al. Multicenter evaluation of an artificial neural network to increase the prostate cancer detection rate and reduce unnecessary biopsies. Clin Chem 2002;48:1279–87.

124. Li J, Zhang Z, Rosenzweig J, Wang YY, Chan DW. Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. Clin Chem 2002;48:1296–304.

125. Krauss T, Emons G, Kuhn W, Augustin HG. Predictive value of routine circulating soluble endothelial cell adhesion molecule measurements during pregnancy. Clin Chem 2002;48:1418–25.

126. Panteghini M, Cuccia C, Bonetti G, Giubbini R, Pagani F, Bonini E.Single-point cardiac troponin T at coronary care unit discharge after myocardial infarction correlates with infarct size and ejection fraction. Clin Chem 2002;48:1432–6.

127. Oda H, Shiina Y, Seiki K, Sato N, Eguchi N, Urade Y. Development and evaluation of a practical ELISA for human urinary lipocalin-type prostaglandin D synthase. Clin Chem 2002;48:1445–53.

128. Carroccio A, Vitale G, Di Prima L, Chifari N, Napoli S, La Russa C, et al. Comparison of anti-transglutaminase ELISAs and an antiendomysial antibody assay in the diagnosis of celiac disease: a prospective study. Clin Chem 2002;48:1546–50.

129. Qu Y, Adam BL, Yasui Y, Ward MD, Cazares LH, Schellhammer PF, et al. Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. Clin Chem 2002;48:835–43.

130. Kauppinen R, von und zu Fraunberg M. Molecular and biochemical studies of acute intermittent porphyria in 196 patients and their families. Clin Chem 2002;48:1891–900.

131. Schwartz GL, Chapman AB, Boerwinkle E, Kisabeth RM, Turner ST. Screening for primary aldosteronism: implications of an increased plasma aldosterone/renin ratio. Clin Chem 2002;48:1919–23.

132. Derhaschnig U, Laggner AN, Roggla M, Hirschl MM, Kapiotis S, Marsik C, et al. Evaluation of coagulation markers for early diagnosis of acute coronary syndromes in the emergency room. Clin Chem 2002;48:1924–30.

133. Kulpa J, Wojcik E, Reinfuss M, Kolodziejski L. Carcinoembryonic antigen, squamous cell carcinoma antigen, CYFRA 21-1, and neuron-specific enolase in squamous cell lung cancer patients.Clin Chem 2002;48:1931–7.

134. Legros FJ, Nuyens V, Minet E, Emonts P, Boudjeltia KZ, Courbe A, et al. Carbohydrate-deficient transferrin isoforms measured by capillary zone electrophoresis for detection of alcohol abuse. Clin Chem 2002;48:2177–86.

135. Ryden I, Pahlsson P, Lindgren S. Diagnostic accuracy of 1-acid glycoprotein fucosylation for liver cirrhosis in patients undergoing hepatic biopsy. Clin Chem 2002;48:2195–201.

136. Penders J, Fiers T, Delanghe JR. Quantitative evaluation of urinalysis test strips. Clin Chem 2002;48:2236–41.

137. van Dalen A, Kessler AC. A multicentre evaluation of tumour marker determinations using the automatic Enzymum-test systems ES 300 and ES 600/700. Eur J Clin Chem Clin Biochem 1996;34:377–84.

138. Bacigalupo MA, Bazzini P, Farina L, Ius A. Evaluation of three immunoassays for detection of toxoplasma-specific immunoglobulin G and M. Eur J Clin Chem Clin Biochem 1996;34:503–5.

139. Matsushita H, Xu J, Kuroki M, Kondo A, Inoue E, Teramura Y, et al. Establishment and evaluation of a new chemiluminescent enzyme immunoassay for carcinoembryonic antigen adapted to the fully automated ACCES System. Eur J Clin Chem Clin Biochem 1996;34:829–35.

140. Schmitt UM, Stieber P, Hasholzner U, Pahl H, Hofmann K, Fateh-Moghadam A. Methodological and clinical evaluation of two automated enzymatic immunoassays as compared with a radioimmunoassays for neuron-specific enolase. Eur J Clin Chem Clin Biochem 1996;34:679–82.

141. Conejo JR, Benedito JE, Jime´nez A, Menche´n M, Cano J, Granizo V, et al. Diagnostic value of three tumour markers determined in pleural effusions. Eur J Clin Chem Clin Biochem 1996;34:139–42.

142. Plebani M, Borghesan F, Bernardi D, Faggian D. Clinical evaluation of a new quantitative method for specific IgE antibodies. Eur J Clin Chem Clin Biochem 1996;34:579–84.

143. Collinson P, Gerhardt W, Katus HA, Mu¨ller-Bardorff M, Braun S, Schricke U, et al. Multicentre evaluation of an immunological rapid test for the detection of troponin T in whole blood samples. Eur J Clin Chem Clin Biochem 1996;34:591–8.

144. Mroczko B, Szmitkowski M, Niklinski J. Granulocyte-colony stimulating factor and macrophage-colony stimulating factor in patients with non-small cell lung cancer. Clin Chem Lab Med 2001;39:374–9.

145. von Eyben FE, Blaabjerg O, Hyltoft-Petersen P, Madsen EL, Amato R, Liu F, et al. Serum lactate dehydrogenase isoenzyme 1 and prediction of death in patients with metastatic testicular germ cell tumors. Clin Chem Lab Med 2001;39:38–44.

146. Shaarawy M, Salem ME. Clinical value of microtransferrinuria and microalbuminuria in the prediction of preeclampsia. Clin Chem Lab Med 2001;39:29–34.

147. Genser B, Truschnig-Wilders M, Stunzner D, Landini MP, Halwachs-Baumann G. Evaluation of five commercial enzyme immunoassays for the detection of human cytomegalovirus-specific IgM antibodies in the absence of a commercially available gold standard. Clin Chem Lab Med 2001;39:62–70.

148. Tello FL, Prats CH, González MD. Free and complexed prostatespecific antigen (PSA) in the early detection of prostate cancer. Clin Chem Lab Med 2001;39:116–20.

149. Glojnaric I, Casl MT, Simic D, Lukac J. Serum amyloid A protein (SAA) in colorectal carcinoma. Clin Chem Lab Med 2001;39:129–33.

150. Mello G, Parretti E, Ognibene A, Mecacci F, Cioni R, Scarselli G, et al. Prediction of the development of pregnancy induced hypertensive disorders in high risk pregnant women by artificial neural networks. Clin Chem Lab Med 2001;39:801–5.

151. Schellenberg F, Mennetrey L, Bacq Y, Pages JC. Carbohydratedeficient transferrin (CDT)determination by nephelometry using a commercial kit. Analytical and diagnostic aspects. Clin Chem Lab Med 2001;39:866–71.

152. Sarno M, Sarno L, Baylink D, Drinkwater B, Farley S, Kleerekoper M, et al. Excretion of sweat and urine pyridinoline crosslinks in healthy controls and subjects with established metabolic bone disease. Clin Chem Lab Med 2001;39:223–8.

153. Siekmeier R, Bierlich A, Jaross W. The white blood cell differential: three methods compared. Clin Chem Lab Med 2001;39:432–45.

154. Giovanella L, Ceriani L, Giardina G, Bardelli D, Tanzi F, Garancini S. Serum cytokeratin fragment 21.1 (CYFRA 21.1) as tumour marker for breast cancer: comparison with

carbohydrate antigen 15.3 (CA 15.3) and carcinoembryonic antigen (CEA). Clin Chem Lab Med 2002;40:298–303.

155. Mroczko B, Szmitkowski M, Okulczyk B. Granulocyte-colony stimulating factor (G-CSF) and macrophage-colony stimulating factor (M-CSF) in colorectal cancer patients. Clin Chem Lab Med 2002;40:351–5.

156. Giovanella L, Ceriani L. High-sensitivity human thyroglobulin (hTG) immunoradiometric assay in the follow-up of patients with differentiated thyroid cancer. Clin Chem Lab Med 2002;40:480–4.

157. Kocna P, Vaný´ckova´ Z, Perus¡icova´ J, Dvora´k M. Tissue transglutaminase—serology markers for coeliac disease. Clin Chem Lab Med 2002;40:485–92.

158. Barlandas-Rendon E, Muller MM, Garcia-Latorre E, Heinschink A.Comparison of urine cell characteristics by flow cytometry and cytology in patients suspected of having bladder cancer. Clin Chem Lab Med 2002;40:817–23.

159. Linuma Y, Senda K, Takakura S, Ichiyama S, Tano M, Abe T, et al. Evaluation of a commercially available serologic assay for antibodies against tuberculosis-associated glycolipid antigen. Clin Chem Lab Med 2002;40:832–6.

160. Al-Daghistani HI, Abdel-Dayem M. Diagnostic value of various urine tests in the Jordanian population with urinary tract infection. Clin Chem Lab Med 2002;40:1048–51.

161. Hernando M, Gonzalez C, Sanchez A, Guevara P, Navajo JA, Papisch W, et al. Clinical evaluation of a new automated antidsDNA fluorescent immunoassay. Clin Chem Lab Med 2002;40:1056–60.

162. Burkhardt H, Bojarsky G, Gladisch R. Diagnostic efficiency of cystatin C and serum creatinine as markers of reduced glomerular filtration rate in the elderly. Clin Chem Lab Med 2002;40:1135–8.

163. Zago L, Dupraz H, Sarchi M, Rý´o M. The molar ratio of retinolbinding protein to transthyretin in the assessment of vitamin A status in adults. Proposal of a cut-off point. Clin Chem Lab Med 2002;40:1301–7.

164. Deville WL, Bezemer PD, Bouter LM. Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy. J Clin Epidemiol. 2000;53:65-9.

165. Pascual E, Tovar J, Ruiz MT. The ordinary light microscope: an apropiate tool for provisional detection and identification of crystals in synovial fluid. Ann Rheum Dis. 1989;48:983-5.

166. Dieppe P., Swan A. Identification of crystals in synovial fluid. Ann Rheum Dis 1999;58:261-3.

167. Reginato AJ, Reginato AM, Fernández - Dápica MP, Ramachandrula A. Familial calcium pyrophosphate crystal deposition disease or calcium pyrophosphate gout. Rev Rhum Engl Ed 1995;62:376-91.

168. Fam AG, MD. What is new about crystals other than monosodium urate? Curr Opin Rheumatol. 2000;12:228-34.

169. Salinas M., Rosas J., Iborra J., Manero H., Pascual E. Comparison of manual and automated cell counts in EDTA preserved synovial fluids. Storage has little influence on the results. Ann Rheum Dis 1997;56:622-6.

170. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HC; STARD group. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Clin Chem Lab Med. 2003;41:68-73.

171. Moher D, Jones A, Lepage L. CONSORT Group (Consolidated Standards for Reporting of Trials). Use of the CONSORT statement and quality of reports of randomised trial. A comparative beforeand-after evaluation. JAMA 2001;285:1912–5.

172. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, et al. The revised CONSORT statement for reporting randomised trials: explanation and elaboration. Ann Intern Med 2001;134:663–94.

173. 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.

174. Burke W, Atkins D, Gwinn M, Guttmacher A, Haddow J, Lau J, Palomaki G, Press N, Richards CS, Wideroff L, Wiesner GL. Genetic test evaluation : information needs of clinicians, policy markers and the public. Am J Epidemiol2002 ;156 :311-18.

175. Ransohoff DF. Rules of evidence for cancer molecular-marker discovery and validation. Nat Rev Cancer. 2004;4:309-14.

10- ANEXOS

10.1.- ANEXO I: PLANTILLA METODOLÓGICA DE REID ET AL [5].

PLANTILLA METODOLÓGICA DE REID ET AL [19]

Criterios de inclusión:

- Pruebas diagnósticas aplicadas al hombre Si No
- Test con utilidad clínica Si No
- Estudio de sensibilidad y especificidad o cociente de verosimilitud Si No

1- REPRODUCIBILIDAD Si No

1.2. Tipo de posible variabilidad Observador Instrumento

1.2.1. Técnica de valoración de la variabilidad del observador:

No Si-acuerdo Si-prueba kappa

1.2.2. Valoración de la variabilidad del instrumento Si No

1.2.2.1.Variación intra o interensayo:

No Si-baja Si-intermedia Si-alta

2- DESCRIPCIÓN DEL ESPECTRO DEL ESTUDIO Si (≥3) No (<3)

2.1. Edad Si No

2.2. Sexo Si No

2.3. Síntomas clínicos o estados de la enfermedad Si No

2.4. Criterios de selección y exclusión Si No

3- PREVENCIÓN DEL SESGO DE SECUENCIA O VERIFICACIÓN Si No

- Cohortes prospectivo Si No
- Cohortes retrospectivo Si No
- Casos y controles Si No

4- PREVENCIÓN DEL SESGO DE REVISIÓN CIEGA Si No

- Cohortes prospectivo Si No
- Cohortes retrospectivo Si No
- Casos y controles Si No

5- SE EMPLEA EL ERROR ESTANDAR O INTERVALOS DE CONFIANZA Si No

- Número de pacientes estudiados:

6- DESCRIPCIÓN DE RESULTADOS POR ESTRATOS Si No

7- PRESENTACIÓN DE RESULTADOS INDETERMINADOS Si No

- Valoración pertinente de su inclusión o exclusión Si No

10.2.- ANEXO II: PLANTILLA METODOLÓGICA DE CRITERIOS PROPIOS DE LOS AUTORES [7, 8]

PLANTILLA METODOLÓGICA DE CRITERIOS PROPIOS [21, 22]

A. OBJETIVOS

A.1. Se especifican los objetivos del estudio: Si No

A.2. Se justifica la valoración de la prueba: Si No

B. PATRÓN DE REFERENCIA

B.1: Se especifica: Si No

B.2. Se indica si es aceptado: Si No

B.3. Incorpora pruebas a valorar: Si No No especifica

B.4. Realizado en toda la serie: Si No No especifica

B.5. ¿Es necesario realizar PR de forma estandarizada?: Si No

C. MÉTODO DE REALIZACIÓN DE LA PRUEBA A VALORAR

C.1. Suficientemente descrito: Si Si con referencia No

C.2. Reproducibilidad: Si No No procede

C.3. Definición de término normal: Si No No procede

C.3.1. Justificación de la definición: Si No No procede

D. DISEÑO DEL ESTUDIO

D.1. Fase del estudio diagnóstico: 0 I II-III IV

D.2. Direccionalidad: Prospectivo Retrospectivo Incierto

D.3. Descripción de la fuente de la población a estudio: Si No Parcialmente

D.4. Se explican los criterios de selección: Si No Parcialmente

D.5. Se explican los criterios de exclusión: Si No Parcialmente

D.6. Descripción del espectro de sujetos: Si No Parcialmente

D.7. Sesgo de secuencia: Seguro Sospechoso Improbable

D.8. Comparación ciega: Si No No se especifica No procede

D.9. Secuencia: 1º PR 1º PD Simultánea Incierta

E. RESULTADOS

E.1. Se expresa: Sensibilidad solo Sensibilidad y Especificidad Sustitución adecuada

E.2. Curvas ROC o coeficiente de verosimilitud: Si No Prescindible Innecesario

E.3. Índice para pruebas conjuntas: Si No Innecesario

E.4. Valor predictivo: Si No Innecesario

E.5. Medición del IC: Si-adecuado Si-inadecuado No

E.6. Empleo de índices confusos: Si No ¿Cuál?

E.7. Resultados por estratos: Si No Innecesario

F. CONCLUSIONES

F.1. Conclusiones: Correctas Alguna incorrección Incorrectas Sin conclusiones

F.2. Interpretación correcta de los índices de exactitud: Si No

G. ASPECTOS FORMALES

G.1. Empleo de términos confusos sin definirlos: Si No

G.2. Errores de cuantificación: Si No

10.3.- ANEXO III: PLANTILLA METODOLÓGICA CONJUNCIÓN DE CRITRIOS PROPIOS DE LOS AUTORES [7, 8] Y DE REID ET AL [5].

PLANTILLA METODOLÓGICA CONJUNCIÓN DE LOS CRITERIOS PROPIOS [21, 22] Y DE REID ET AL [19](*)

Criterios de inclusión:

- Pruebas diagnósticas aplicadas al hombre Si No
- Test con utilidad clínica Si No
- Estudio de sensibilidad y especificidad o cociente de verosimilitud Si No

1. OBJETIVOS

1.1. Se especifican los objetivos del estudio Si No
1.2. Se justifica la valoración de la prueba Si No

2. PATRÓN DE REFERENCIA

2.1. Se especifica Si No
2.2. Incorpora pruebas a valorar Si No
2.3. Realizado en toda la serie Si No

3. MÉTODO DE REALIZACIÓN DE LA PRUEBA A VALORAR

3.1. Suficientemente descrito Si No
3.2. Reproducibilidad (*) Si No
3.2.1. Tipo de posible variabilidad Observador Instrumento
3.2.1.1. Técnica de valoración de la variabilidad del observador:
No Si-acuerdo Si-prueba kappa
3.2.1.2. Valoración de la variabilidad del instrumento Si No
3.2.1.2.1.Variación intra o interensayo:
No Si-baja Si-intermedia Si-alta
3.3. Definición de término normal Si No No procede

4. DISEÑO DEL ESTUDIO

4.1. Descripción del origen de la población a estudio Si (≥2) No (<2)
4.1.1. Descripción de forma de acceso al centro asistencial Si No
4.1.2. La muestra a estudio refleja la población clínica Si No
4.1.3. Criterio de la selección final de la muestra Si No
4.2. Descripción del espectro del estudio (*) Si (≥3) No (<3)
4.2.1. Edad Si No
4.2.2. Sexo Si No
4.2.3. Síntomas clínicos o estados de la enfermedad Si No
4.2.4. Criterios de selección y exclusión Si No
4.3. Prevención del sesgo de secuencia o verificación (*) Si No
- Cohortes prospectivo Si No
- Cohortes retrospectivo Si No
- Casos y controles Si No

4.4. Prevención del sesgo de revisión o comparación ciega (*) Si No
- Cohortes prospectivo Si No
- Cohortes retrospectivo Si No
- Casos y controles Si No

4. RESULTADOS

5.1. Son susceptibles de expresarse de forma continua Si No
5.1.1.Se expresan los resultados como
Curvas ROC Coc. de verosimilitud Sensi y especi en puntos de corte
5.2. Se valora más de una prueba con mismo objetivo diagnóstico Si No
5.2.1. Indice para pruebas conjuntas Si No
5.3. Se calculan los valores predictivos Si No
5.4. Se emplea el error estándar o íntervalos de confianza (*) Si No
- Número de pacientes estudiados:

5.5. Descripción de resultados por estratos (*) Si No
5.6. Presentación de resultados indeterminados (*) Si No

10.4.- ANEXO IV: CRITERIOS DE STARD [25]

–CRITERIOS STARD [32]

TITULO / RESUMEN:

1- Identifica el artículo como un estudio de exactitud diagnóstica.

INTRODUCCIÓN.

2- Objetivos del estudio, tales como estimar la exactitud diagnóstica o la comparación de exactitud entre las pruebas o grupos participantes.

METODOS.

Participantes.

3- Población a estudio: Criterios de inclusión y exclusión y lugar (área geográfica) donde se recogen los datos.
4- Recogida de pacientes: sobre la base de sus síntomas, resultados de pruebas previas o por haber recibido la prueba a valorar o el patrón de referencia.
5- Muestras: ¿la población a estudio es una serie consecutiva? Si no, especifica cómo se seleccionan los pacientes.
6- Datos: la recogida de los datos se planifica antes de que se realizara la prueba y el PR (estudio prospectivo) o después (estudio retrospectivo).
7- PR y su definición.
8- Especificaciones técnicas incluyen cómo y cuándo se recogieron las medidas y/o otras
9- Definición de las unidades, y/o categorías de los resultados de las pruebas diagnósticas y el PR.
10- Número, entrenamiento y experiencia de las personas que realizan e interpretan los tests y el PR.
11- Revisión ciega.

Métodos estadísticos.

12- Métodos para calcular y comparar medidas de exactitud diagnóstica y los métodos estadísticos para cuantificar el error (95%).
13- Métodos para calcular la reproducibilidad de los tests.

RESULTADOS

14- Cuándo se ha realizado el estudio, incluyendo el inicio y final de la recogida de los datos.
15- Características clínicas y demográficas de la población a estudio (por ejemplo, sexo, edad, espectro de los síntomas, morbilidad, tratamientos, centros de recogida...)
16- Número de participantes que satisficieron el criterio de inclusión y que luego recibieron o no las pruebas diagnósticas y/o el PR; describir por qué los pacientes no recibieron cada prueba (diagrama de flujo).

17- Intervalo de tiempo desde la prueba diagnóstica y el PR y cualquier otro tratamiento administrado entre ellos.
18- Distribución de la severidad de la enfermedad (criterio que defina) en aquellos con la condición a estudio; otros diagnósticos en los participantes sin la condición a estudio.
19- Una tabla de los resultados de las pruebas diagnósticas (incluyendo valores indeterminados o perdidos) con los resultados del PR; para resultados continuos, distribución de los resultados de al prueba diagnóstica con los resultados del PR.
20- Cualquier efecto adverso de la realización de la prueba diagnóstica o el PR.
21- Estimación de la exactitud diagnóstica y medidas de error estadístico (por ejemplo intervalos de confianza del 95%).
22- Cómo se tratan los resultados indeterminados, los valores perdidos y otros resultados.
23- Estimación de la variabilidad de la exactitud diagnóstica entre los subgrupos de participantes, investigadores o centros.
24- Estimación de la reproducibilidad.

DISCUSION

25- Discusión de la aplicabilidad clínica de los hallazgos del estudio

10.5.- Anexo V: Lumbreras Lacarra B, Ramos-Rincon JM, Hernandez-Aguado I. Valoración de la Metodología en la investigación sobre pruebas diagnósticas de laboratorio en Revista Clínica Española y Medicina Clínica. Rev Clin Esp 2004;204:472-6.

COMUNICACIÓN BREVE

Valoración de la metodología en la investigación sobre pruebas diagnósticas de laboratorio en *Revista Clínica Española* y *Medicina Clínica*

B. Lumbreras Lacarra[a], J. M. Ramos Rincón[b] e I. Hernández Aguado[c]

[a] *Servicio de Análisis Clínicos. Hospital General Universitario de Alicante.* [b] *Servicio de Medicina Interna. Hospital General Universitario de Elche. Alicante.* [c] *Departamento de Salud Pública. Historia de la Ciencia y Ginecología. Universidad Miguel Hernández. Alicante.*

Fundamento. **La identificación de las limitaciones metodológicas de la investigación diagnóstica de pruebas de laboratorio es la base para guiar su perfeccionamiento y facilitar la aplicación de los resultados a las decisiones clínicas.**
Métodos. **Se analizaron 17 artículos publicados en *Revista Clínica Española* y *Medicina Clínica* de 1997 a 2000, en los que se determinaba la sensibilidad y especificidad de una prueba diagnóstica de laboratorio. Se aplicaron, de forma independiente por tres investigadores, unos criterios de calidad metodológica estandarizados.**
Resultados. **Se observó una alta frecuencia en el cumplimiento de algunos aspectos metodológicos clave: descripción adecuada del patrón de referencia (88%), de la prueba investigada (88%), de la muestra clínica a estudio (71%) o en la prevención del sesgo de secuencia (82%). La valoración de resultados indeterminados (12%), la reproducibilidad de la prueba (12%) o la descripción del origen de la población a estudio (6%) fue infrecuente.**
Conclusión. **Los defectos metodológicos de la investigación diagnóstica de laboratorio podrían evitarse mediante la colaboración con los servicios clínicos y modificaciones sencillas de los diseños de los estudios.**

PALABRAS CLAVE: pruebas diagnósticas, laboratorio, metodología.

Lumbreras Lacarra B, Ramos Rincón JM, Hernández Aguado I. Valoración de la metodología en la investigación sobre pruebas diagnósticas de laboratorio en Revista Clínica Española y Medicina Clínica. *Rev Clin Esp 2004;204(9):472-6.*

Assessment of research methology on diagnostic laboratory test in *Revista Clínica Española* and *Medicina Clínica*

Basis. **The identification of the methodological limitations of diagnostic research on laboratory tests is the basis in order to guide its enhancement and facilitate the application of the results to the clinical decisions.**
Methods. **Seventeen articles published in *Revista Clínica Española* and *Medicina Clínica* from 1997 to 2000 were analyzed, in which the sensitivity and specificity of a diagnostic laboratory test were determined. Some standardized criteria of methodological quality were independently applied by three investigators.**
Results. **A high frequency in the fulfillment of some key methodological aspects was observed: adequate description of the standard of reference (88%), of the test investigated (88%), of the clinical sample evaluated (71%), and of the prevention of the sequence bias (82%). The assessment of indeterminate results (12%), the test reproducibility (12%), or the description of the origin of the evaluated population (6%) were infrequent.**
Conclusion. **The methodological defects of diagnostic laboratory research could be prevented through the collaboration with clinical services and with simple modifications of the trial's designs.**

KEY WORDS: diagnostic test, laboratory, methodology.

Introducción

La evaluación rigurosa de las pruebas diagnósticas debería ser un paso obligado antes de su introducción a la práctica clínica [1]. Esta área de investigación es menos permeable a la mejora en el rigor científico a juzgar por los resultados de las revisiones de los trabajos publicados en las mejores revistas médicas, que han evidenciado serias limitaciones metodológicas [2-4]. El campo del laboratorio (bioquímica, microbiología, inmunología) reúne gran parte de la investigación diagnóstica; es por ello clave identificar las carencias metodológicas más frecuentes en este ámbito a fin de orientar la mejora de la calidad de la investigación, y por ende la aplicabilidad de sus resultados a la toma de decisiones clínicas.

Este estudio se diseñó con el propósito de describir aquellos aspectos de la investigación de técnicas de la-

Correspondencia: I. Hernández Aguado.
Departamento de Salud Pública, Historia de la Ciencia y Ginecología.
Campus de San Juan.
Facultad de Medicina de la Universidad Miguel Hernández.
Ctra. de Valencia, km 8,7.
03550 San Juan de Alicante (Alicante).
Correo electrónico: ihernandez@umh.es

Aceptado para su publicación el 16 de enero de 2003

boratorio más deficientes y que por tanto más comprometen su validez y aplicabilidad.

Material y métodos

Se ha realizado una valoración metodológica con estándares aceptados y empleados anteriormente de los estudios que versan sobre pruebas diagnósticas de laboratorio, publicados desde 1997 a 2000 en dos revistas clínicas de ámbito nacional incluidas en Science Citation Index: *Revista Clínica Española* y *Medicina Clínica*.

Selección de los artículos

Los artículos se identificaron mediante el sistema bibliográfico informático MEDLINE, con las palabras clave «sensibilidad y especificidad» (*sensitivity and specificity*) que debían aparecer en el título, resumen o MeSH. Se incluyeron todos aquellos artículos que comunicaron investigaciones sobre pruebas diagnósticas de laboratorio aplicadas a humanos con utilidad clínica.

Criterios calidad metodológica

Se aplicó a cada artículo una ficha metodológica (anexo 1) basada en la conjunción de los 7 estándares de Reid et al [2] y de la guía de calidad de Hernández y García [3]. En cuestiones comunes con el cuestionario de Reid et al se siguen las definiciones de éste, ya que se han aplicado conjuntamente. Se consideran 5 apartados:

1) Objetivos. Se considera especificado cuando de forma precisa se indica el problema diagnóstico y la prueba a investigar para solucionarlo, y justificado si los autores señalan explícitamente que se trata de una prueba nueva, de una usada anteriormente con aplicación novedosa o de una antigua con controversias en su aplicación.

2) Patrón de referencia. Se juzga especificado cuando se detalla en qué consiste la aplicación de esta prueba estándar, tanto en los sujetos como en los controles (en caso de que los haya). Se previene el sesgo de incorporación cuando en su definición no está incluido alguno de los resultados de la prueba diagnóstica a valorar. Se considera realizado en toda la serie cuando se especifica que se realiza de manera sistemática en todos los pacientes del estudio.

3) Método de realización de la prueba a valorar. En pruebas nuevas se considera descrito si se detalla lo suficiente como para poder reproducirse. Se mide la reproducibilidad de la prueba si se calcula la variabilidad interobservador o de la técnica.

4) Diseño del estudio. Se considera descrito el origen de la población a estudio cuando se cumplen dos de las siguientes tres condiciones: descripción de las formas de acceso de los pacientes, la muestra refleja la población clínica real y se explica el procedimiento de selección final de la muestra. Se cumple el criterio de descripción del espectro de los sujetos cuando se describen tres de las siguientes circunstancias: distribución por edad, por sexo, resumen de síntomas clínicos o estadios de la enfermedad y criterios de selección y exclusión. En la prevención del sesgo de secuencia o verificación diagnóstica en estudios de cohortes se acepta el criterio si a todos los sujetos se les practica la prueba diagnóstica y el patrón de referencia. En los estudios de casos y controles si la prueba diagnóstica precede a la prueba de referencia se cumple el estándar si la verificación de la enfermedad se realiza con independencia del resultado de la prueba diagnóstica; si la prueba de referencia precede a la diagnóstica el crédito se obtiene cuando los resultados de ésta están estratificados de acuerdo a los factores clínicos que induzcan a la realización de la prueba de referencia.

En la prevención del sesgo de revisión para estudios de cohortes prospectivos en los que el paciente recibe primero la prueba diagnóstica, el estándar es aceptado si la prueba de referencia se evalúa independientemente. En estudios prospectivos si la prueba de referencia precede a la diagnóstica y en las series de casos y controles, se cumple si aparece una afirmación de la independencia de la interpretación de los resultados.

5) Resultados. Si la prueba diagnóstica es susceptible de expresar sus resultados de forma continua o discreta en varias categorías se valora si se presentan los resultados en forma de curvas ROC, cociente de verosimilitud o cálculo de sensibilidad y especificidad en distintos puntos de corte. Cuando se evalúa más de una prueba con el mismo objetivo diagnóstico se valora si se presentan índices de exactitud para expresar los resultados conjuntamente. El cálculo del valor predictivo de una prueba es necesario cuando la muestra a estudio representa a la población clínica real y por tanto debe calcularse. Se consideraran que los resultados aparecen distribuidos por estratos cuando los índices de exactitud se expresan según distintos subgrupos clínicos o demográficos de la población. El criterio de la precisión de los resultados se cumple cuando se presenta la precisión estadística de las estimaciones de los índices de exactitud mediante el cálculo del intervalo de confianza o error estándar. Deben describirse todos los resultados tanto los positivos y negativos como los indefinidos, especificando si hay o no resultados indeterminados o dudosos y comunicar cómo se analizan.

Variabilidad de los observadores

Los trabajos fueron revisados independientemente por los tres evaluadores participantes en el estudio. A continuación se ponían en común los resultados para cada uno de los ítems, y en caso de discordancia en la respuesta se llegaba a un consenso. El porcentaje de concordancia simple fue del 81% (en los criterios de Reid et al se obtuvo un 86% de acuerdo simple y en los criterios propios un 73%).

La creación, gestión de la base de datos y los análisis estadísticos de los resultados se efectuaron con el programa EpiInfo 2000.

Resultados

Durante el período de estudio (de 1997 a 2000) se han publicado un total de 1.917 artículos en la revista *Medicina Clínica* y 949 trabajos en la *Revista Clínica Española*. De éstos, y tras aplicar los criterios de inclusión y exclusión previamente comentados, analizamos un total de 9 trabajos de la primera y 8 de la segunda revista.

Los procedimientos evaluados fueron de microbiología (7; 45%), seguido de bioquímica (5; 30%), inmunología, hormonas y hematología. El número medio de pacientes/muestras fue de 224, con un rango comprendido entre 33 y 1.022.

En la tabla 1 se muestran los resultados obtenidos. Los objetivos se especifican en todos los artículos, pero en sólo 8 (47%) se justifica la necesidad de evaluar la prueba. Los aspectos relacionados con el patrón de referencia y la prueba diagnóstica mostraron una calidad alta, particularmente en la especificación de sus descripciones. Sin embargo, pocos estudios (12%) analizaron la reproducibilidad de la prueba evaluada.

LUMBRERAS LACARRA B, ET AL. VALORACIÓN DE LA METODOLOGÍA EN LA INVESTIGACIÓN SOBRE PRUEBAS DIAGNÓSTICAS DE LABORATORIO EN *REVISTA CLÍNICA ESPAÑOLA* Y *MEDICINA CLÍNICA*

ANEXO 1

Criterios para la valoración de estudios sobre pruebas diagnósticas				
Criterios de inclusión (*)				
Pruebas diagnósticas aplicadas al hombre	Sí	No		
Pruebas con utilidad clínica	Sí	No		
Estudio de sensibilidad y especificidad o cociente de verosimilitud	Sí	No		
1) Objetivos				
Se especifican los objetivos del estudio	Sí	No		
Se justifica la valoración de la prueba	Sí	No		
2) Patrón de referencia				
Se especifica	Sí	No		
Incorpora pruebas a valorar	Sí	No		
Realizado en toda la serie	Sí	No		
3) Método de realización de la prueba a valorar				
Suficientemente descrito	Sí	No		
Reproducibilidad (*)	Sí	No		
Tipo de posible variabilidad	Observador	Instrumento		
Técnica de valoración de la variabilidad del observador:	No	Sí acuerdo	Sí prueba kappa	
Valoración de la variabilidad del instrumento	Sí	No		
Variación intra o interensayo	No	Sí baja	Sí intermedia	Sí alta
Definición de término normal	Sí	No	No procede	
4. Diseño del estudio				
Descripción del origen de la población a estudio	Sí (< 2)	No (< 2)		
Descripción de forma de acceso al centro asistencial	Sí	No		
La muestra a estudio refleja la población clínica	Sí	No		
Criterio de la selección final de la muestra	Sí	No		
Descripción del espectro del estudio (*)	Sí (3)	No (< 3)		
Edad	Sí	No		
Sexo	Sí	No		
Síntomas clínicos o estados de la enfermedad	Sí	No		
Criterios de selección y exclusión	Sí	No		
Prevención del sesgo de secuencia o verificación (*)	Sí	No		
Cohortes prospectivo	Sí	No		
Cohortes retrospectivo	Sí	No		
Casos y controles	Sí	No		
Prevención del sesgo de revisión o comparación ciega (*)	Sí	No		
Cohortes prospectivo	Sí	No		
Cohortes retrospectivo	Sí	No		
Casos y controles	Sí	No		
5) Resultados				
Son susceptibles de expresarse de forma continua o discreta en dos o más categorías	Sí	No		
Se expresan los resultados como	Curvas ROC	Coc. de verosimilitud	Sensibilidad y especificidad en puntos de corte	
Se valora más de una prueba con mismo objetivo diagnóstico	Sí	No		
Índice para pruebas conjuntas	Sí	No		
Se calculan los valores predictivos	Sí	No		
Se emplea el error estándar o intervalos de confianza (*)	Sí	No		
Número de pacientes estudiados:				
Descripción de resultados por estratos (*)	Sí	No		
Presentación de resultados indeterminados (*)	Sí	No		
Valoración pertinente de su inclusión o exclusión	Sí	No		

* Criterios de Reid MC, et al[2].

Del diseño del estudio destaca la adecuada descripción del espectro de pacientes (sobre todo los parámetros de sexo y edad) y la escasa especificación por parte de los autores del origen de la población a estudio.

En un elevado porcentaje de artículos (71%) se calcularon los valores predictivos correctamente, mientras que pocos de los trabajos emplearon curvas ROC o cocientes de verosimilitud por puntos de corte cuando era apropiado hacerlo. La especificación de los resultados indeterminados fue infrecuente.

Ninguno de los estudios cumplió los 7 estándares de Reid et al; sólo dos cumplieron 5 y la mayoría cumplían tres.

Al comparar los estándares metodológicos de Reid et al aplicados a los artículos analizados en esta revisión y los encontrados en las revisiones realizadas por nuestro grupo en la revista *Medicina Clínica* durante los años 1992-1995 y en la revista *Enfermedades Infecciosas y Microbiología Clínica* durante los años 1990-1996 (tabla 2) se puede comprobar una mejora en varios criterios metodológicos, destacando el cálculo de la precisión de los resultados. También se incrementó la frecuencia en la descripción de la composición del espectro y en la prevención del sesgo de secuencia.

Discusión

La investigación sobre pruebas diagnósticas de laboratorio tiene una calidad metodológica aceptable en España a juzgar por la calidad de los trabajos publicados en los últimos años en *Revista Clínica Española* y *Medicina Clínica*. En comparación con períodos

TABLA 1

Descripción del cumplimiento de los estándares propuestos por los autores y por Reid et al (+) en los 17 artículos analizados de las dos publicaciones: *Revista Clínica Española* y *Medicina Clínica*

Nº artículos	Total (%)
Objetivos	
Especificados	17 (100%)
Justificación de la prueba	8 (47%)
Patrón de referencia	
Especificado	15 (88%)
Incorporación de la prueba a valorar	6 (35%)
Realización en toda la serie	12 (71%)
Prueba diagnóstica	
Descripción suficiente	15 (88%)
Definición de la normalidad	14 (82%)
(+) Reproducibilidad	2 (12%)
Diseño del estudio	
Descripción del origen de la población	1 (6%)
(+) Composición del espectro	12 (71%)
Distribución por edad	12 (71%)
Distribución por sexo	13 (76%)
Síntomas clínicos/estadios de la enfermedad	10 (60%)
Criterios selección y exclusión	8 (47%)
(+) Prevención sesgo de secuencia	14 (82%)
(+) Prevención sesgo de revisión	3 (18%)
Resultados	
Curvas ROC o cocientes de verosimilitud o exactitud	
Por puntos de corte* (n = 12)	3 (18%)
Índice pruebas conjuntas** (n = 14)	3 (18%)
Cálculo de valores predictivos	12 (71%)
(+) Análisis por subgrupos	7 (41%)
(+) Precisión de los resultados	9 (53%)
(+) Resultados indeterminados	2 (12%)

*En pruebas susceptible de expresión en forma continua. **Para estudios que analizan más de una prueba con el mismo cometido diagnóstico.

TABLA 2

Comparación del cumplimiento de los criterios de REID et al en el estudio actual con los resultados obtenidos anteriormente en dos publicaciones nacionales: *Medicina Clínica* (1992-1995) y *Enfermedades Infecciosas y Microbiología Clínica* (1990-1996)

Nº artículos	Actual trabajo (n = 17) (%)	*Enferm Infecc Microbiol Clin* (1990-1996) (Ref 1) (n = 41) (%)	*Med Clin* (1992-1995) (Ref 2) (n = 42) (%)
Composición del espectro	12 (71%)	25 (56%)	21 (50%)
Distribución por edad	12 (71%)		22 (52%)
Distribución por sexo	13 (76%)		28 (67%)
Síntomas clínicos/estadios de la enfermedad	10 (60%)		26 (62%)
Criterios de selección y exclusión	8 (47%)		15 (36%)
Análisis por subgrupos	7 (41%)	24 (53%)	21 (50%)
Prevención sesgo de secuencia	14 (82%)	30 (67%)	29 (69%)
Prevención sesgo de revisión	3 (18%)	2 (6%)	15 (36%)
Precisión de los resultados	9 (53%)	1 (2%)	7 (17%)
Resultados indeterminados	2 (12%)	5 (11%)	5 (12%)
Reproducibilidad	2 (12%)	5 (11%)	8 (19%)

anteriores [3,4] y con lo observado en revistas clínicas internacionales de prestigio [2] se puede afirmar que hay avances significativos. Quedan aún algunas insuficiencias que dificultan la aplicación de los resultados de la investigación a la práctica clínica.

Entre los aspectos metodológicos que precisan de particular atención destaca la ausencia de justificación en la investigación de la prueba que se está realizando. Es preciso que los autores se esfuercen en ser más convincentes sobre la novedad que aporta su trabajo y que precisen mejor qué cometido clínico proponen para la prueba que investigan. Este esfuerzo redundará en una mayor relevancia de este campo de investigación. Se ha argumentado que la distancia que a veces hay entre los investigadores que valoran las pruebas diagnósticas o los profesionales que las realizan con los clínicos que toman decisiones puede motivar la falta de relevancia clínica de los objetivos de los estudios diagnósticos [5]. Esta distancia puede ser también la explicación a la escasa atención prestada en los trabajos revisados al origen de la población a estudio. Esto conlleva que el clínico no sepa si los resultados descritos en un artículo de una prueba determinada se pueden aplicar a la práctica diaria de una población cualquiera. La situación ideal en la valoración de una prueba diagnóstica es realizarla en una población similar a la que luego se pretende aplicar la prueba. Para valorar la comparabilidad entre la población clínica y los pacientes a los que aplicar los resultados, los estudios publicados deben informar sobre el ámbito sanitario en el que se realiza la investigación, describir los filtros asistenciales o formas de acceso de los pacientes al centro asistencial, explicar si la población a estudio es una muestra consecutiva aleatoria de aquella población clínica real en la que se pretende aplicar la prueba diagnóstica, y finalmente cuantificar la población candidata al estudio y comunicar cuántos de ellos fueron finalmente seleccionados para el mismo. La investigación diagnóstica en el laboratorio precisa de una mejor conexión con la clínica para que esta información esté disponible.

Algunas de las carencias metodológicas observadas pueden corregirse fácilmente. Nos referimos, por ejemplo, a la reproducibilidad de la prueba. Su medición es necesaria porque variaciones elevadas en los procedimientos de laboratorio o entre los observadores condicionan la relevancia del estudio de su exactitud. Las evaluaciones de pruebas diagnósticas deben incluir el estudio de su variabilidad. De igual forma es preciso que los investigadores eviten el sesgo de revisión asegurando la valoración independiente del patrón de referencia y la prueba que se evalúa.

La presentación de los resultados de una investigación debe ser lo suficientemente completa como para que su aplicación práctica se vea favorecida. Por ejemplo, cuando se evalúan varias pruebas con el mismo cometido diagnóstico los autores deben presentar la exactitud que tiene la interpretación conjunta de las pruebas. También es imprescindible presentar la exactitud por puntos de corte en pruebas cuyos resultados se expresan de forma continua. Estos defectos metodológicos detectados son de sencilla corrección, y por tanto cabe esperar una mejora si se consigue una mayor exigencia metodológica en la investigación diagnóstica.

La muestra analizada no es muy amplia, pero es un reflejo de la mejor investigación diagnóstica publicada recientemente en España y suficiente para confirmar, una vez comparados los resultados con las últimas revisiones que databan del año 1995, que se está apreciando una notable mejora en la calidad de este tipo de investigación.

Las guías metodológicas empleadas se han usado en revisiones previas y están, por tanto validadas para su uso. Hay otras guías disponibles, algunas ampliamente empleadas cuando se trata de valorar la aplicabilidad de los resultados de una investigación diagnóstica [6-7], y otras más adecuadas para evaluar la validez interna de los estudios [8]. Ninguna es completamente satisfactoria, lo que motivó el uso de un doble conjunto de criterios con ánimo de exhaustividad. Entendemos que los criterios empleados abracan las cuestiones claves de la metodología de investigación en diagnóstico.

Aunque son muchos y variados los retos de la investigación diagnóstica [9], la adecuación de los objetivos a las necesidades del conocimiento clínico y el rigor metodológico son imprescindibles para una mayor conexión entre la ciencia y el cuidado clínico y para un mejor aprovechamiento mediante el metaanálisis de los estudios diagnósticos [10-11]. La mejora observada en las dos revistas evaluadas es alentadora, particularmente por tratarse del laboratorio de análisis clínico, pero aún quedan pasos por dar para que esta investigación permita mejorar sistemáticamente la calidad de la atención clínica.

BIBLIOGRAFÍA

1. Hernández Aguado I, García García AM. La evaluación de pruebas diagnósticas en España. Revisiones en Salud Pública 1993;3:243-62.
2. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. JAMA 1995;274:645-51.
3. Ramos Rincón JM, Hernández Aguado I. Investigación sobre pruebas diagnósticas en Medicina Clínica. Valoración de la metodología. Med Clin (Barc) 1998;111:129-34.
4. Ramos JM, Hernández I. Métodos de valoración de pruebas diagnósticas en enfermedades infecciosas y microbiología clínica. Enferm Infecc Microbiol Clin 1998;16:179-84.
5. Hernández Aguado I. The winding road towards evidence based diagnoses. J Epidemiol Community Health 2002;56:323-5.
6. Jaeschke R, Guyatt GH, Sackett DL, for the Evidenced-based Medicine Working Group. Users' guides to the medical literature. III: how to use an article about a diagnostic test, A: are the results of the study valid? JAMA 1994;271:389-91.
7. Jaeschke R, Guyatt GH, Sackett DL, for the Evidenced-based Medicine Working Group. Users' guides to the medical literature. III: how to use an article about a diagnostic test, B: what are the results and will the help me in caring for my patients? JAMA 1994;271:703-7.
8. Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC, et al. Guidelines for meta-analysis evaluating diagnostic tests. Ann Intern Med 1994;120:667-70.
9. Feinstein AR. Misguided efforts and future challenges for research on «diagnostic tests». J Epidemiol Community Health 2002;56:330-2.
10. Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. J Clin Epidemiol 1995;48:119-30.
11. Reid MC, Lane DA, Feinstein AR. Academic calculations versus clinical judgements: practicing physicians' use of quantitative measures of test accuracy. Am J Med 1988;104:374-80.

Anexo 2

ARTÍCULOS ANALIZADOS

1. Orozco D, Gil VF, Pedrera V, Buigues F, Medina E, Merino J. Validez de la determinación de la glucemia basal en el control de los pacientes diabéticos no dependientes de insulina. Med Clin (Barc) 1997;108:325-9.
2. Papo M, Quer JC, Pastor R. Anticuerpos anticitoplasma del neutrófilo en la enfermedad inflamatoria del intestino. Med Clin (Barc) 1998;110:11-5.
3. Gómez Martínez del Val M, Gallardo FG, Marrch Laguna JF. Rastreo con galio-67 en la tuberculosis y las infecciones por *Mycobacterium avium*-M intracelulare en pacientes con infección por el VIH. Med Clin (Barc) 1998; 110:570-3.
4. Niubó J, Pérez JL, Manito N, García A. La reacción en cadena de la polimerasa como marcador de la infección por citomegalovirus en los receptores de un transplante cardiaco. Med Clin (Barc) 1999;112:121-4.
5. Costa A, Conget I, Treserras R, Gomis R. Utilidad de la glucemia basal y de la hemoglobina glucosilada para la detección de la tolerancia anormal a la glucosa en familiares de pacientes con diabetes tipo 2. Med Clin (Barc) 1999;112:241-4.
6. Gil VG, Peinado E, Obrador E, Pascual R. Validez de las pruebas diagnósticas para confirmar o descartar un infarto agudo de miocardio. Med Clin (Barc) 2000;114;11-3.
7. Gil VG, Peinado E, Obrador E, Pascual R. Validez de las pruebas diagnósticas para confirmar o descartar una apendicitis aguda. Med Clin (Barc) 2000;114;48-51.
8. Sánchez Carbayo M, Urrittia M, Hernández ML, González de Buitrago JM. Citoqueratinas (UBC y CIFRA 21-1) y proteínas de la matriz nuclear (NMP") como marcadores tumorales en la orina en el diagnóstico del cáncer vesical. Med Clin (Barc) 2000;114:361-6.
9. Romero Gómez M, Vargas J, Grande L. Utilidad de la detección de antígenos de *Helicobacter pylori* en heces en el diagnóstico de infección y en el control de la erradicación tras el tratamiento. Med Clin (Barc) 2000;114:571-3.
10. Muñoz Méndez J, Alfajeme I, Hernández Borge J. Enzima conversora de la angiotensina en tromboembolia pulmonar como un marcador de lesión vascular. Rev Clin Esp 1997;84-91.
11. Casal M, Gutiérrez J, Vaquero M. Interés clínico de un nuevo sistema simple para el aislamiento de *Mycobacterium tuberculosis*. Rev Clin Esp 1997;197:148-51.
12. García Pachón E, Padilla Navas I. Utilidad diagnóstica de la colinesterasa en los exudados pleurales. Rev Clin Esp 1997;197:402-5.
13. Santo-Domingo J, Rubio G, Martín JJ. Transferrina deficiente en carbohidratos y otros marcadores de consumo de alcohol en el hospital general. Rev Clin Esp 1997;197:627-30.
14. López Bartolomé O, Morán Vasallo A, Ramírez Armengol JA. Diagnóstico microbiológico de *Helicobacter pylori* y su resistencia a los antimicrobianos. Rev Clin Esp 1998;198:420-3.
15. García Enguídanos A, Crespo Azanza E, Díez Recasens J. Valoración de las pruebas diagnósticas no invasivas frente al *second look* en el cáncer epitelial de ovario. Rev Clin Esp 1998;502-5.
16. Casal M, Gutiérrez J, Vaquero M. Evaluación clínica de un nuevo sistema automático no radiométrico para el diagnóstico rápido de tuberculosis. Rev Clin Esp 1998;198:651-4.
17. Bermejo San José F, Bobxeda de Miguel D, Gisbert JP. Eficacia de cuatro técnicas de amplio uso para el diagnóstico de la infección por *Helicobacter pylori* en la enfermedad ulcerosa gástrica. Rev Clin Esp 2000; 200:475-9.

10.6.- Anexo VI: Editorial. Clin Chem. 2004;50:465-6.

Editorial

The Quality of Reporting in Diagnostic Test Research: Getting Better, Still Not Optimal

In this era of evidence-based medicine, clinicians and other decision-makers turn to the scientific literature for high-quality evidence about the usefulness, precision, and accuracy of diagnostic tests. Such evidence is needed more than ever because the list of diagnostic tests is growing exponentially, and even more biomarkers, proteomics, and applications of gene expression profiling will be added in the years to come.

Studies of diagnostic accuracy can provide the necessary data. Rigorous methodologic standards in research about diagnostic tests have been slower to develop than standards for therapy studies. During the past decade, our knowledge about study design features that are associated with bias and lack of applicability in diagnostic studies has grown. Diagnostic studies with deficiencies in specific design features have been shown to be associated with biased, optimistic estimates of diagnostic accuracy compared with studies without such deficiencies *(1)*.

Given this potential for bias, it is of paramount importance that study reports include a proper description of the study methods, in particular those design features that have been most clearly associated with bias. Within the spirit of evidence-based medicine, readers should take these features into account when examining a study and its results, pondering the decision on whether to implement changes in practice based on conclusions from the study.

Unfortunately, the quality of reporting of studies of diagnostic accuracy is often poor, making judgments about validity, bias, and applicability to patients in clinical settings difficult. The poor quality of reporting has been documented and lamented many times. One of the best known examples is a study by Reid et al. *(2)*, who showed that even in high-impact medical journals many design features and patient characteristics were not sufficiently described.

This issue of the Journal contains an article by Lumbreras-Lacarra et al. *(3)* that shows the results of a study on the quality of reporting in *Clinical Chemistry* and *Clinical Chemistry and Laboratory Medicine*. For papers published in 1996, the results in terms of completeness of reporting were found to be comparable to those of the Reid study. This means that *Clinical Chemistry* and *Clinical Chemistry and Laboratory Medicine* performed as well—or as badly—as the general clinical journals in that same time window. No less than 14 of 18 papers did not describe the eligibility criteria for the study, a similar number failed to report the avoidance of review bias, and only 4 reported measures of imprecision.

Over the past decades many journals have recognized the importance of quality of reporting, and several initiatives have been developed to improve the efficiency of scientific communication. This evolution has brought us the structured paper, the "Vancouver" Uniform Requirements for Manuscripts Submitted to Biomedical Journals *(4)*, and the structured abstract *(5)*. A more recent phenomenon has been the development of checklists for specific study types. The first was the development of the Consolidated Standards of Reporting Trials (CONSORT) *(6)*. This is a single-page checklist to be used by authors and reviewers to guarantee that essential features of the design and results of randomized clinical trials are well reported. The CONSORT statement has been adopted by the International Committee of Medical Journal Editors, the Council of Science Editors, and the World Association of Medical Editors.

In 1996, a group of clinical chemists developed a list of items to include in a checklist for studies of diagnostic accuracy. This list was first published in *Clinical Chemistry* for comment, and an amended version was published in the Journal in 2000 and incorporated in the Information for Authors *(7)*. The checklist has been used in the review of most studies of diagnostic accuracy published in the Journal during 2001 and 2002 (personal communication, David E. Bruns, University of Virginia Medical Center, Charlottesville, VA). In the first issue of 2003, this Journal published a more general checklist, the Standards for the Reporting of Diagnostic Accuracy studies (STARD) *(8, 9)*. This same STARD statement has been published in more than a dozen other international journals, accompanied in many of them by editorials and by the editorial decision to include STARD into the instructions for authors.

Does a checklist make a difference? Papers in journals that promoted CONSORT (*BMJ*, *JAMA*, and *Lancet*) showed greater improvement in quality of reporting than did papers in a journal that did not advocate its use (*New England Journal of Medicine*) *(10)*. Many feel that the reporting of diagnostic studies has also improved, although some of this may be attributable to a growing sophistication among researchers in general. The study by Lumbreras-Lacarra et al. *(3)* provides additional evidence to suggest a role for checklists. In addition to their analysis of the papers published in 1996, these authors present an overview of papers published in *Clinical Chemistry* and in *Clinical Chemistry and Laboratory Medicine* in 2001 and 2002. Although the numbers are small, their analysis shows clear signs of improvement between 1996 and 2002 for *Clinical Chemistry*, with the average number of desired features increasing from two to four. A similar improvement could not yet be observed for *Clinical Chemistry and Laboratory Medicine*, which published the STARD statement this year, but did not use a similar checklist before *(11)*.

Although these are encouraging results, referees and editors of journals need to do a better job in ensuring that their authors implement the STARD recommendations. For example, journals still receive—and publish—papers that fail to report on the sex and age distributions of the study participants, papers that do not report on the eligibility criteria, and papers that do a poor job in

describing how and when the diagnosis was verified. All of this matters in judging both the potential for bias and the applicability of study findings.

On a previous occasion, the editor of this Journal pointed out that a checklist works primarily by reminding authors to add information that often strengthens their conclusions but has been omitted *(12)*. A checklist can "ensure clear and transparent reporting of studies, which will, as always, depend for their importance on the creativity and insights and effort of their authors" *(12)*. As authors, reviewers, and readers, we can improve the quality of reporting, and hence the quality of the evidence base, for diagnostic tests. In the end, better reporting of the results of diagnostic accuracy studies has the promise to improve the efficiency and quality of healthcare.

References

1. 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.
2. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. JAMA 1995;274:645–51.
3. 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.
4. International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals: writing and editing for biomedical publication [updated November 2003]. http://www.icmje.org.
5. Haynes RB. More informative abstracts: current status and evaluation. J Clin Epidemiol 1993;46:595–7.
6. Moher D, Schulz KF, Altman D. The CONSORT statement: revised recommendations for improving the quality of reports of parallel group randomized trials. JAMA 2001;285:1987–91.
7. Bruns DE, Huth EJ, Magid E, Young DS. Toward a checklist for reporting of studies of diagnostic accuracy of medical tests. Clin Chem 2000;46:893–5.
8. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Clin Chem 2003;49:1–6.
9. 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.
10. Moher D, Jones A, Lepage L, for the CONSORT group. Use of the CONSORT statement and quality of reports of randomized trials. A comparative before-and-after evaluation. JAMA 2001;285:1992–5.
11. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Clin Chem Lab Med 2003;41:68–73.
12. Bruns DE. The STARD initiative and the reporting of studies of diagnostic accuracy. Clin Chem 2003;49:19–20.

Patrick M.M. Bossuyt

Academic Medical Center
University of Amsterdam
Amsterdam, NA 1105 AZ The Netherlands
Fax 31-31-20-6912683
E-mail p.m.bossuyt@amc.uva.nl

DOI: 10.1373/clinchem.2003.029736

10.7.- Anexo VII: 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.

Clinical Chemistry 50:3
530–536 (2004)

Evidence-based
Laboratory Medicine
and Test Utilization

Methodology in Diagnostic Laboratory Test Research in *Clinical Chemistry* and *Clinical Chemistry and Laboratory Medicine*

Blanca Lumbreras-Lacarra,[1] José Manuel Ramos-Rincón,[2] and Ildefonso Hernández-Aguado[3*]

Background: The application of epidemiologic principles to clinical diagnosis has been less developed than in other clinical areas. Knowledge of the main flaws affecting diagnostic laboratory test research is the first step for improving its quality. We assessed the methodologic aspects of articles on laboratory tests.

Methods: We included articles that estimated indexes of diagnostic accuracy (sensitivity and specificity) and were published in ***Clinical Chemistry*** or ***Clinical Chemistry and Laboratory Medicine*** in 1996, 2001, and 2002. ***Clinical Chemistry*** has paid special attention to this field of research since 1996 by publishing recommendations, checklists, and reviews. Articles were identified through electronic searches in Medline. The strategy combined the Mesh term "sensitivity and specificity" (exploded) with the text words "specificity", "false negative", and "accuracy". We examined adherence to seven methodologic criteria used in the study by Reid et al. (***JAMA*** 1995;274:645–51) of papers published in general medical journals. Three observers evaluated each article independently.

Results: Seventy-nine articles fulfilled the inclusion criteria. The percentage of studies that satisfied each criterion improved from 1996 to 2002. Substantial improvement was observed in reporting of the statistical uncertainty of indices of diagnostic accuracy, in criteria based on clinical information from the study population (spectrum composition), and in avoidance of workup bias. Analytical reproducibility was reported frequently (68%), whereas information about indeterminate results was rarely provided. The mean number of methodologic criteria satisfied showed a statistically significant increase over the 3 years in ***Clinical Chemistry*** but not in ***Clinical Chemistry and Laboratory Medicine.***

Conclusions: The methodologic quality of the articles on diagnostic test research published in ***Clinical Chemistry*** and ***Clinical Chemistry and Laboratory Medicine*** is comparable to the quality observed in the best general medical journals. The methodologic aspects that most need improvement are those linked to the clinical information of the populations studied. Editorial actions aimed to increase the quality of reporting of diagnostic studies could have a relevant positive effect, as shown by the improvement observed in ***Clinical Chemistry.***

The recently published STARD initiative concerning the publication of studies on diagnostic accuracy is an important step forward in improving the quality of diagnostic investigation. Its relevance is even greater if we consider the many challenges of research on diagnostic tests and that application of epidemiologic principles to the field of clinical diagnosis has been less developed than other clinical areas, or perhaps less effective *(1, 2)*.

In 1996, Reid et al. *(3)* pointed out serious methodologic limitations of research on diagnostic tests published in the most prestigious international scientific clinical journals. More recently, serious methodologic limitations have been brought to light in reports on genetic testing *(4, 5)*.

Clinical Chemistry has been paying special attention to this field of research. From 1996, *Clinical Chemistry* has included in the instructions for authors the seven criteria used by Reid et al. *(3)* in their study of the methodologic

[1] Department of Clinic Analysis, General University Hospital of Alicante, Alicante, Spain.
[2] Department of Internal Medicine, General University Hospital of Elche, Elche, Spain.
[3] Department of Public Health, University of Miguel Hernández, San Juan de Alicante, Spain.
*Address correspondence to this author at: Department of Public Health, Facultad de Medicina, University of Miguel Hernández, Carretera de Valencia Km. 8.7, 03550 San Juan de Alicante, Spain. Fax 34-96-5919551; e-mail ihernandez@umh.es.
Received April 4, 2003; accepted December 12, 2003.
Previously published online at DOI: 10.1373/clinchem.2003.019786

quality of studies of diagnostic accuracy studies published in general medical journals (*Annals of Internal Medicine*, *JAMA*, *Lancet*, and *New England Journal of Medicine*). In 1997 and 2000, *Clinical Chemistry* published drafts of a checklist for reporting of studies of diagnostic accuracy *(6, 7)*, and in 2001, a 28-item checklist became part of the Information for Authors. Furthermore, *Clinical Chemistry* has published methodologic reviews on diagnostic research and evidence-based laboratory medicine *(8, 9)*. In 2003, *Clinical Chemistry* published the STARD Initiative for reporting studies of diagnostic accuracy *(1, 2, 10)*, a guideline that was also published or promoted in several other journals, including *Academic Radiology*, *American Journal of Clinical Pathology*, *Annals of Internal Medicine*, *British Medical Journal*, *Clinical Biochemistry*, *Clinical Chemistry and Laboratory Medicine*, *JAMA*, *Journal of Clinical Microbiology*, *Lancet*, and *Radiology*.

Despite outstanding efforts to improve the quality of methodologic aspects of diagnostic investigations, there is a lack of specific information concerning the quality of the investigation of tests carried out in the laboratory. Knowledge of particular weaknesses in this research field could guide progress toward evidence-based laboratory diagnosis.

The objective of our study was to analyze the reporting and methodologic quality of published studies of diagnostic accuracy of laboratory tests. We considered diagnostic studies published in two journals dealing with clinical chemistry, included in the *Journal Citation Reports* (*Clinical Chemistry* and *Clinical Chemistry and Laboratory Medicine*), and applied the criteria proposed by Reid et al. *(3)*.

Materials and Methods

Clinical Chemistry and *Clinical Chemistry and Laboratory Medicine* (previously called *European Journal of Clinical Chemistry and Clinical Biochemistry*) are journals specializing in clinical laboratory studies that, in addition to having a major impact, frequently publish articles on the evaluation of diagnostic tests in the clinical chemistry laboratory. We reviewed the articles on diagnostic tests published in the two journals in 1996, 2001, and 2002. To select the articles we carried out a search in PubMed, using the most accurate strategy described by Devillé et al. *(11)*. The strategy combined the Mesh term "sensitivity and specificity" (exploded) with the text words "specificity", "false negative", and "accuracy". To improve sensitivity we expanded the search including the Mesh term "area under the curve" and the text words "diagnostic odds ratio" and "likelihood ratios".

Articles were accepted for further review if they fulfilled the inclusion criteria described by Reid et al. *(3)*: humans were tested, the test was intended for clinical use, and indexes of accuracy were provided with both sensitivity and specificity or with the counterpart likelihood ratio. Additionally, we reviewed papers that provided ROC curves. Only original articles showing an abstract in Medline were finally reviewed. Two authors checked the abstracts for eligibility criteria, and in case of doubt the full article was reviewed.

APPLICATION OF METHODOLOGIC CHECKLIST

We applied the seven methodologic criteria recommended by Reid et al. *(3)*, which we reproduce below literally:

(1) Spectrum composition: This standard was met if at least three of the following four descriptors were provided: sex distribution, age distribution, summary of presenting clinical symptoms and/or disease stage, and eligibility criteria for study subjects.
(2) Analysis of pertinent subgroups: This standard was fulfilled when results for indexes of accuracy were cited for any pertinent demographic or clinical subgroup of the investigated population.
(3) Avoidance of workup bias: For cohort studies, this standard was met if all subjects were assigned to receive both diagnostic testing and gold standard verification, either by direct procedure or by suitable clinical follow-up. In case-control studies, credit depends on whether the diagnostic test preceded or followed the gold standard procedure. If the diagnostic test preceded, credit was given if disease verification was obtained for a consecutive series of study subjects regardless of their diagnostic test result. If the diagnostic test followed, credit was given if the results were stratified according to the clinical factor that evoked the gold standard procedure.
(4) Avoidance of review bias: For prospective cohort studies in which patients always receive the diagnostic test first, credit was given if the gold standard procedures were evaluated independently. A statement about independence in interpreting both the test and the gold standard procedure was required for prospective studies in which the gold standard procedure was sometimes done before the diagnostic test and for case-control studies in which the test preceded the gold standard procedure. In case-control studies in which the diagnostic test followed disease verification, a statement was required to indicate an independent evaluation of the diagnostic test.
(5) Precision of results for test accuracy: This standard was met if standard error or confidence intervals, regardless of magnitude, were reported for test sensitivity and specificity or likelihood ratios.
(6) Presentation of indeterminate test results: To meet this standard, a study had to report all of the appropriate positive, negative, and indeterminate results generated during evaluation of the diagnostic test and whether indeterminate results had been included or excluded when indexes of accuracy were calculated.
(7) Test reproducibility: For test requiring observer interpretation, at least some of the tests subjects should have been evaluated for a summary measure of observer variability. For tests performed without ob-

server interpretation, credit was given for a summary measure of instrument variability.

Each article was evaluated independently by at least two observers. The observer agreement in this phase was 87%. In a second step, all disagreements were evaluated by the third author and solved by consensus. Data analyses were carried out with Epiinfo 2000.

Results

Clinical Chemistry published 440 articles in 1996, 436 in 2001, and 417 in 2002. *Clinical Chemistry and Laboratory Medicine* published 162 articles in 1996, 210 in 2001, and 232 in 2002. Of the 1897 articles published in both journals, PubMed searches identified 460 on test evaluation. Of those 460 articles, only 358 were original reports. Finally, 79 papers fulfilled the eligibility criteria: 52 were from the journal *Clinical Chemistry* (11 were published in 1996, 17 in 2001, and 24 in 2002) *(12–63)* and 27 from *Clinical Chemistry and Laboratory Medicine* (7 published in 1996, 10 in 2001, and 10 in 2002) *(64–90)*.

Regarding the diagnostic procedures evaluated in these studies, biochemistry was the field most frequently referred to (55%), followed by immunology (22%), microbiology (8%), hormones (9%), genetics (4%), and hematology (2%). The mean number of patients/samples included in the studies was 331 (range, 10–2971 patients/samples).

The mean number of methodologic criteria satisfied was 2.3 in 1996, 2.7 in 2001, and 3.4 in 2002 ($P = 0.047$). For *Clinical Chemistry*, the means were 2.0 in 1996, 2.9 in 2001, and 4.0 in 2002 ($P = 0.001$). For *Clinical Chemistry and Laboratory Medicine*, the means were 2.7 in 1996, 2.4 in 2001, and 1.9 in 2002 ($P = 0.66$).

Fulfillment of individual methodologic criteria ranged from 0% to 83% for studies published in 1996, from 15% to 81% for studies published in 2001, and from 6% to 82% for studies published in 2002 (Table 1). Articles published in 2002 showed better fulfillment than those published in 1996 or 2001 in all but two criteria. Presentation of data for test reproducibility was met by most studies in 1996 (83%); this value did not change in 2001 (81%), but in 2002 changed to 68%. Forty-four percent of the studies in 1996, 26% in 2001, and 38% in 2002 satisfied the second standard, estimation of accuracy in pertinent subgroups. Among the criteria that appeared to have improved in 2002, the statistical uncertainty of test indexes changed from 22% in articles published in 1996 to 44% in 2001 and to 65% in 2002. The spectrum composition, from 22% in articles published in 1996 to 37% in 2001 and 71% in 2002. Unexpectedly, the incidence and handling of indeterminate test results were hardly discussed.

We compared our results in 1996 with those observed by Reid et al. *(3)* in the subsample of articles published between 1990 and 1993 from a group of relevant clinical journals (Table 1). In spite of the reduced sample sizes, some statistically significant differences were observed. The articles published in 1996 in *Clinical Chemistry* more frequently met the second (accuracy in subgroups) and seventh standards (reproducibility). The articles studied by Reid et al. *(3)* did better in the presentation of indeterminate results.

Discussion

Judging by the articles published in *Clinical Chemistry* and *Clinical Chemistry and Laboratory Medicine*, both with a high impact factor, the quality of the methodology used in research on diagnostic laboratory tests is comparable to the most important international clinical journals. In certain aspects of the methodology, specifically the reproducibility of test results, the articles published in labora-

Table 1. Fulfillment of the methodologic criteria used by Reid et al. *(3)* in the articles published in 1996, 2001, and 2002 in *Clinical Chemistry* and *Clinical Chemistry and Laboratory Medicine* and in articles published in general medical journals and reviewed by Reid et al.

	Number (%) of articles			
	Year 1996[a] (n = 18)	Year 2001 (n = 27)	Year 2002 (n = 34)	Reid et al. (1990–1993)[b] (n = 34)
Spectrum composition	4 (22)	10 (37)	24 (71)	11 (32)
Age distribution	9 (50)	18 (67)	28 (82)	
Sex distribution	9 (50)	19 (70)	26 (76)	
Clinical symptoms and/or disease stage	6 (33)	11 (41)	12 (35)	
Study eligibility criteria	4 (22)	10 (37)	18 (53)	
Accuracy in subgroups	8 (44)	7 (26)	13 (38)	4 (12)[c]
Avoidance of workup bias	6 (33)	13 (48)	24 (71)	21 (62)
Avoidance of review bias	4 (22)	10 (37)	15 (44)	16 (47)
Test accuracy precision	4 (22)	12 (44)	22 (65)	8 (24)
Indeterminate tests results	0	4 (15)	2 (6)	13 (38)[c]
Test reproducibility	15 (83)	22 (81)	21 (68)	11 (32)[c]

[a] Differences between 1996 and 2001 were not statistically significant.
[b] Articles from Reid et al. *(3)* were restricted to those published in 1990–1993, the closest dates to 1996.
[c] $P < 0.05$ in Epiinfo Fisher exact test comparing proportions between 1996 and articles reviewed by Reid et al. *(3)*.

tory journals appear to be of higher quality. On the other hand, there are some methodologic flaws that would be easy to correct and would substantially improve the clinical applicability of the results of the investigations. On average the quality of the articles published in laboratory journals is high, but very few articles comply with all or almost all of the methodologic standards.

It is difficult to evaluate whether our analysis was more or less strict than that performed by Reid et al. *(3)*. Although the criteria for application of some of the methodologic standards were not clearly specified by Reid et al., we did not observe any great interobserver variation in the results of the evaluation, and this gives us a certain degree of confidence in our findings.

One of the main discrepancies with the results obtained by Reid et al. *(3)* is the frequency with which the authors described the reproducibility of the results of the diagnostic test they were evaluating. Our results show that the laboratory studies report this characteristic much more frequently [83% in 1996, 81% in 2001, and 68% in 2002 vs 32% in Reid et al. *(3)*]. This result may have been expected because the experts in diagnostic laboratory tests usually pay more attention to analytical imprecision than do others.

A key shortcoming of the diagnostic studies done until 2001 in the laboratory is the lack of a full description of the spectrum of patients or samples studied (22% in 1996, 37% in 2001, and 71% in 2002). The improvement observed in 2002, particularly in *Clinical Chemistry*, has two important positive consequences. On the one hand, readers can better judge whether the population studied is comparable to other populations, which favors the applicability of results. On the other hand, the description of the clinical spectrum allows the presentation of findings by strata so that the reader can judge whether the accuracy of the test evaluated may change depending on the clinical or sociodemographic characteristics of the patients studied. Not only do we believe that authors should describe the spectrum of patients or samples studied and present the analysis by strata, but also that the STARD criteria *(1, 2)* should be applied, and it should be necessary to provide a more detailed explanation of the scope of the study, the way of presenting the patients, and the type of sampling. This would be a decisive step forward in improving the applicability of the results.

Regarding the presence of workup and review biases, it seems that in our series they have been prevented to an extent similar to that in the series studied by Reid et al. *(3)*. However, it has to be pointed out that many of the publications evaluated appeared to be free of those biases such that potential for them might be said not to exist. The lack of explicit information, however, makes it impossible to guarantee that the study was indeed free of such biases.

Few articles took indeterminate results into account when evaluating the diagnostic test (none in 1996, 15% in 2001, and 6% in 2002) compared with 38% in the study by Reid et al. *(3)*. It is possible that many of the articles analyzed in our study did not have any indeterminate results, but even so, they do not comply with the criterion because they do not say so explicitly, as required by the criteria of Reid et al.

Undoubtedly, the recommendations of the STARD initiative published in *Clinical Chemistry (1, 2)* include more exhaustive requirements than those proposed by Reid et al. *(3)*. Use of the latter enabled us to compare the papers from clinical laboratory journals with those described in the most important international clinical journals. In addition, it allowed us to analyze the trend. For example, we studied the articles published a year before the appearance in *Clinical Chemistry* of the first article dealing with the importance of methodology in evaluation of diagnostic tests (1996) *(5)* and 5–6 years later (2001–2002). This enabled us to study the change produced in the literature and, therefore, the effect of publication in the journal *Clinical Chemistry* of the recommendations to be followed in the study of laboratory tests.

In 1996, the journal *Clinical Chemistry* included criteria similar to those of Reid et al. *(3)* in its instructions to authors, and in 1997 and in 2000 it published preliminary versions of a 28-item checklist, whereas *Clinical Chemistry and Laboratory Medicine* began to advocate the use of the STARD criteria in 2003 only. It therefore seems logical that *Clinical Chemistry* improved between 1996, 2001, and 2002, whereas *Clinical Chemistry and Laboratory Medicine* did not. Similar results were observed by Moher et al. *(91)* and Altman et al. *(92)*, who demonstrated that the quality of reports in journals that promoted the Consolidated Standards of Reporting Trials (CONSORT) (*British Medical Journal*, *JAMA*, and *Lancet*) showed greater improvement than in a journal that did not advocate its use (*New England Journal of Medicine*). The results presented in these two reports, as well as ours, support the conclusion that editors and reviewers, in adopting criteria such as STARD or CONSORT, can play a key role in improving the quality of published reports of studies of this sort.

Both the STARD initiative and the appearance of the first-ever publication dedicated specifically to diagnostic investigation *(93)* may lead to future improvement of the quality of this type of investigation. Indeed, in the case of *Clinical Chemistry*, this is already evident. In the future, improvement in the quality of the methodology should be monitored to confirm the possible beneficial effects. Furthermore, there are other aspects of diagnostic investigation, with objectives other than accuracy, that will require development and new standards *(94, 95)*. These objectives include both the use of clinical trials of diagnostic tests and investigation of the undesirable effects on health of unwanted information given by these tests.

We thank Judith Williams help in preparing the manuscript; we also thank the two anonymous reviewers and Dr. Joseph Watine for their useful comments during the peer reviewing of the manuscript.

References

1. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Standards for Reporting of Diagnostic Accuracy. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD statement. Clin Chem 2003;49:1–6.
2. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, 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.
3. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. JAMA 1995;274:645–51.
4. Bogardus ST Jr, Concato J, Feinstein AR. Clinical epidemiological quality in molecular genetic research. The need for methodological standards. JAMA 1999;281:1919–26.
5. 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.
6. Bruns DE. The clinical chemist. Clin Chem 1997;43:2211–2.
7. Bruns DE, Huth EJ, Magid E, Young DS. Towards a checklist for reporting of studies of diagnostic accuracy of medical tests. Clin Chem 2000;46:893–5.
8. Price CP. Evidence-based laboratory medicine. Supporting decision making. Clin Chem 2000;46:1041–50.
9. McQueen MJ. Overview of evidence based medicine challenges for evidence based laboratory medicine. Clin Chem 2001;47:1536–46.
10. Bruns DE. The STARD initiative and the reporting of studies of diagnostic accuracy. Clin Chem 2003;49:19–20.
11. Deville WL, Bezemer PD, Bouter LM. Publications on diagnostic test evaluations in family medicine journals: an optimal search strategy. J Clin Epidemiol 2000;53:65–9.
12. Schellenberg F, Martin M, Caces E, Benard JY, Weill J. Nephelometric determination of carbohydrate deficient transferrin. Clin Chem 1996;42:551–7.
13. Woltge HW, Seibel MJ, Ziegler R. Comparison of total and bone-specific alkaline phosphatase inpatients with non skeletal disorders or metabolic bone diseases. Clin Chem 1996;42:1796–804.
14. Gerard B, Peponnet C, Brunie G, Cave H, Denamur E, d'Auriol L, et al. Fluorometric detection of HIV-I genome through use of an internal control, inosine-substituted primers, and microtiter plate format. Clin Chem 1996;42:696–703.
15. Laurino JP, Bender EW, Kessimian N, Chang J, Pelletier T, Usategui M. Comparative sensitivities and specificities of the mass measurements of CK-MB2, CK-MB, and myoglobin for diagnosing acute myocardial infarction. Clin Chem 1996;42:1454–9.
16. Villena V, Navarro-Gonzalvez JA, Garcia-Benayas C, Manzanos JA, Echave J, Lopez-Encuentra A, et al. Rapid automated determination of adenosine deaminase and lysozyme for differentiating tuberculous and non tuberculous pleural effusion. Clin Chem 1996;42:218–21.
17. Sacchetti L, Ferrajolo A, Salerno G, Esposito P, Lofrano MM, Orlani G, et al. Diagnostic value of various serum antibodies detected by diverse methods in childhood celiac disease. Clin Chem 1996;42:1838–42.
18. Rohlfs EM, Chaing SH, Chapman JF. Analytical and clinical evaluation of refractive index-matched anomalous diffraction (RIMAD) for assessment of fetal lung maturation. Clin Chem 1996;42:1861–8.
19. Castaldo G, Intrieri M, Calcagno G, Cimino L, Budillon G, Sacchetti L, et al. Ascitic pseudiuridine discriminates between hepatocarcinoma-derived ascites and cirrhotic ascites. Clin Chem 1996;42:1843–6.
20. Marquet PY, Daver A, Sapin R, Bridgi B, Muratet JP, Hartmann DJ, et al. Highly sensitive immunoradiometric assay for serum thyroglobulin with minimal interference from autoantibodies. Clin Chem 1996;42:258–62.
21. Helander A, Beck O, Jones AW. Laboratory testing for recent alcohol consumption: comparison of ethanol, methanol, and 5-hydroxytryptophol. Clin Chem 1996;42:618–24.
22. Guéchot J, Laudat A, Loria A, Serfaty L, Poupon R, Giboudeau J. Diagnostic accuracy of hyaluronan and type III procollagen amino-terminal peptide serum assays as markers of liver fibrosis in chronic viral hepatitis C evaluated by ROC curve analysis. Clin Chem 1996;42:558–63.
23. Bizzaro N, Mazzanti G, Tonutti E, Villalta D, Tozzoli R. Diagnostic accuracy of the anti-citrulline antibody assay for rheumatoid arthritis. Clin Chem 2001;47:1089–93.
24. Hayashi N, Kawamoto T, Mukai M, Morinobu A, Koshiba M, Kondo S, et al. Detection of antinuclear antibodies by use of an enzyme immunoassay with nuclear Hep-2 cell extract and recombinant antigens: comparison with immunofluorescence assay in 307 patients. Clin Chem 2001;47:1649–59.
25. Norberg T, Klaar S, Lindqvist L, Lindahl T, Ahlgren J, Bergh J. Enzymatic mutation detection method evaluated for detection of p53 mutations in cDNA from breast cancers. Clin Chem 2001;47:821–8.
26. Christenson RH, Duh SH, Sanhai WR, Wu AH, Holtman V, Painter P, et al. Characteristics of an albumin cobalt binding test for assessment of acute coronary syndrome patients: a multicenter study. Clin Chem 2001;47:464–70.
27. Turpeinen U, Methuen T, Alfthan H, Laitinen K, Salaspuro M, Stenman UH. Comparison of HPLC and small column (CDTect) methods for disialotransferrin. Clin Chem 2001;47:1782–7.
28. Umapathysivam K, Hopwood JJ, Meikle PJ. Determination of acid α-glucosidase activity in bloodspots as a diagnostic test for Pompe disease. Clin Chem 2001;47:1378–83.
29. Cassinat B, Darsin D, Guardiola P, Toubert ME, Rain JD, Gluckman E, et al. Intermethod discordance for α-fetoprotein measurements in Fanconi anemia. Clin Chem 2001;47:1405–9.
30. Bijwaard KE, Aguilera NS, Monczak Y, Trudel M, Taubenberger JK, Lichy JH. Quantitative real-time reverse transcription PCR assay for cyclin D1 expression: utility in the diagnosis of mantle cell lymphoma. Clin Chem 2001;47:195–201.
31. Rickert M, Seissler J, Dangel W, Lorenz H, Richter W. Fusion protein for combined analysis of autoantibodies to the 65-kDa isoform of glutamic acid decarboxylase and islet antigen-2 in insulin-dependent diabetes mellitus. Clin Chem 2001;47:926–34.
32. Jensen UG, Brandt NJ, Christensen E, Skovby F, Norgaard-Pedersen B, Simonsen H. Neonatal screening for galactosemia by quantitative analysis of hexose monophosphates using tandem mass spectrometry: a retrospective study. Clin Chem 2001;47:1364–72.
33. Lempinen M, Kylänpää-Bäck ML, Stenman UH, Puolakkainen P, Haaplainen R, Finne P, et al. Predicting the severity of acute pancreatitis by rapid measurement of trypsinogen-2 in urine. Clin Chem 2001;47:2103–7.
34. Nurmikko P, Pettersson K, Piironen T, Hugosson J, Lilja H. Discrimination of prostate cancer from benign disease by plasma measurement of intact, free prostate-specific antigen lacking an internal cleavage site at Lys^{145}–Lys^{146}. Clin Chem 2001;47:1415–23.
35. Sillanaukee P, Olsson U. Improved diagnostic classification of alcohol abusers by combining carbohydrate-deficient transferrin and γ-glutamyltransferase. Clin Chem 2001;47:681–5.
36. Mourad M, Malaise J, Chalb Eddour D, De Meyer M, König J, Schepers R, et al. Pharmacokinetic basis for the efficient and safe

use of low-dose mycophenolate mofetil in combination with tacrolimus in kidney transplantation. Clin Chem 2001;47:1241–8.
37. Anton RF, Dominic KC, Bigelow M, Westby C, CDTect Research Group. Comparison of Bio-Rad %CDT TIA and CDTect as laboratory markers of heavy alcohol use and their relationships with γ-glutamyltransferase. Clin Chem 2001;47:1769–75.
38. Jonsson M, Carlson J, Jeppsson JO, Simonsson P. Computer-supported detection of M-components and evaluation of immunoglobulins after capillary electrophoresis. Clin Chem 2001;47:110–7.
39. Andersen JM, Hedström J, Kemppainen E, Finne P, Poulakkainen P, Stenman UH. The ratio of trypsin-2-α1-antitrypsin to trypsinogen-1 discriminates biliary and alcohol-induced acute pancreatitis. Clin Chem 2001;47:231–6.
40. Richard S, Mlossec V, Moreau JF, Taupin JL. Detection of oligoclonal immunoglobulins in cerebrospinal fluid by an immunofixation-peroxidase method. Clin Chem 2002;48:167–73.
41. Ferre N, Camps J, Prats E, Vilella E, Paul A, Figuera L, et al. Serum paraoxonase activity: a new additional test for the improved evaluation of chronic liver damage. Clin Chem 2002;48:261–8.
42. Weber LT, Shipkova M, Armstrong VW, Wagner N, Schutz E, Mehls O, et al. Comparison of the Emit immunoassay with HPLC for therapeutic drug monitoring of mycophenolic acid in pediatric renal-transplant recipients on mycophenolate mofetil therapy. Clin Chem 2002;48:517–25.
43. Filler G, Priem F, Lepage N, Sinha P, Vollmer I, Clark H, et al. β-Trace protein, cystatin C, β_2-microglobulin, and creatinine compared for detecting impaired glomerular filtration rates in children. Clin Chem 2002;48:729–36.
44. Fantz CR, Powell C, Karon B, Parvin CA, Hankins K, Dayal M, et al. Assessment of the diagnostic accuracy of the TDx-FLM II to predict fetal lung maturity. Clin Chem 2002;48:761–5.
45. Orlando R, Mussap M, Plebani M, Piccoli P, De Martin S, Floreani M, et al. Diagnostic value of plasma cystatin C as a glomerular filtration marker in decompensated liver cirrhosis. Clin Chem 2002;48:850–8.
46. Wasmuth JC, Oliver y Minarro D, Homrighausen A, Leifeld L, Rockstroh JK, Sauerbruch T, et al. Phospholipid autoantibodies and the antiphospholipid antibody syndrome: diagnostic accuracy of 23 methods studied by variation in ROC curves with number of clinical manifestations. Clin Chem 2002;48:1004–10.
47. Xu SQ, He M, Yu HP, Wang XY, Tan XL, Lu B, et al. Bioluminescent method for detecting telomerase activity. Clin Chem 2002;48:1016–20.
48. Poon TC, Mok TS, Chan AT, Chan CM, Leong V, Tsui SH, et al. Quantification and utility of monosialylated α-fetoprotein in the diagnosis of hepatocellular carcinoma with nondiagnostic serum total α-fetoprotein. Clin Chem 2002;48:1021–7.
49. Martinez M, Espana F, Royo M, Alapont JM, Navarro S, Estelles A, et al. The proportion of prostate-specific antigen (PSA) complexed to α_1-antichymotrypsin improves the discrimination between prostate cancer and benign prostatic hyperplasia in men with a total PSA of 10 to 30 μg/L. Clin Chem 2002;48:1251–6.
50. Stephan C, Cammann H, Semjonow A, Diamandis EP, Wymenga LF, Lein M, et al. Multicenter evaluation of an artificial neural network to increase the prostate cancer detection rate and reduce unnecessary biopsies. Clin Chem 2002;48:1279–87.
51. Li J, Zhang Z, Rosenzweig J, Wang YY, Chan DW. Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. Clin Chem 2002;48:1296–304.
52. Krauss T, Emons G, Kuhn W, Augustin HG. Predictive value of routine circulating soluble endothelial cell adhesion molecule measurements during pregnancy. Clin Chem 2002;48:1418–25.
53. Panteghini M, Cuccia C, Bonetti G, Giubbini R, Pagani F, Bonini E. Single-point cardiac troponin T at coronary care unit discharge after myocardial infarction correlates with infarct size and ejection fraction. Clin Chem 2002;48:1432–6.
54. Oda H, Shiina Y, Seiki K, Sato N, Eguchi N, Urade Y. Development and evaluation of a practical ELISA for human urinary lipocalin-type prostaglandin D synthase. Clin Chem 2002;48:1445–53.
55. Carroccio A, Vitale G, Di Prima L, Chifari N, Napoli S, La Russa C, et al. Comparison of anti-transglutaminase ELISAs and an anti-endomysial antibody assay in the diagnosis of celiac disease: a prospective study. Clin Chem 2002;48:1546–50.
56. Qu Y, Adam BL, Yasui Y, Ward MD, Cazares LH, Schellhammer PF, et al. Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. Clin Chem 2002;48:1835–43.
57. Kauppinen R, von und zu Fraunberg M. Molecular and biochemical studies of acute intermittent porphyria in 196 patients and their families. Clin Chem 2002;48:1891–900.
58. Schwartz GL, Chapman AB, Boerwinkle E, Kisabeth RM, Turner ST. Screening for primary aldosteronism: implications of an increased plasma aldosterone/renin ratio. Clin Chem 2002;48:1919–23.
59. Derhaschnig U, Laggner AN, Roggla M, Hirschl MM, Kaplotis S, Marsik C, et al. Evaluation of coagulation markers for early diagnosis of acute coronary syndromes in the emergency room. Clin Chem 2002;48:1924–30.
60. Kulpa J, Wojcik E, Reinfuss M, Kolodziejski L. Carcinoembryonic antigen, squamous cell carcinoma antigen, CYFRA 21-1, and neuron-specific enolase in squamous cell lung cancer patients. Clin Chem 2002;48:1931–7.
61. Legros FJ, Nuyens V, Minet E, Emonts P, Boudjeltia KZ, Courbe A, et al. Carbohydrate-deficient transferrin isoforms measured by capillary zone electrophoresis for detection of alcohol abuse. Clin Chem 2002;48:2177–86.
62. Ryden I, Pahlsson P, Lindgren S. Diagnostic accuracy of α_1-acid glycoprotein fucosylation for liver cirrhosis in patients undergoing hepatic biopsy. Clin Chem 2002;48:2195–201.
63. Penders J, Fiers T, Delanghe JR. Quantitative evaluation of urinalysis test strips. Clin Chem 2002;48:2236–41.
64. van Dalen A, Kessler AC. A multicentre evaluation of tumour marker determinations using the automatic Enzymum-test systems ES 300 and ES 600/700. Eur J Clin Chem Clin Biochem 1996;34:377–84.
65. Bacigalupo MA, Bazzini P, Farina L, Ius A. Evaluation of three immunoassays for detection of toxoplasma-specific immunoglobulin G and M. Eur J Clin Chem Clin Biochem 1996;34:503–5.
66. Matsushita H, Xu J, Kuroki M, Kondo A, Inoue E, Teramura Y, et al. Establishment and evaluation of a new chemiluminescent enzyme immunoassay for carcinoembryonic antigen adapted to the fully automated ACCES System. Eur J Clin Chem Clin Biochem 1996;34:829–35.
67. Schmitt UM, Stieber P, Hasholzner U, Pahl H, Hofmann K, Fateh-Moghadam A. Methodological and clinical evaluation of two automated enzymatic immunoassays as compared with a radioimmunoassays for neuron-specific enolase. Eur J Clin Chem Clin Biochem 1996;34:679–82.
68. Conejo JR, Benedito JE, Jiménez A, Menchén M, Cano J, Granizo V, et al. Diagnostic value of three tumour markers determined in pleural effusions. Eur J Clin Chem Clin Biochem 1996;34:139–42.
69. Plebani M, Borghesan F, Bernardi D, Faggian D. Clinical evaluation of a new quantitative method for specific IgE antibodies. Eur J Clin Chem Clin Biochem 1996;34:579–84.
70. Collinson P, Gerhardt W, Katus HA, Müller-Bardorff M, Braun S, Schricke U, et al. Multicentre evaluation of an immunological rapid test for the detection of troponin T in whole blood samples. Eur J Clin Chem Clin Biochem 1996;34:591–8.

71. Mroczko B, Szmitkowski M, Nikliński J. Granulocyte-colony stimulating factor and macrophage-colony stimulating factor in patients with non-small cell lung cancer. Clin Chem Lab Med 2001;39:374–9.
72. von Eyben FE, Blaabjerg O, Hyltoft-Petersen P, Madsen EL, Amato R, Liu F, et al. Serum lactate dehydrogenase isoenzyme 1 and prediction of death in patients with metastatic testicular germ cell tumors. Clin Chem Lab Med 2001;39:38–44.
73. Shaarawy M, Salem ME. Clinical value of microtransferrinuria and microalbuminuria in the prediction of preeclampsia. Clin Chem Lab Med 2001;39:29–34.
74. Genser B, Truschnig-Wilders M, Stunzner D, Landini MP, Halwachs-Baumann G. Evaluation of five commercial enzyme immunoassays for the detection of human cytomegalovirus-specific IgM antibodies in the absence of a commercially available gold standard. Clin Chem Lab Med 2001;39:62–70.
75. Tello FL, Prats CH, González MD. Free and complexed prostate-specific antigen (PSA) in the early detection of prostate cancer. Clin Chem Lab Med 2001;39:116–20.
76. Gignaric I, Casl MT, Simic D, Lukac J. Serum amyloid A protein (SAA) in colorectal carcinoma. Clin Chem Lab Med 2001;39:129–33.
77. Mello G, Parretti E, Ognibene A, Mecacci F, Cioni R, Scarselli G, et al. Prediction of the development of pregnancy induced hypertensive disorders in high risk pregnant women by artificial neural networks. Clin Chem Lab Med 2001;39:801–5.
78. Schellenberg F, Mennetrey L, Bacq Y, Pages JC. Carbohydrate-deficient transferrin (CDT)determination by nephelometry using a commercial kit. Analytical and diagnostic aspects. Clin Chem Lab Med 2001;39:866–71.
79. Sarno M, Sarno L, Baylink D, Drinkwater B, Farley S, Kleerekoper M, et al. Excretion of sweat and urine pyridinoline crosslinks in healthy controls and subjects with established metabolic bone disease. Clin Chem Lab Med 2001;39:223–8.
80. Siekmeier R, Bierlich A, Jaross W. The white blood cell differential: three methods compared. Clin Chem Lab Med 2001;39:432–45.
81. Giovanella L, Ceriani L, Giardina G, Bardelli D, Tanzi F, Garancini S. Serum cytokeratin fragment 21.1 (CYFRA 21.1) as tumour marker for breast cancer: comparison with carbohydrate antigen 15.3 (CA 15.3) and carcinoembryonic antigen (CEA). Clin Chem Lab Med 2002;40:298–303.
82. Mroczko B, Szmitkowski M, Okulczyk B. Granulocyte-colony stimulating factor (G-CSF) and macrophage-colony stimulating factor (M-CSF) in colorectal cancer patients. Clin Chem Lab Med 2002;40:351–5.
83. Giovanella L, Ceriani L. High-sensitivity human thyroglobulin (hTG) immunoradiometric assay in the follow-up of patients with differentiated thyroid cancer. Clin Chem Lab Med 2002;40:480–4.
84. Kocna P, Vanícková Z, Perusicová J, Dvorák M. Tissue transglutaminase—serology markers for coeliac disease. Clin Chem Lab Med 2002;40:485–92.
85. Barlandas-Rendon E, Muller MM, Garcia-Latorre E, Heinschink A. Comparison of urine cell characteristics by flow cytometry and cytology in patients suspected of having bladder cancer. Clin Chem Lab Med 2002;40:817–23.
86. Iinuma Y, Senda K, Takakura S, Ichiyama S, Tano M, Abe T, et al. Evaluation of a commercially available serologic assay for antibodies against tuberculosis-associated glycolipid antigen. Clin Chem Lab Med 2002;40:832–6.
87. Al-Daghistani HI, Abdel-Dayem M. Diagnostic value of various urine tests in the Jordanian population with urinary tract infection. Clin Chem Lab Med 2002;40:1048–51.
88. Hernando M, Gonzalez C, Sanchez A, Guevara P, Navajo JA, Papisch W, et al. Clinical evaluation of a new automated anti-dsDNA fluorescent immunoassay. Clin Chem Lab Med 2002;40:1056–60.
89. Burkhardt H, Bojarsky G, Gladisch R. Diagnostic efficiency of cystatin C and serum creatinine as markers of reduced glomerular filtration rate in the elderly. Clin Chem Lab Med 2002;40:1135–8.
90. Zago L, Dupraz H, Sarchi M, Rio M. The molar ratio of retinol-binding protein to transthyretin in the assessment of vitamin A status in adults. Proposal of a cut-off point. Clin Chem Lab Med 2002;40:1301–7.
91. Moher D, Jones A, Lepage L. CONSORT Group (Consolidated Standards for Reporting of Trials). Use of the CONSORT statement and quality of reports of randomised trial. A comparative before-and-after evaluation. JAMA 2001;285:1912–5.
92. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, et al. The revised CONSORT statement for reporting randomised trials: explanation and elaboration. Ann Intern Med 2001;134:663–94.
93. Knottnerus JS. The evidence base of clinical diagnosis. London: BMJ Books, 2002:237pp.
94. Feinstein AR. Misguided efforts and future challenges for research on "diagnostic tests". J Epidemiol Community Health 2002;56:330–2.
95. Hernández Aguado I. The winding road towards evidence based diagnoses. J Epidemiol Community Health 2002;56:323–5.

10.8.- Anexo VII: Lumbreras B., Pascual E., Frasquet J., González-Salinas J., Rodriguez E., Hernandez-Aguado I. Analysis for crystals in synovial fluid: training of the analysts results in high consistency. Annals of the Rheumatic Diseases. (en prensa)

Analysis for crystals in synovial fluid: training of the analysts results in high consistency

Blanca Lumbreras (1), Eliseo Pascual (2), Julia Frasquet (1), Josefina González-Salinas (1), Enrique Rodríguez (1), Ildefonso Hernández-Aguado (3)

(1) Department of Clinic Analysis, General University Hospital of Alicante, Alicante Spain

(2) Department of Rheumatology, General University Hospital of Alicante, Alicante Spain

(3) Department of Public Health, University of Miguel Hernández, San Juan de Alicante, Spain.

Abstract.

Introduction. Identification of monosodium urate crystals (MSU) and calcium pyrophosphate dehydrate crystals (CPPD) in synovial fluid (SF) samples is diagnostic of CPPD crystal related arthropathy and of gout. A number of studies have shown poor consistency in results of crystal analysis. We aim to determine whether training of the analysts increases the consistency.

Materials and methods. An expert rheumatologist gave a course on crystal detection and identification. Then the 4 trained observers blindly and independently examined SF samples previously classified by the expert, and obtained from patients with both crystal arthropathies, and other non-crystal related inflammatory joint conditions.

Results. A total of 194 observations were done on 64 SF samples; 96 on samples without crystals (49.4%); 55 with CPPD crystal (28.4%) and 43 with MSU crystals (22.2%). We obtained a sensitivity of 95,9% and a specificity of 86.5% in the crystal detection (presence or absence of crystals), a sensitivity for identification of MSU of 95,3%, with specificity of 97,2, and for CPPD crystals a sensitivity for identification of 92,7% with specificity of 92.1%. The kappa index of agreement between the observers with the reference standard was 0,84 in crystal detection; in crystal identification: MSU crystal samples, kappa = 0,93 and CPPD crystal samples kappa = 0,79. **Discussion**. For trained observers, the analysis of crystals in SF is a consistent procedure.

INTRODUCTION

Identification of monosodium urate crystals (MSU) and calcium pyrophosphate crystals in SF samples obtained from inflamed joints allows the precise diagnosis of gout and the calcium pyrophosphate dehydrate (CPPD) crystal related arthropathy.

MSU and CPPD crystals show different characteristics of birefringence, and the polarized light microscope constitutes the standard method for SF analysis in the search of crystals [1-3]; shape and appearance of MSU and CPPD crystals under the light microscope are different and contribute to their differentiation [4, 5]. Although MSU crystals are strongly birefringent and easily seen with the polarized light microscope, CPPD crystals are poorly birefringence, and many are not [6].

Diverse studies have shown that crystal identification suffers from a lack of consistency between different observers: four samples of different SF were sent to 25 laboratories for analysis and crystal identification, and a sensitivity of 78% for the detection of MSU crystals and only 12% for CPPD crystals was obtained [7]. In another study [8] eleven different aliquots of SF were distributed to 3 laboratories, and discrepancies were found in 7 of the 11 samples. In a third study, 16 SF samples without crystals, and samples with different concentrations of CPPD crystals (41 samples) and MSU crystals (42 samples) were analysed; although the correct identification of crystals increased with its concentration, the authors reported for MSU crystals a sensitivity of 69% and a specificity of 97% and for CPPD crystals, a sensitivity of 82% and a specificity of 78% [9]; In another study aliquots of SF containing MSU crystals and others SF with materials such as cholesterol crystals or starch particles, were sent to 25 laboratories; although all samples with MSU crystals were identified correctly, there was a 24% of false-positive results since other materials were taken as MSU crystals [10]. Finally in another study four samples of synovial fluid were distributed to several clinical laboratories between 1989 and 1996: If MSU crystals were abundant, the rate of false-positive results was 0-38%, and increased to 67% when these crystals were scanty [11]. All these results show a lack of consensus on the routine of SF analysis. [12-17]

The lack of consistency in SF analysis could be explained by a) the incapacity of the technique to give quality results, or b) it can be attributed to the observer, since crystal identification requires subjective interpretation, and the training of the observer

MATERIALS AND METHODS

Source of the SF samples: The study was prospective: 64 SF samples of patients with crystal related and other inflammatory arthropaties were collected at the clinics of the Rheumatology Section of the Hospital General Universitario de Alicante, Spain, between September 2001 and June 2003.

Participants: Four residents of the Department of Clinic Analysis. For the purposes of the study, the participating residents, which have not previous experience in SF analysis, received a training course described in the following paragraph.

Formative course: the course was given by one of us (EP), with experience in SF analysis. First the morphological and characteristics of birefringence of MSU and CPPD crystals were reviewed in a session, and then aliquots of SF samples with and without crystals were examined blindly by the residents along a three months period, at which time the trainer considered that the trainees identified both types of crystals properly. The guidelines followed in the course are described in the following paragraph.

Basic guidelines of the training course: At least for teaching purposes, we consider that the analysis for crystals has to be approached in two consecutive steps: a) *crystal detection* (to assure that crystals will not be missed), and b) *crystal identification*: to determine the identity of the crystals detected in the previous step. The reason to do so is that we feel that the *detection* MSU crystals is better done by means of the uncompensated polarized microscope, where all crystals show strong birefringence. By contrast, only about 20% of CPPD crystals show birefringence (and it is weaker than that shown by MSU crystals), and are better seen by means of the ordinary microscope (frequently they are intracellular, and should be searched there), which is the tool that should be used for their detection [6]. Then, in a search for crystals in SF, the samples should be observed by both means, or CPPD crystals have a probability of being missed. If crystals are detected, they should be *identified* by means of the compensated polarized microscope. It has to be remembered that the appearance of MSU and CPPD crystals is very different, and with few exceptions observers with experience have little difficulty differentiating them even with the ordinary light microscope [19]. We also considered that the

most common confounding element in crystal analysis are the common artefacts, and the trainees were familiarized with them; apatite, was not considered, Cholesterol was shown at the training course, but not seen afterwards, and steroid crystal were not present in any of the samples analyzed as a part of the study.

Procedure of the study: All samples were examined and classified by the reference rheumatologist (EP) who knew their origin. Then it was divided in aliquots and blindly and independently observed by the participant residents, who determined: a) presence or absence of crystals, and b) crystal type (MSU or CPPD), and annotated independently the results, which were not unblinded and analysed until the end of the study. All observations were done in the first two hours after extraction of the fluid. All the observations were made on a microscope Olympus BH to 400 x.

Statistical analyses: Statistical descriptions were performed with the statistical package for Windows 11,0 (SPSS, Inc., Chicago, Illinois). Kappa index was used for the analysis of concordance (statistical package STATA 8,0); the degree of concordance was expressed as a numerical value for k that can range from 0.0, indicating absolute discordance, to 1.0, indicating perfect concordance. (A value over 0,61 indicates that the agreement force is good).

RESULTS

Samples: A total of 64 samples of SF were analysed. The clinical diagnoses of the patients were gout (twelve patients), CPPD related arthropathy (sixteen patients), rheumatoid arthritis (twelve patients), and other inflammatory arthritis, including juvenile idiopathic arthritis, psoriatic arthritis, spondyloarthropathies and unclassified poliarthritides (twenty-three patients).

The four analysts made a total of 194 observations on SF; 96 observations on samples without crystals (49.4%); 55 observations on CPPD crystal samples (28.4%) and 43 observations on MSU crystal samples (22.2%).

Participants: Not all the observers examined the 64 samples: observer 1, 38 SF (59.4%); observer 2, 56 SF (87.5%); observer 3, 44 SF (68.8%) and observer 4, 56 SF (87.5%). That is, 4 observers analysed 20 samples (31.3%); 3 observers, 30 samples (46.8%); 2 observers, 10 samples (15.6%) and 1 observer, 4 samples (6.3%).

Crystal analysis

· Crystal detection (presence or absence of crystals): After examining each preparation with both ordinary light and uncompensated polarized microscope, we obtained a sensitivity of 95.9%, a specificity of 86.5% and 13 false-positive results (6.7%): 10 SF without crystals were identified as CPPD crystal samples and 3 SF without crystals as MSU crystal samples. (Values for each observer are summarized in table 1).

· Crystal identification (after detecting the presence of crystals they have to be identified as MSU, CPPD or eventually other type): We made 98 observations on SF samples with crystals. In the 43 observations on MSU crystal samples, we obtained a sensitivity of 95.3% and a specificity of 97.2%. In the 55 observations on CPPD crystal samples, we obtained a sensitivity of 92.7% and a specificity of 92.1%

· Agreement between observers: The index kappa of agreement between the observers with the reference standard was 0,85. In crystal detection, kappa was 0,84; in crystal identification kappa was 0,93 in MSU crystal samples and kappa was 0.79 on CPPD crystal samples. (Table 2)

·Time between the obtaining of the sample and its evaluation: Time average was 48 minutes (sd: 54).

DISCUSSION

The present data shows that when observers have been trained in crystal detection and identification in synovial fluid samples, their results are consistent. The two steps procedure followed by us for crystal analysis (1st crystal detection, 2nd identification of the crystal detected) may be determinant of these good results. CPPD crystals, where problems in detection and identification found in previous studies are more common, more often are non-refringent [6], and we consider that their detection has to rely in morphological identification under ordinary light microscopy, while MSU crystals are very well detected under uncompensated polarized microscopy, where they shine brightly on the dark field [4]. Compensated polarised microscopy allows adequate identification of the detected crystals. It has to be kept in mind that with rare exceptions the crystals responsible of arthritis are only of these two types [14]. Artefacts constitute an important confusing element for those inexperienced or untrained in synovial fluid analysis.

In interpreting our results, it has to be kept in mind that the observers had no previous experience with synovial fluid analysis, and that they carried out the observations that constitute the base of this study only after a short period of formal training. We feel that with additional experience the results would have been better. In fact, most of the misclassification occurred in the initial observations.

The concept of crystal detection (that is, if either CPPD or MSU crystals are present in a SF sample, their presence will not be missed) has not been used previously, since till now, a one step simultaneous detection and identification was the routine. Thus, we can not compare the results of our *detection* step with other studies, but our good sensitivity and specificity for crystal *detection* appears to support the scheme of approaching the search of crystals in the two steps strategy.

Our sensitivity and specificity for crystal identification has been higher than that found in previous studies. For MSU containing SF samples, we obtained a sensitivity of 95.3% and specificity for proper identification of 97.2%, in comparison with the sensitivity of 69% and specificity of 97% found previously [9]. For CPPD crystals we found a sensitivity of 92.7% and specificity for identification of 92.1%; in a previous study, a sensitivity of 12% was found [7]; another study produced better results, with a sensitivity of 82% and a specificity of 78% [9]. The experience in crystal analysis and level of training in it of the observers participating in these studies was not mentioned in the reports. We had some false positive results, 13 (6.7%); these are lower than those previously reported [10, 11].

It is in the CPPD crystals where the higher difficulty lies: when we used polarized microscopy for crystal, they were seen in only 52.7% of the samples that contained them. By contrast ordinary light microscopy detected the crystals in 94.5% of the SF analyzed. These data are in keeping with data previously reported, and outline the weight that the frequent absence of sufficient birefringence of the CPPD crystals has on their detection by polarized microscopy [6]. We detected well MSU, as also previously reported, and although in our results the detection rate was slightly higher with ordinary light, our feeling is that polarized light allows an easier detection, and should rutinarely be used for this purpose.

Final identification of the detected crystals was recorded only after observation with both ordinary and compensated microscopes, and we did not record a tentative identification at the *detection* step. Nevertheless, It is our feeling though that in most occasions, after observation with ordinary light the type of the crystal was already clear, as previously reported. By its rhomboidal or parallelepipedic shape, CPPD crystals are easily identified; by shape, only acicular CPPD crystals can be taken for the needle shaped MSU crystals, and it is here where the compensated polarized microscope has a definitive role [13]. None of the samples included in this report contained both MSU and CPPD crystals. According to our results, if a polarized light microscope is not at hand, a trained observer can do well in CPPD and MSU crystal detection and by means of the generally available ordinary light microscope.

Our results show a high level of concordance between the results obtained by the observers after their training, and the expert who had classified the SF samples for the study: global kappa was 0,85 (because the number of ratings per subject vary, we can not calculate tests statistics). Concordance was highest for MSU crystals (kappa 0,93), then for the samples with CPPD crystals (kappa 0,79), outlining again that it is in this type of crystals where the major difficulty lies.

A quality control program in crystal identification needs to be initiated, with special attention in sensitivity and specificity of the method. Ideally the quality program would have to incorporate the formation of the analysts, and of the procedure of examination. Quality control programmes now monitor most laboratory tests, but the interest showed in general on synovial fluid analysis by clinical laboratories has been low, and crystal analysis is not established as a sound routine in most of them. Rheumatologists themselves perform the analysis of the fluids that they draw; proper formal training to do so appear essential to assure the consistency of their results.

REFERENCES

1- Schumacher H R. Synovial fluid analysis. In: Kelley W N, Harris E D, Ruddy S, Sledge C B, eds. *Texbook of rheumatology*. 2nd ed. Philadelphia: Saunders, 1985:561-7.

2- McCarty DJ. Synovial fluid. In: McCarty D J, ed. *Arthritis and allied conditions*. 10 th ed. Philadelphia: Lea and Febiger, 1985: 54-75.

3- Gatter R A. *A practical handbook of joint fluid analysis*. Philadelphia: Lea and Febiger, 1984: 37-49.

4- McCarty DJ, Hollander JL: Identification of urate crystals in gouty synovial fluid. Ann Intern Med 1961;54:452-460.

5- Kohn NN, Hughes RE, Mc Carthy DJ Jr, Faires JS. The significance of calcium pyrophosphate crystals in the synovial fluid of arthritis patients: the “pseudogout syndrome”. II. Identification of crystals. Ann Intern Med. 1962;56:738-45.

6- Ivorra J, Rosas J, Pascual E. Most calcium pyrophosphate crystals appear as non-birefringent. Ann Rheum Dis 1999;58:582-84.

7- Hasselbacher P. Variation in synovial fluid analysis by hospital laboratories. Arthritis Rheum 1987;30:637-42.

8- Schumacher HR Jr, Sieck MS, Rothfuss S, Clayburne GM, Baumgarten DF, Mochan BS, Kant JA.Reproducibility of sinovial fluid analysis. Arthritis Rheum 1986;29:770-774.

9- Gordon C, Swan A, Dieppe P. Detection of crystals in synovial fluid crystal by light microscopy: sensitivity and reliability. Ann Rheum Dis. 1989; 48: 737-42.

10- Von Essen R, Holtta AM. Quality control of the laboratory diagnosis of gout by synovial fluid microscopy. Scand J Rheumatol 1990;19:232-4.

11- Von Essen R, Hölttä AMH, Pikkarainen R. Quality control of synovial fluid crystal identification. Ann Rheum Dis 1998;57:107-9.

12- Snaith ML. ABC of Rheumatology: gout, hyperuricaemia and crystal arthrtis. BMJ. 1995 Feb 25;310(6978):521-4.

13- Pascual E. The diagnosis of gout and CPPD crystal arthropathy. Br J Rheum 1996; 35:306-8.

14- Jaccard YB, Gerster JC, Calame L. Mixed monosodium urate and calcium pyrophosphate crystal-induced arthropathy. Rev Rhum Engl Ed. 1996;63:331-5.

15- Fam AG, MD. What is new about crystals other than monosodium urate? Curr Opin Rheumatol. 2000;12:228-34.

16- Dieppe P., Swan A. Identification of crystals in synovial fluid. Ann Rheum Dis 1999;58:261-3.

17- Segal B, Albert D. Diagnosis of crystal-induced arthritis by sinovial fluid examination for crystals: lessons from an imperfect test. Arthritis Care Res. 1999;12:376-80.

18- Buntinx F, Schouten HJ, Knottnerus JA, Crebolder HF, Essed GG. Interobserver variation in the assessment of the sampling quality of cervical smears. J Clin Epidemiol. 1993 Apr;46(4):367-70.

19- Pascual E, Tovar J, Ruiz MT. The ordinary light microscope: an appropriate tool for provisional detection and identification of crystals in synovial fluid. Ann Rheum Dis. 1989;48:983-5.

Table 1: Sensitivity and specificity in the crystal detection (to determine the crystal presence or absence) in 64 of synovial fluid samples analysed:

	OBSERVER 1	OBSERVER 2	OBSERVER 3	OBSERVER 4	TOTAL
TP	21	28	19	26	94
TN	13	25	21	24	83
FP	4	3	2	4	13
FN	0	0	2	2	4
Sensitivity (%)	100.0 (97.6 – 100.0)	100.0 (98.2 – 100.0)	90.5 (88.0 – 92.9)	92.8 (91.0 – 94.7)	95.9 (95.4 – 96.5)
Specificity (%)	76.7 (73.4 – 79.5)	89.3 (87.4 – 91.1)	91.3 (89.1 – 93.6)	85.7 (83.9 – 87.6)	86.5 (85.9 – 87.0)
PPV (%)	84.0 (81.9 – 86.1)	90.3 (88.7 – 92.0)	90.5 (88.0 – 92.9)	86.6 (84.9 – 88.4)	87.9 (87.4 – 88.4)
NPV (%)	100.0 (96.2 – 100.0)	100.0 (98.0 – 100.0)	91.3 (89.1 – 93.6)	92,3 (90.3 – 94.3)	95.4 (94.8 – 96.0)

TP= True positive; TN= True negative; FP= False positive; FN= False negative; PPV= positive predictive value; NPV= Negative predictive value.

Table 2: Index kappa of agreement between the four observers with reference standard for detection and identification crystal:

Observer	Agreement (%)	Expected agreement (%)	Kappa	Se*	CI**
1	86.84	33.45	0.80	0.11	[0.58-1.02]
2	96.43	36.10	0.94	0.09	[0.76-1.13]
3	90.91	39.36	0.85	0.11	[0.63-1.06]
4	89.29	36.54	0.83	0.09	[0.64 – 1.01]

**Se= Statistical error.*
***CI= Confidential interval.*