Cochrane Library

Cochrane Database of Systematic Reviews

Biomarkers for diagnosis of Wilson's disease (Review)

Ryan A, Nevitt SJ, Tuohy O, Cook P

Ryan A, Nevitt SJ, Tuohy O, Cook P.
Biomarkers for diagnosis of Wilson's disease.
Cochrane Database of Systematic Reviews 2019, Issue 11. Art. No.: CD012267.
DOI: 10.1002/14651858.CD012267.pub2.

www.cochranelibrary.com

WILEY

TABLE OF CONTENTS

HEADER 1
ABSTRACT 1
PLAIN LANGUAGE SUMMARY 2
SUMMARY OF FINDINGS 4
BACKGROUND 7
OBJECTIVES 9
METHODS 9
Figure 1. 11
RESULTS 13
Figure 2. 14
Figure 3. 15
Figure 4. 17
Figure 5. 18
DISCUSSION 19
AUTHORS' CONCLUSIONS 20
ACKNOWLEDGEMENTS 20
REFERENCES 21
CHARACTERISTICS OF STUDIES 24
DATA 35
Test 1. Caeruloplasmin (Leipzig criteria threshold 0.1 g/L). 36
Test 2. Caeruloplasmin (Leipzig criteria threshold 0.2 g/L). 36
Test 3. Caeruloplasmin (other threshold 0.14 g/L). 36
Test 4. Caeruloplasmin (other threshold 0.15 g/L). 36
Test 5. Caeruloplasmin (other threshold 0.18 g/L). 36
Test 6. Caeruloplasmin (other threshold 0.19 g/L). 36
Test 7. Hepatic copper (Leipzig criteria threshold > 4 µmol/g). 37
Test 8. Hepatic copper (other threshold 1.2 µmol/g). 37
Test 9. Hepatic copper (other threshold 1.5 µmol/g). 37
Test 10. Hepatic copper (other threshold 3.3 µmol/g). 37
Test 11. 24-hour urinary copper (Leipzig criteria threshold 0.64 µmol/24 hours). 37
Test 12. 24-hour urinary copper (Leipzig criteria threshold 1.6 µmol/24 hours). 37
Test 13. 24-hour urinary copper (threshold 0.8 µmol/24 hours). 37
Test 14. 24-hour urinary copper (threshold 1.06 µmol/24 hours). 37
ADDITIONAL TABLES 37
APPENDICES 46
CONTRIBUTIONS OF AUTHORS 52
DECLARATIONS OF INTEREST 52
SOURCES OF SUPPORT 52
DIFFERENCES BETWEEN PROTOCOL AND REVIEW 52

[Diagnostic Test Accuracy Review]

Biomarkers for diagnosis of Wilson's disease

Aidan Ryan[1], Sarah J Nevitt[2], Orla Tuohy[3], Paul Cook[1]

[1]Department of Clinical Biochemistry, University Hospital Southampton NHS Foundation Trust, Southampton, UK. [2]Department of Biostatistics, University of Liverpool, Liverpool, UK. [3]Wessex Neurological Centre, University Hospital Southampton, Southampton, UK

Contact address: Aidan Ryan, Department of Clinical Biochemistry, University Hospital Southampton NHS Foundation Trust, 17 Tremona Road, Southampton, SO16 6YD, UK. aidan.ryan@uhs.nhs.uk.

Editorial group: Cochrane Cystic Fibrosis and Genetic Disorders Group
Publication status and date: New, published in Issue 11, 2019.

Citation: Ryan A, Nevitt SJ, Tuohy O, Cook P. Biomarkers for diagnosis of Wilson's disease. *Cochrane Database of Systematic Reviews* 2019, Issue 11. Art. No.: CD012267. DOI: 10.1002/14651858.CD012267.pub2.

ABSTRACT

Background

Wilson's disease, first described by Samuel Wilson in 1912, is an autosomal recessive metabolic disorder resulting from mutations in the *ATP7B* gene. The disease develops as a consequence of copper accumulating in affected tissues.

There is no gold standard for the diagnosis of Wilson's disease, which is often delayed due to the non-specific clinical features and the need for a combination of clinical and laboratory tests for diagnosis. This delay may in turn affect clinical outcome and has implications for other family members in terms of diagnosis. The Leipzig criteria were established to help standardise diagnosis and management. However, it should be emphasised that these criteria date from 2003, and many of these have not been formally evaluated; this review examines the evidence behind biochemical testing for Wilson's disease.

Objectives

To determine the diagnostic accuracy of three biochemical tests at specified cut-off levels for Wilson's disease. The index tests covered by this Cochrane Review are caeruloplasmin, 24-hour urinary copper and hepatic copper content. These tests were evaluated in those with suspected Wilson's disease and appropriate controls (either healthy or those with chronic liver disease other than Wilson's). In the absence of a gold standard for diagnosing Wilson's disease, we have used the Leipzig criteria as a clinical reference standard.

To investigate whether index tests should be performed in all individuals who have been recommended for testing for Wilson's disease, or whether these tests should be limited to subgroups of individuals.

Search methods

We identified studies by extensive searching of, e.g. the Cochrane Central Register of Controlled Trials (CENTRAL), PubMed, Embase, the Web of Science and clinical trial registries (29 May 2019).

Date of the most recent search of the Cochrane Cystic Fibrosis and Genetic Disorders Inborn Errors of Metabolism Register: 29 May 2019.

Selection criteria

We included prospective and retrospective cohort studies that assessed the diagnostic accuracy of an index test using the Leipzig criteria as a clinical reference standard for the diagnosis of Wilson's disease.

Data collection and analysis

Two review authors independently reviewed and extracted data and assessed the methodological quality of each included study using the QUADAS-2 tool. We had planned to undertake meta-analyses of the sensitivity, specificity at relevant cut-offs for each of the biochemical

tests for Wilson's, however, due to differences in the methods used for each biochemical index test, it was not possible to combine the results in meta-analyses and hence these are described narratively.

Main results

Eight studies, involving 5699 participants (which included 1009 diagnosed with Wilson's disease) were eligible for inclusion in the review. Three studies involved children only, one adults only and the four remaining studies involved both children and adults. Two evaluated participants with hepatic signs and six with a combination of hepatic and neurological signs and symptoms of Wilson's disease, as well as pre-symptomatic individuals. The studies were of variable methodological quality; with high risk if bias for participant selection and the reference standard used being of greatest methodological concern. Key differences between studies include differences in assay methodology, different cut-off values for diagnostic thresholds, different age and ethnicity groups. Concerns around study design imply that diagnostic accuracy figures may not transfer to populations outside of the relevant study.

Index test: caeruloplasmin

Five studies evaluated various thresholds of caeruloplasmin (4281 participants, of which 541 had WD). For caeruloplasmin a cut-off of 0.2 g/L as in the Leipzig criteria achieved a sensitivity of 77.1% to 99%, with variable specificity of 55.9% to 82.8%. Using the cut-off of 0.1 g/L of the Leipzig criteria seemed to lower the sensitivity overall, 65% to 78.9%, while increasing the specificity to 96.6% to 100%.

Index test: hepatic copper

Four studies evaluated various thresholds of hepatic copper (1150 participants, of which 367 had WD). The hepatic copper cut-off of 4 µmol/g used in the Leipzig criteria achieved a sensitivity of 65.7% to 94.4%, with a variable specificity of 52.2% to 98.6%.

Index test: 24-hour urinary copper

Three studies evaluated various thresholds of 24-hour urinary copper (268 participants, of which 101 had WD). For 24-hour urinary copper, a cut-off of 0.64 to 1.6 µmol/24 hours used in the Leipzig criteria achieved a variable sensitivity of 50.0% to 80.0%, with a specificity of 75.6% to 98.3%.

Authors' conclusions

The cut-offs used for caeruloplasmin, 24-hour urinary copper and hepatic copper for diagnosing Wilson's disease are method-dependent and require validation in the population in which such index tests are going to be used. Binary cut-offs and use of single-test strategies to rule Wilson's disease in or out is not supported by the evidence in this review. There is insufficient evidence to inform testing in specific subgroups, defined by age, ethnicity or clinical subgroups.

PLAIN LANGUAGE SUMMARY

Laboratory blood, urine tests and liver biopsy used for the diagnosis of Wilson's disease in children and adults

Why is improving Wilson's disease diagnosis important?

Wilson's disease is an inherited disease that leads to a build-up of copper in affected parts of the body. Diagnosis usually occurs in children or young adults, but has been seen in adults over 60 years of age. Copper build-up begins in the liver progressing over time to affect the brain; however, the challenge for doctors is that liver disease in Wilson's disease has non-specific features and standard liver blood tests may be normal, even with advanced scarring of the liver or cirrhosis. Early diagnosis allows earlier treatment, however, other causes of chronic liver disease may cause false-positive results and, depending on cut-off values used for testing, may result in further unnecessary testing. Conversely, false-negative results may also arise when a single-test strategy for diagnosis is used, possibly leading to a delay in treatment.

What is the aim and what was included in this review?

We aimed to examine the accuracy of three commonly used diagnostic tests to correctly identify Wilson's disease. These tests are: caeruloplasmin (a protein that carries copper in blood); copper in the urine; and copper in the liver. Initial evaluation usually involves checking an individual's eyes for signs of Wilson's disease and a blood test for caeruloplasmin, as this is the most widely accessible biochemical test for Wilson's disease. However, the pathway to diagnosing Wilson's disease is highly variable. Follow-up testing depends on results of initial testing, plus the ability to access relevant tests and the likelihood with which the doctor believes the individual has Wilson's disease.

What are the main results in the review?

We found eight studies (5699 participants), of whom 1009 were diagnosed with Wilson's disease. One study assessed all three biochemical tests, three assessed caeruloplasmin, one assessed 24-hour urinary copper, two assessed hepatic copper and one assessed both urine and hepatic copper.

Four studies evaluated adults and children, three evaluated children and adolescents and one evaluated adults. The clinical presentation of Wilson's disease also varied: six studies evaluated individuals with both liver and neurological symptoms of Wilson's disease in addition to individuals who had not yet developed symptoms; and two studies evaluated individuals with liver symptoms only.

The ability of the three tests evaluated to detect those with Wilson's disease (termed sensitivity) was variable (50% to 94.4%); the ability to detect those without disease (termed specificity) was also variable (52.2% to 98.3%). No single test was capable of diagnosing Wilson's disease in isolation. There was also not enough evidence to determine the accuracy of the tests within different age groups or Wilson's disease subgroups (e.g. those with liver or neurological symptoms).

How reliable are the results of the studies in this review?

Since there is no gold standard test for diagnosing Wilson's disease, we selected a clinical and laboratory standard (the Leipzig criteria) to determine the diagnosis of the disease. Results of this review suggest that part of the variability in test sensitivity and specificity at the cut-offs in the Leipzig criteria is likely to be influenced by the method used to undertake the diagnostic tests. However, there were some problems with how the included studies were conducted. This may result in the caeruloplasmin, urine or liver copper appearing more accurate than it is, increasing the number of positive results (sensitivity).

What are the implications of this review?

Limited evidence from the included studies support the use of multiple-index testing as outlined in the Leipzig criteria. The diagnostic thresholds used in this criteria will vary with laboratory test, with the method used to conduct the laboratory test, and with the individuals in the included studies (who varied by age, ethnicity and clinical presentation of disease). These factors should therefore be taken into account when interpreting the results. High sensitivity (true-positive rate) for each of the laboratory tests is possible at particular cut-off values; however, when used in isolation, each laboratory test may have a false-positive or false-negative rate. Limitations in study design may exaggerate test accuracy.

How up-to-date is this review?

The authors searched for and used studies published up to 29 May 2019.

SUMMARY OF FINDINGS

Summary of findings 1. Summary of findings table - caeruloplasmin

Biomarkers for diagnosis of Wilson's disease

Population: people suspected of WD

Prior testing: Unclear

Setting: outpatients presenting with liver or neurological symptoms, or asymptomatic individuals

Index tests: caeruloplasmin

Target condition: WD

Reference standard: Leipzig criteria

Importance: to determine whether caeruloplasmin used an index test could result in earlier diagnosis and earlier treatment of WD, as well as reduce further necessary testing

Studies: observational (cohort, case-control) studies that assessed the diagnostic accuracy of an index test in the clinical context of the diagnosis of WD

Threshold	Summary / accuracy	Number of participants (studies)	Prevalence median (range)	Implications	Quality and comments
0.2 g/L	Sensitivity 77.1% - 99% Specificity 55.9% - 82.8%	4120 (4)	0.36 (0.08 - 0.46)	Single test strategy inadequate to rule out or rule in WD	Overall methodological quality of included studies was variable with high risk of bias for participant selection and reference standard and low risk of bias for index test and flow and timing. Concerns around study design imply that diagnostic accuracy figures may not transfer to populations outside of the relevant study.
0.1 g/L	Sensitivity 65% - 78.9% Specificity 96.6% - 100%	293 (2)	0.36 (0.33 - 0.41)	Single test strategy inadequate to rule out or rule in WD	Overall methodological quality of included studies was variable with high risk of bias for participant selection and reference standard and low risk of bias for index test and flow and timing. Concerns around study design imply that diagnostic accuracy figures may not transfer to populations outside of the relevant study.

Abbreviations: WD: Wilson's disease

Summary of findings 2. Summary of findings table - hepatic copper

Biomarkers for diagnosis of Wilson's disease					
Population: people suspected of WD **Prior testing**: unclear **Setting**: outpatients presenting with liver or neurological symptoms, or asymptomatic individuals **Index tests**: hepatic copper **Target condition**: WD **Reference standard**: Leipzig criteria **Importance**: to determine whether hepatic copper used an index test could result in earlier diagnosis and earlier treatment of WD, as well as reduce further necessary testing **Studies**: observational (cohort, case-control) studies that assessed the diagnostic accuracy of an index test in the clinical context of the diagnosis of WD					
Threshold	**Summary / accuracy**	**Number of participants (studies)**	**Prevalence** **median (range)**	**Implications**	**Quality and comments**
4 µmol/g	Sensitivity 65.7-94.4% Specificity 52.2-98.6%	1150 (4)	0.38 (0.28 - 0.46)	Single test strategy inadequate to rule out or rule in WD	Overall methodological quality of included studies was variable with high risk of bias for participant selection and reference standard and low risk of bias for index test and flow and timing. Concerns around study design imply that diagnostic accuracy figures may not transfer to populations outside of the relevant study.

Abbreviations: WD: Wilson's disease

Summary of findings 3. Summary of findings table - 24-hour urinary copper

Biomarkers for diagnosis of Wilson's disease
Population: people suspected of WD **Prior testing**: unclear **Setting**: outpatients presenting with liver or neurological symptoms, or asymptomatic individuals **Index tests**: 24-hour urinary copper

Target condition: WD

Reference standard: Leipzig criteria

Importance: to determine whether 24-hour urinary copper used an index test could result in earlier diagnosis and earlier treatment of WD, as well as reduce further necessary testing

Studies: observational (cohort, case-control) studies that assessed the diagnostic accuracy of an index test in the clinical context of the diagnosis of WD

Threshold	Summary / accuracy	Number of participants (studies)	Prevalence median (range)	Implications	Quality and comments
0.64 µmol/24 hours	Sensitivity 78.9% Specificity 87.9%	96 (1)	0.40 (NA)	Single-test strategy inadequate to rule out or rule in WD	Overall methodological quality of included studies was variable with high risk of bias for participant selection and reference standard and low risk of bias for index test and flow and timing. Concerns around study design imply that diagnostic accuracy figures may not transfer to populations outside of the relevant study.
1.6 µmol/24 hours	Sensitivity 50% - 80% Specificity 75.6% - 98.3%	268 (3)	0.41 (0.28 - 0.46)	Single-test strategy inadequate to rule out or rule in WD	Overall methodological quality of included studies was variable with high risk of bias for participant selection and reference standard and low risk of bias for index test and flow and timing. Concerns around study design imply that diagnostic accuracy figures may not transfer to populations outside of the relevant study.

Abbreviations: NA: not applicable; WD: Wilson's disease

BACKGROUND

Wilson's disease (WD), first described by Samuel Wilson in 1912, is an autosomal recessive metabolic disorder resulting from mutations in the *ATP7B* gene which encodes a protein pathway involved in copper hepatic metabolism (Bull 1993; Compston 2009). This secretory pathway involves both copper excretion into bile and its incorporation into apocaeruloplasmin for the synthesis of functional caeruloplasmin (Davis 1996). The disease develops as a consequence of the accumulation of copper in affected tissues and therefore hepatic disease presents earlier than neurological disease (Ferenci 2019).

The diagnosis of WD is often delayed, it may take over three years due to the non-specific clinical features and the requirement for combination testing for diagnosis (Ferenci 2019). This delay may in turn affect outcome and has implications for other family members in terms of diagnosis. Recent work has shown that up to 40% of children but 58% of adults were cirrhotic at diagnosis (Ferenci 2019). The diagnosis of WD depends on a combination of clinical, biochemical, histological and genetic testing and analysis. The Leipzig criteria (Table 1) were established to help standardise diagnosis and management of WD (Ferenci 2003). However, it should be emphasised that these criteria date from 2003, and many of these have not been formally evaluated; this review examines the evidence behind biochemical testing for WD (Mak 2008).

Target condition being diagnosed

The clinical presentation of WD can vary widely in terms of symptoms, signs at presentation and age of onset of such features, with clinically evident liver disease often preceding neurological disease by up to a decade (EASL 2012). The key clinical diagnostic features used to form the basis of the Leipzig criteria include liver disease, motor and neuropsychiatric disturbances, corneal Kayser-Fleischer (KF) rings and acute haemolysis (in association with acute liver failure) (Ferenci 2003). The original classical neurological presentations as described by Wilson, were characterised by a movement disorder in the setting of characteristic biochemical abnormalities and often contrast with the non-specific protean hepatic manifestations.

Index test(s)

The most commonly used initial diagnostic test for WD is caeruloplasmin, with a concentration of less than 0.2 g/L considered as the conventional diagnostic cut-off (EASL 2012). However, the lower reference limit can vary with different assay types and age; this may be reduced in other causes of chronic liver disease, copper deficiency or in protein-losing states. Caeruloplasmin was originally described as an acute-phase protein with diverse functions (Hellman 2002). The initial protein produced is an inactive, unstable non-copper bound form, apocaeruloplasmin. Following the addition of copper by *ATP7B*, the functional more stable product holocaeruloplasmin is formed. The type of assay used has important implications for WD diagnosis. Immunoassays are commonly used for measuring caeruloplasmin, and measure both apo- and holocaeruloplasmin forms; however, caeruloplasmin oxidase-based methods only measure the holocaeruloplasmin form. Therefore, immunoassays may theoretically lead to an overestimate as compared with the caeruloplasmin oxidase-based method. However, the lack of widespread availability of the oxidase method limits it's use (Gnanou 2006; Walshe 2003).

Turbidimetry and nephelometry are common immunoassay methods used to measure many plasma proteins including caeruloplasmin (Tietz 2012a). In principle an individual's sample is added to a combination of assay reagent and antibodies to the analyte or antigen of interest (in this case caeruloplasmin), which results in precipitation of an immune complex that increases the turbidity (cloudiness) of the sample. By shining a light through the sample and with appropriate calibration, the level of analyte can be determined. In turbidimetry the absorbance of the light by the sample is measured, whereas in nephelometry the light scatter is measured at a fixed angle. In general the analyte concentration is inversely proportional to the transmitted light signal. Published data, based on UK external quality assurance scheme results, have shown that for caeruloplasmin (depending on chosen method) the same sample may vary +/-20% depending on the chosen immunoassay method; with Olympus, Roche Intergra showing a negative bias and Beckman Immage, Dade Behring, Abbott platforms showing a positive bias (Zegers 2013).

24-hour urinary copper studies are often used as a follow-up to abnormal caeruloplasmin testing. In the absence of renal impairment, urinary copper reflects the amount of non-caeruloplasmin bound copper. Cut-off values of more than 1.2 times the upper limit of normal (ULN) or more than 2 ULN have been suggested as indicating possible WD (EASL 2012). Use of such cut-off values is problematic and method-dependent, with up to 25% of people with WD (especially children) having urinary copper levels less than this (Nicastro 2010). Urinary copper excretion may also be increased in other causes of chronic liver disease (LaRusso 1976).

Hepatic copper accumulation is the hallmark and earliest manifestation of WD. Copper distribution within hepatic parenchyma may not be homogenous, may be susceptible to sampling error by biopsy and may be elevated in other liver disorders, particularly those involving cholestasis (Roberts 2008). Cut-off values have been suggested, but again these are method-dependent and as yet not fully validated (EASL 2012). The use of specific stains, e.g. rhodamine or orcein, reveal focal copper deposition in less than 10% as these stains only detect lysosomal deposition. Previous studies have suggested that a level of hepatic copper greater than 4 µmol/g is considered the best evidence for a diagnosis of WD; however, there is some evidence that such a threshold may need to be lower in order to increase sensitivity (Yang 2015).

Two of the more common techniques that are used to measure urinary and hepatic copper are atomic absorption and inductively coupled plasma mass spectrometry. Atomic absorption spectroscopy uses the absorption of light to measure the concentration of gas phase atoms, which is achieved by vaporising copper in a flame. The atoms absorb the energy generated from the flame, making transition to higher energy levels with the copper concentration being determined from the amount of absorption (Tietz 2012b). Inductively coupled plasma mass spectrometry combines a high temperature source (inductively coupled plasma (ICP)) with a mass spectrometer. The ICP converts the atoms of the relevant element in the sample, in this case copper, to ions and these ions are separated and detected using mass/charge ratio by a mass spectrometer (Tietz 2012c). This has been shown to be more sensitive than atomic absorption with lower limits of detection, larger linear range and is increasingly become the method of choice to measure trace elements such as copper. Again, associating any cut-off value with the relevant method is important for urinary copper, with

the recent UK trace element quality assurance scheme showing a variability of +/-17% depending on the choice of atomic absorption or ICP mass spectrometry (UK NEQAS 2018). Unfortunately, due to limited sample availability, such schemes are not available for hepatic copper to show inter-assay variability.

Clinical pathway

In the original paper outlining the Leipzig criteria, there is no agreed clinical reference standard pathway for the diagnosis of WD, and hence this has not been documented (Ferenci 2003). There is a diagnostic algorithm in the European Association for the Study of the Liver (EASL) guidelines on WD, but again this is not a complete pathway that outlines test selection and stages of testing proceeding to diagnosis (EASL 2012). The EASL guidelines do comment that a combination of tests are required and that, reflecting the challenge of diagnosing WD, no single clinical sign or laboratory test is diagnostic in isolation.

A recent cohort study of 1359 people with WD has helped improve our understanding of how WD presents clinically (Ferenci 2019). The study participants were 702 children and 655 adults (679 males and 678 females). Initially, the authors discovered an asymptomatic hepatic involvement, which may progress and become symptomatic with 39.5% of children and 58% of adults affected with cirrhosis. Neurological disease occurred later and was more common in males, with hepatic disease being more common in females. Overall, the mean age for presentation with chronic liver disease was 17.8 years, for decompensated cirrhosis was 25.9 years, and for those with a neurological presentation was 23.8 years. The authors also noted that delays of over three years for diagnosis were not uncommon and that often individuals were diagnosed following an initial presentation with asymptomatic aminotransferase elevation (Ferenci 2019).

The mainstay of WD therapy is copper chelation (e.g. penicillamine) or medication that interferes with copper absorption, such as zinc (Członkowska 2018). Those individuals with cirrhosis may be considered candidates for hepatic transplantation. Medical treatment for WD is lifelong, with success being dependent on clinical features at presentation, with more than 90% improvement in those treated before advanced liver disease or neurological involvement, dropping to 50% in those with neurological disease at presentation (EASL 2012).

Alternative test(s)

This autosomal recessive disorder of copper transport is due to mutations in the *ATP7B* gene. The worldwide prevalence of WD has previously been cited as 1 in 30,000, with a carrier frequency of 1 in 90; however, these figures pre-date the discovery of the *ATP7B* gene and more recent work has cited a higher frequency of 1 in 7000 for genetic diagnosis (Coffey 2013). Following extensive linkage and positional cloning studies, the *ATP7B* gene was located on chromosome 13q14.3 (Bull 1993). The gene has 21 exons with more than 10000 base pairs. The molecular analysis of individuals and families affected by WD have demonstrated that, to date, there are up to 500 disease-causing mutations (Coffey 2013).

The problem with the collation of such mutations and variants is the lack of control participants tested in studies, which then inaccurately reported new variants. Recommendations of a minimum of 100 normal chromosomes from the same ethnic population to be tested are often not followed (Kenney 2007). Whilst most of the pathogenic mutations discovered to date are rare and only reported in single families, some of these are more common and account for large numbers of WD cases. Most affected individuals are compound heterozygotes. These mutations mainly affect the transmembrane region and largely consist of missense and stop mutations (Kenney 2007; Thomas 1995). However, strict genotype-phenotype remains unproven, even within families; and therefore, other genetic modifiers are believed to be at play (Czlonkowska 2009; Huster 2012). A recent study by Ferenci evaluating 1359 people with WD failed to show any link between genotype and clinical presentation, suggesting that factors such as age and sex are more important in how individuals present (Ferenci 2019).

In the presence of definite clinical or biochemical abnormalities, the identification of only one of the two disease-causing genes may be adequate to confirm diagnosis. However, if the significance of the initial identified mutation is doubtful, the second mutation should be identified (EASL 2012). It should be noted that in order to infer pathogenicity, a mutation must clearly be disease-causing and not just a common missense variant. Hence, the importance of normal ethnic controls. Developments in next-generation sequencing may allow faster sequencing and better coverage; however, large numbers of variants of uncertain significance may be generated and relevant standardised functional methods to test these remain to be clearly established.

As discussed earlier, it is the failure of incorporation of copper into its carrier proteins the leads to low caeruloplasmin and low serum copper, however, the proportion of unbound copper is increased. Historically, this was calculated as non-caeruloplasmin copper (NCC (µmol/L) = total copper (µmol/L) - n(µmol/mg) x caeruloplasmin (mg/L) where n is the factor for copper bound/mg of ceruloplasmin) as it was not possible to measure this (Twomey 2005). This NCC was used for diagnosing and managing people with WD on chelation therapy. However, recent work has questioned the variability and reliability of NCC for both diagnosis and therapeutic monitoring (Duncan 2016; Pfeiffenberger 2019). The latter study concluded that for therapeutic monitoring, NCC offered no benefit over 24-hour urine copper alone.

Exchangeable copper and its derived relative exchangeable copper (REC) have recently been proposed as a new biomarker for diagnosing WD as a method of measuring free copper described above. Specfically, REC has been shown to provide a high sensitivity and specificity for WD (El Balkhi 2011). The exchangeable copper corresponds to the labile fraction of copper bound mainly to albumin as well as free unbound copper. An increase in this fraction above normal is thought to reflect a blood and tissue copper overflow into the blood due to hepatic damage. This test has, however, only been evaluated in small groups and further validation will be required, particularly its specificity in other causes of chronic liver disease. The convenience of a reliable serum marker for diagnostic purposes is highly desirable for use in WD work up. A follow-up paper by the same group has shown that exchangeable copper may be of use in differentiating neurological from hepatic WD, with those having neurological WD having higher exchangeable copper (Poujois 2017).

A key clinical feature in the diagnosis of WD is the presence of KF rings, occurring in 100% of individuals with neurological disease and less frequently in those with liver disease (Taly 2007). The phenomenon arises as a consequence of copper deposition in the Descemets membrane and indicates that free copper has been re-

leased into the individual's circulation. Visualisation requires the use of slit-lamp amplification (Walshe 2011).

Due to the invasive nature of liver biopsy, this is no longer commonly undertaken for routine diagnosis of WD; previous studies have shown that up to 40% of individuals at presentation may have cirrhosis (Merle 2007). Early histological changes of WD are non-specific and represent a spectrum that may include hepatic steatosis, chronic hepatitis and fibrosis and may add to diagnostic delay.

Acute liver failure due to WD is an important diagnosis to make early, affecting both the management of the individual and also enabling screening and diagnosis of other family members (Ostapowicz 2002). Acute hepatic failure in WD gives rise to many characteristic biochemical and haematological abnormalities, due to the toxic effect of an acute copper release from hepatocyte lysis. The laboratory findings of fulminant WD previously described have included Coomb's negative haemolytic anaemia, low serum alkaline phosphatase and increased aspartate to alanine aminotransferase ratios (Berman 1999; Korman 2008; Lee 1998; Wilson 1987).

Rationale

Consensus guidelines for diagnosing WD exist. However, many of the criteria have not been formally evaluated and issues such as sensitivity and specificity for index tests remain to be fully explored.

OBJECTIVES

To determine the diagnostic accuracy of the index tests for WD. The index tests covered by this Cochrane Review are caeruloplasmin, 24-hour urinary copper and hepatic copper content.

Secondary objectives

We have two main secondary objectives, to investigate whether index tests should be performed in all individuals who have been recommended for testing for WD and whether these tests should be limited to subgroups of individuals (see Methods investigation of heterogeneity). We discuss differences in cut-off values and assay types as these are likely to have the most influence on heterogeneity. We anticipated that the study reports and number of papers selected would lack the necessary detail and volume to undertake meaningful subgroup analysis.

METHODS

Criteria for considering studies for this review

Types of studies

We included observational (cohort, case-control) studies that assessed the diagnostic accuracy of an index test in the clinical context of the diagnosis of WD. These studies included those with WD or suspected WD versus a normal population but also included heterozygotes where genetic testing was available. We excluded studies that evaluated the index test in a normal population without a WD comparator group or in use of diseases other than WD.

Participants

Children and adults of all ages with suspected WD evaluated by the Leipzig criteria were eligible for inclusion (Ferenci 2003). Studies that did not use the Leipzig criteria or failed to define how WD was defined and those with the acute fulminant form were excluded.

Index tests

The diagnostic accuracy of caeruloplasmin, urinary copper and liver copper content was evaluated for diagnosing WD. The thresholds for a positive score of each of these index tests, according to the Leipzig criteria, are provided in an additional table (Table 1).

Target conditions

WD as defined by the Leipzig criteria (Ferenci 2003); details are presented in an additional table (Table 1).

Reference standards

The clinical reference standard is the diagnosis of WD as outlined by the Leipzig criteria (Ferenci 2003).

Search methods for identification of studies

We searched for all relevant published and unpublished studies without restrictions on language, year or publication status.

Electronic searches

We searched for relevant studies from the Cochrane Cystic Fibrosis and Genetic Disorders Group's Inborn Errors of Metabolism Trials Register using the term: Wilson*:kw.

The Inborn Errors of Metabolism Trials Register is compiled from electronic searches of the Cochrane Central Register of Controlled Trials (CENTRAL) (updated with each new issue of *The Cochrane Library*), weekly searches of MEDLINE and the prospective handsearching of one journal - Journal of Inherited Metabolic Disease. Unpublished work is identified by searching through the abstract books of the Society for the Study of Inborn Errors of Metabolism conference and the SHS (Scientific Hospital Supplies) Inborn Error Review Series. For full details of all searching activities for the register, please see the relevant section of the Cochrane Cystic Fibrosis and Genetic Disorders Group's website.

Date of last register search: 29 May 2019.

We also searched the following databases and trial registries:

- Cochrane Central Register of Controlled Trials (CENTRAL; 2017 Issue 8) and Cochrane Database of Systematic Reviews (CDSR) in the Cochrane Library www.cochranelibrary.com (searched 29 May 2019);
- PubMed (www.ncbi.nlm.nih.gov/pubmed (1946 to 29 May 2019);
- CINAHL EBSCO (1982 to 29 May 2019);
- Embase Ovid (1982 to 29 May 2019);
- Science Citation Index via the Web of Science (1898 to 29 May 2019);
- Web of Science's Conference Proceedings Citation Index (CPCI; 1900 to 29 May 2019);
- British Library's ZETOC (zetoc.jisc.ac.uk/wzgw?db=etoc; 1993 to 29 May 2019) for conference abstracts;
- PROSPERO (International Prospective Register of Systematic Reviews www.crd.york.ac.uk/prospero/search.asp; searched 29 May 2019);
- US National Institutes of Health Ongoing Trials Register Clinicaltrials.gov (www.clinicaltrials.gov; searched 29 May 2019);

- World Health Organization International Clinical Trials Registry Platform (WHO ICTRP) (apps.who.int/trialsearch; searched 29 May 2019).

For details of our search strategies, please see Appendix 1.

Searching other resources

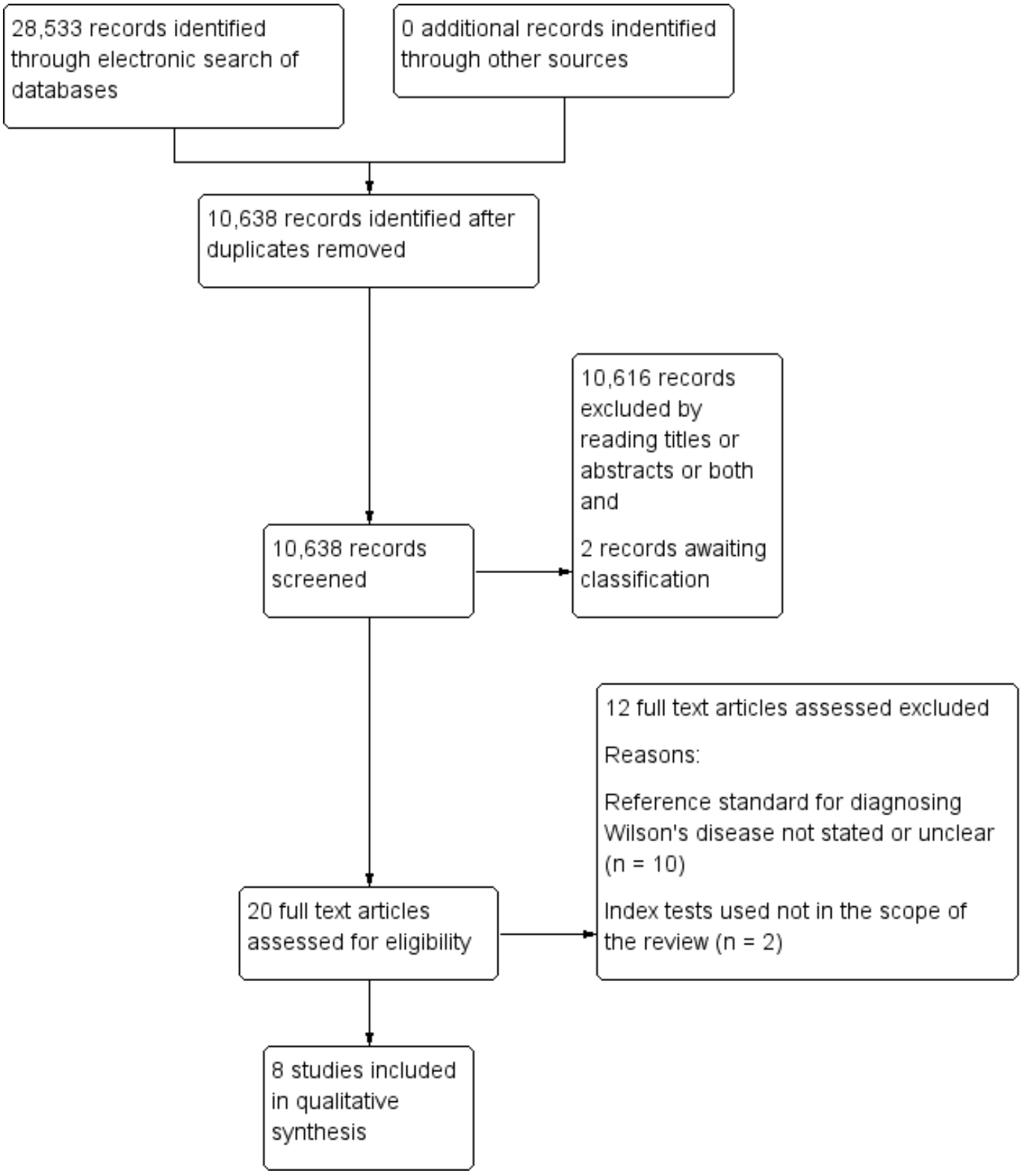
The reference lists of all included articles and relevant systematic reviews were reviewed to identify additional studies not found through the electronic review.

Data collection and analysis

Selection of studies

Two authors (AR, OT) independently reviewed the titles and abstracts of articles found in the Electronic searches for potentially eligible studies for review. The same two authors independently assessed full manuscripts against the inclusion criteria and discrepancies were resolved by discussion. We present a PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram to outline the decision-making process for including studies in the review (Figure 1; PRISMA 2009).

Figure 1. Study flow diagram.



Data extraction and management

Two authors (AR, SN) independently extracted the following data from published articles using a pre-determined extraction form, with discrepancies being resolved by discussion.

- First study author and year (of primary reference)
- Study eligibility
- Participant and method characteristics
- Number of participants

- Clinical and demographic characteristics (age, clinical presentation (hepatic versus neurological), ethnicity)
- Details of index text (assay type, control, cut-off values)
- Details of the reference standard
- Methodological quality of included studies

We created 2 x 2 tables for each method of the index test described in this review, cross-tabulating index test results with presence of the target condition (reference standard), please refer to the relevant appendix for the the format.

The data extraction form incorporated a quality assessment section comprising items from Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) (Whiting 2011).

Where a study did not present all relevant data for creating a 2 x 2 table, we planned to contact the study authors directly to request this information. If study authors were unable to provide the information, we planned to retain the study in the narrative section of the review, but not include it in any meta-analysis.

Assessment of methodological quality

Two authors (AR, SN) independently assessed the methodological quality of each included study using the QUADAS-2 tool (Whiting 2011) as recommended by Cochrane. We resolved any disagreements by discussion. The tool is made up of four domains:

- participant selection;
- index test;
- reference standard;
- participant flow and timing.

We assessed each domain in terms of risk of bias, with the first three domains also considered in terms of applicability concerns. We present review-specific signalling questions and appropriate terms concerning the applicability of primary studies related to this review, together with guidance about ratings, in an appendix (Appendix 2).

Statistical analysis and data synthesis

For all included studies, we used the data in the 2 x 2 tables to calculate two statistics for each index test in each study, as detailed above.

For the definitions of all abbreviations used in this section, please see the table below.

Sensitivity = number of TP / number of participants with the target condition present (TP + FN). The higher the sensitivity of a test at a particular cut off value, the better the test is at correctly identifying individuals who have the target condition.

Specificity = number of TN / number of participants without the target condition present (TN + FP). The higher the specificity of a test at a particular cut off value, the better the test is at correctly identifying individuals who do not have the target condition.

LR+ = sensitivity/1-specificity. This is the ratio of TP to FP for a particular test at a particular cut off value with LR+ greater than 1 being preferred.

LR- = 1-sensitivity/specificity. This is the ratio of FN to TN for a particular test at a particular cut off value with LR- closer to 0 being preferred.

PPV = number of TP/number of TP+FP. This measures the probability that a person with a positive test result has the disease.

NPV = number of TN/number of TN+FN. This measures the probability that a person with a negative test result does not have the disease.

We intended that these estimates would be used to create ROC and forest plots for all studies, however given the heterogeneity of index tests and cut-offs used in the identified studies, we consider that it is more appropriate to describe the above statistics narratively.

Given the lack of validated cut-offs of the index tests (*see* Index tests), we expected variability in cut-off points chosen in the included studies. As this review uses a clinical reference standard based on the Leipzig criteria, we propose to record diagnostic accuracy figures for each index test where available in each of the studies evaluated. Other cut off values based on ROC curve analysis evaluated in each of the studies will also be considered.

Therefore, we proposed to meta-analyse pairs of sensitivity and specificity using the HSROC model (Rutter 2001), which would allow for the possibility of variation in threshold between studies, while also accounting for variation within and between studies and any potential correlation between sensitivity and specificity. However, the number of studies for each index test method was limited and the methods used for the Index test assays were varied (see Table 2). Therefore, we deemed that it was not appropriate to pool any results in meta-analysis and results of the review are described narratively.

Abbreviation	Term
FN	false negatives
FP	false positives
HSROC	hierarchical summary receiver operating characteristic
LR+	positive likelihood ratio
LR-	negative likelihood ratio

NPV	negative predictive value
PPV	positive predictive value
ROC	receiver operating characteristic
TN	true negatives
TP	true positives

Investigations of heterogeneity

We planned to investigate the following subgroups:

- age (to include all ages);
- gender;
- ethnicity;
- clinical features (pre-symptomatic, hepatic and or neurological);
- index test method;
- different study designs.

In exploratory analyses, we planned to visually examine forest plots of sensitivity and specificity for each index test, and summary ROC plots to explore the effect of each of the factors of interest. If there were sufficient studies, we planned to perform meta-regression by including each potential source of heterogeneity as a covariate in the HSROC model. However, again we were unable to carry out our planned analyses owing to an insufficient number of studies and heterogeneous nature of the index test methods which could not be combined in meta-analyses.

Sensitivity analyses

If appropriate, we planned to perform sensitivity analyses excluding studies which are at a high risk of bias for at least one domain of the QUADAS-2 tool (see Assessment of methodological quality). Again, given the small number of studies, we did not attempt to perform any sensitivity analyses.

Assessment of reporting bias

We did not formally plan to investigate reporting bias via existing analytical tools such as funnel plots due to current uncertainty around interpretation of such tools in this setting. Instead, we performed systematic electronic searches and detailed searches of other published and unpublished sources (see relevant sections above) in order to retrieve as many eligible studies as possible for inclusion in the review.

Summary of findings of the review

We have summarised the results of the review for each index test in a summary of findings table for pre-specified thresholds of each index test, based on the clinical reference standard (Leipzig criteria) (Summary of findings table 1; Summary of findings table 2; Summary of findings table 3). The tables summarise the following information: threshold, summary and accuracy, number of participants and studies, prevalence, implications, quality and comments.

RESULTS

Results of the search

Out of 10,638 records (following removal of duplicates), we excluded 10,616 clearly irrelevant records. We obtained and scrutinised a total of 20 full-text reports to assess their eligibility for inclusion in this review and two further reports await classification (Aksu 2018; Zhou 2019) as illustrated in the PRISMA flow diagram (Figure 1). We excluded 12 studies which did not meet the eligibility criteria and have provided details of the reasons for exclusion of each of these studies in the characteristics of excluded studies table. In brief a common reason for exclusion was the fact that studies failed to define how participants were diagnosed with WD or did not use the Leipzig criteria for WD case definition (Causa 1997; Frommer 1981; Gibbs 1979; Gnanou 2006; Lech 2007; Li 1983; Liggi 2013; Mahjoub 2012; Markowitz 1955; Mzhel'skaia 1994). Another common reason was that the method used in the paper were not index tests in the current review (Prasad 1998; Siotto 2014).

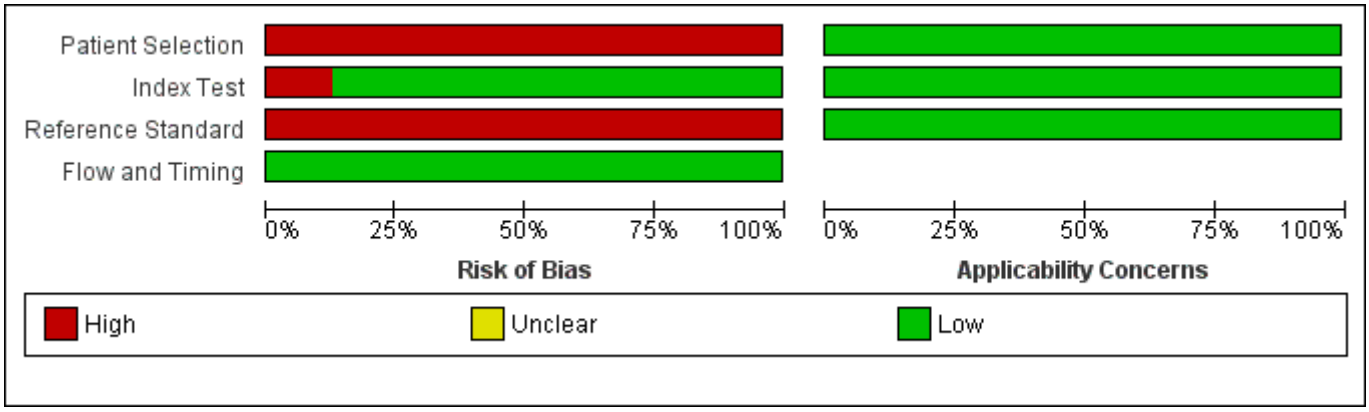
We have included eight studies in the review, however, due to the limited number of studies and methodological differences such as use of different thresholds, we have been unable to undertake any meta-analysis (Ferenci 2005; Mak 2008; Merle 2009; Lu 2010; Nicastro 2010; Sezer 2014; Yang 2015; Xu 2018). These studies were case-control or cross-sectional in nature, evaluating sensitivity, specificity, positive and negative predictive value of the particular test for the diagnosis of WD.

Four studies evaluated hepatic copper using flame atomic absorption spectroscopy (Ferenci 2005; Nicastro 2010; Sezer 2014; Yang 2015), three evaluated caeruloplasmin using nephelometry (Mak 2008; Merle 2009; Xu 2018) and one using radial immunodiffusion (Nicastro 2010). Two studies evaluated 24-hour urine copper using atomic absorption spectroscopy (Nicastro 2010; Sezer 2014), another study used inductively coupled mass spectrometry (Lu 2010). The characteristics of the included studies are summarised in an additional table (Table 2).

Methodological quality of included studies

We judged the studies to be of overall low risk of bias in two of the domain categories of QUADAS-2 (index test, and flow and timing) and to be of high risk of bias for the domains patient selection and reference standard (Figure 2; Appendix 2).

Figure 2. Risk of bias and applicability concerns graph: review authors' judgements about each domain presented as percentages across included studies



The main concern was around participant selection bias with convenience sampling being the most common method of participant selection. However, this is not uncommon in clinical studies particularly with uncommon diseases such as WD where considerable time may be taken for both diagnosis and to build up positive diagnoses. At the same time, more detail regarding phenotype of cases (beyond generic hepatic or neurological) and controls (to include severity-synthetic function in those with chronic liver disease) would assist with mitigating this selection bias.

Due to the nature of testing for WD which depends on clinical, biochemical and genetic results to generate the Leipzig score, there is potentially a risk of bias being aware of the clinical reference standard prior to the conduct of the index test. However, this risk is heavily dependent on the subjective nature of the index test which does not apply to quantitative biochemical testing evaluated in this review. For seven of the eight studies, threshold values were prespecified prior to analysis as outlined for each index test under the Leipzig criteria. The importance of this is that pre-specification limits its potential for over fitting diagnostic accuracy figures that can limit external validity of the study. For one study there was no prespecified index test threshold and cut-off value was optimised after analysis of the index test and on the basis of this has been classified as high risk bias for index test part of QUADAS-2 (Merle 2009) (Figure 2; Appendix 2).

The biochemical index tests evaluated in this review contribute to the Leipzig score of each of the participants evaluated in the relevant study. As a consequence of being a clinical reference standard prior knowledge of the index result is likely to have occurred in order to generate the Leipzig score. As outlined in this review no single biochemical diagnostic test is capable of diagnosing WD in isolation. The Leipzig criteria depends on a combination of clinical, biochemical and genetic testing. Closer inspection of the Leipzig criteria (Table 1) shows that one test (mutational analysis) on its own had potential to bias the Leipzig testing pathway, which could (for a mutation on two chromosomes) generate a score of four, which establishes a WD diagnosis and thus could in theory stop further testing. This, however, under appreciates that clinicians using the Leipzig score pathway will have to use combination testing generally to make a diagnosis and none of the biochemical testing would have the same effect as genetic testing in terms of potential bias. That being said, we acknowledge that some of the criteria for the Leipzig criteria are subjective, namely the clinical and histological criteria and that as a consequence, knowledge of these test results prior calculating the Leipzig score could bias the more subjective elements of the Leipzig criteria.

We grouped QUADAS-2 quality assessment items into four domains: participant selection, index test, reference standard, flow and timing. The main source of bias arose from patient selection and reference standard as discussed above (see Figure 2 and Appendix 2).With regard to flow and timing, whilst specific percentage of follow-up was not calculated for most, it was easy to follow participants to study completion. Therefore, it is unlikely that there were enough losses to follow up to have introduced significant bias.

Findings

Eight studies met the eligibility criteria and we have included them in the review (Ferenci 2005; Mak 2008; Merle 2009; Lu 2010; Nicastro 2010; Sezer 2014; Yang 2015; Xu 2018). Details of these studies can be found in the tables (Characteristics of included studies; Table 2).

Below we narratively summarise review findings including details of the study populations, the analyte measured, the assay used and the diagnostic test accuracy results for all reported thresholds by study.

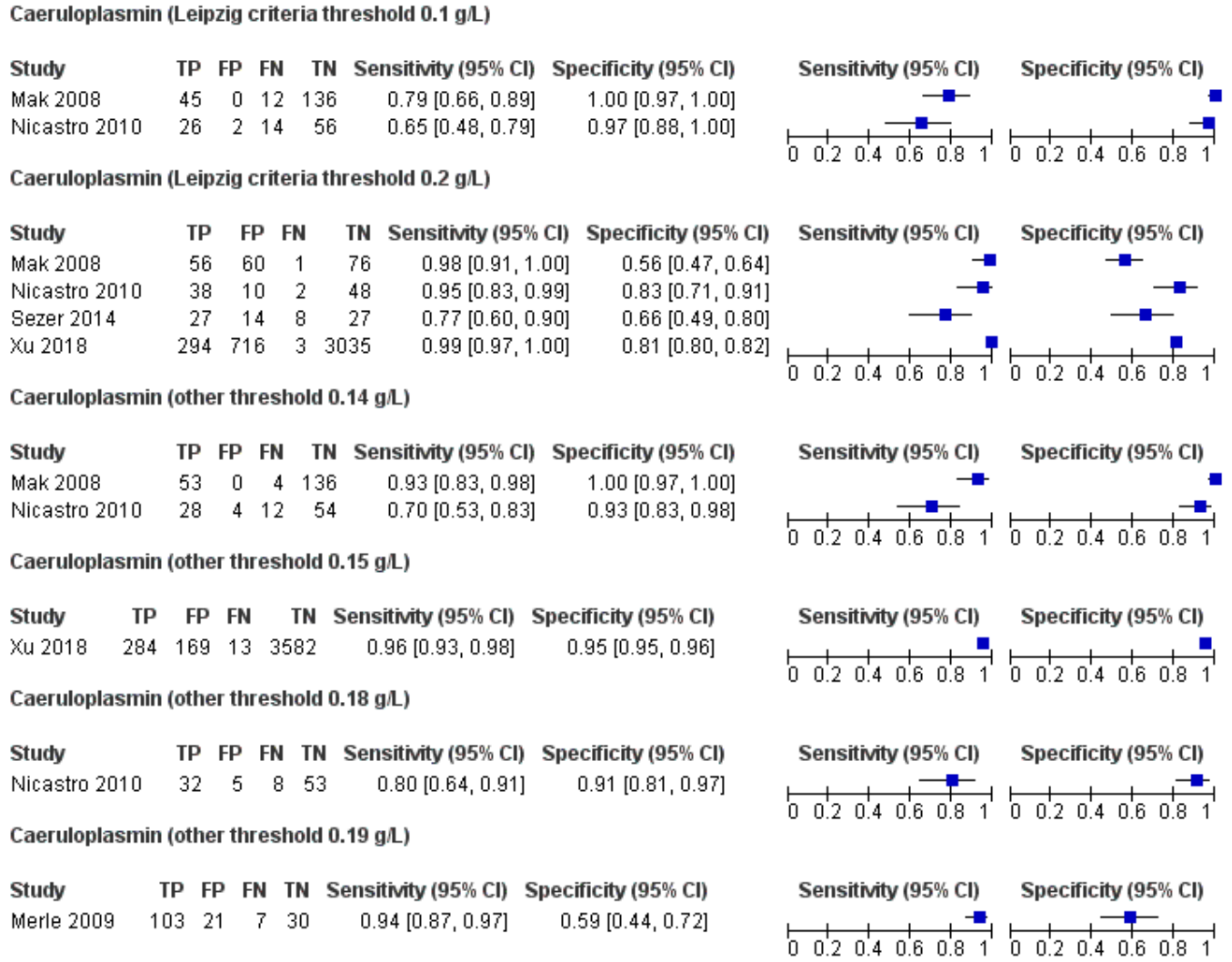
Index test - caeruloplasmin

Five studies evaluated various thresholds of ceruloplasmin (4281 participants, of whom 541 had WD) (Mak 2008; Merle 2009; Nicastro 2010; Sezer 2014; Xu 2018). Three studies were based in Europe (one in Germany (Merle 2009), one in Italy (Nicastro 2010), one in Turkey (Sezer 2014)) and two in China (Mak 2008; Xu 2018). Two studies were conducted in children (Nicastro 2010; Sezer 2014), one in an adult population (Merle 2009) and two in children and adults (Mak 2008; Xu 2018).

Cut-offs defined by the Leipzig criteria (0.1 g/L and 0.2 g/L) (Table 1), were reported by four of the five studies (Mak 2008; Nicastro 2010; Sezer 2014; Xu 2018) and four of the studies reported other thresholds ranging from 0.14 g/L to 0.19 g/L, mostly determined as the 'optimal' or 'most useful' threshold by ROC curve analysis (Mak 2008; Merle 2009; Nicastro 2010; Xu 2018). Sensitivity and specificity results for all thresholds are presented in a figure (Figure 3); 2 x 2 tables, sensitivity, specificity, PPV, NPV, LR+ and LR- for all thresholds are presented in an additional table (Table 3) and the results for the

cut-offs defined by the Leipzig criteria are presented in a summary of findings table (Summary of findings 1).

Figure 3. Forest plot of tests: 1 Caeruloplasmin (Leipzig criteria threshold 0.1g/L), 2 Caeruloplasmin (Leipzig criteria threshold 0.2g/L), 3 Caeruloplasmin (other threshold 0.14g/L), 4 Caeruloplasmin (other threshold 0.15g/L), 5 Caeruloplasmin (other threshold 0.18g/L), 6 Caeruloplasmin (other threshold 0.19g/L).



One study by Mak evaluated the cut-offs for ceruloplasmin used by the Leipzig criteria in a mixed clinical (hepatic and neurological) population of children and adults (Table 1) (Mak 2008). Serum caeruloplasmin was measured using nephelometry, with testing in 59 people with WD pre-treatment, 71 family members (49 heterozygotes, 22 wild type homozygotes), a validation group of 25 with suspected WD and 690 normal controls. The age range in people with WD was 4 years to 50 years. For the cut-offs used in the Leipzig criteria 0.2 g/L and 0.1 g/L, sensitivity was 98.3% (95% CI 90.6% to 100%) and 78.9% (95% CI 66.1% to 88.6%), respectively (Mak 2008). Specificity was 55.9% (95% CI 47.1% to 64.4%) and 100% (95% CI 97.3% to 100%), respectively (Table 3). ROC curve analysis conducted in this study showed that a cut-off of 0.14 g/L gave maximal sensitivity and specificity with an area under the ROC curve of 0.99 (95% CI 0.97 to 1.01). We constructed a 2 x 2 table based on this cut-off and calculated diagnostic test accuracy statistics (Table 3).

In the Merle study, caeruloplasmin measured by caeruloplasmin oxidase, was compared with that as measured by nephelometry in 110 people with WD with mixed clinical features (71 hepatic, 29 neurological, 10 pre-symptomatic), 52 healthy controls and 51 with cirrhosis not due to WD (Merle 2009). Median age of people with WD was 37 years (IQR 27 to 46.5), 45 were male (40.9%) and 65 were female (59.1%). Diagnosis was based on the Leipzig criteria, only adults were evaluated and within the WD group 14 were treatment-naive. As ceruloplasmin oxidase is not part of the Leipzig criteria (Table 1) this will not be considered further and hence the findings recorded from this study relate to caeruloplasmin measured

by nephelometry. The study authors did not quote diagnostic accuracy figures for caeruloplasmin as quoted in the Leipzig criteria but did undertake ROC curve analysis to generate a a cut-off with maximal sensitivity and specificity. ROC curve analysis conducted by the study authors for the 110 people with WD, 52 healthy controls and 51 with cirrhosis not due to WD, showed that a cut-off of 0.19 g/L gave maximal sensitivity and specificity with an area under the ROC curve of 0.93 (95% CI 0.897 to 0.962) (Merle 2009). We constructed a 2 x 2 table based on this cut-off and calculated diagnostic test accuracy statistics comparing the WD group with those with cirrhosis not due to WD (Table 3).

In the Nicastro study, 40 children with treatment-naive WD (elevated transaminases or family screening) and 58 age- and gender-matched controls with non-WD-related chronic liver disease were evaluated (Nicastro 2010). In the WD group there were 26 males, 14 females, age range 1.1 to 20.9 with a median of 6.1 years. Caeruloplasmin was measured by radial-immunodiffusion, WD was defined by the Leipzig criteria (Table 1). ROC curve analysis conducted by the authors showed that a cut-off of 0.2 g/L gave maximal sensitivity and specificity with an area under the ROC curve of 0.94 (95% CI 0.88 to 0.99) (Table 3). This cut-off is also one of the caeruloplasmin cut-offs used in the Leipzig criteria (Table 1). We constructed a 2 x 2 table based on this cut-off and calculated diagnostic test accuracy statistics: sensitivity 95.0% (95% CI 83.1% to 99.4%); and specificity 82.8% (95% CI 70.6% to 82.2%) (Table 3).

The Sezer study was conducted in a paediatric population with 35 children with WD (treatment-naive) as defined by the Leipzig criteria (Table 1) and 41 age, gender matched with non-WD chronic liver disease (Sezer 2014). In the WD group mean age was 10.2 years, with a range of 8 years to 16.5 years and 57.1% (20) were male and 42.9% (15) were female. Serum caeruloplasmin was measured using immuno-turbidimetry. At the 0.2 g/L cut off used in the Leipzig criteria (Table 1), sensitivity was 77.1% (95% CI 59.9% to 89.6%) and specificity 65.9% (95% CI 49.4% to 79.9%) (Table 3). The authors concluded that they did not undertake ROC curve analysis to define a clear cut-off due to overlap in caeruloplasmin values between WD and non-WD group, with 22% in the former group > 0.2 g/L and 29% in the latter group < 0.2 g/L (Sezer 2014).

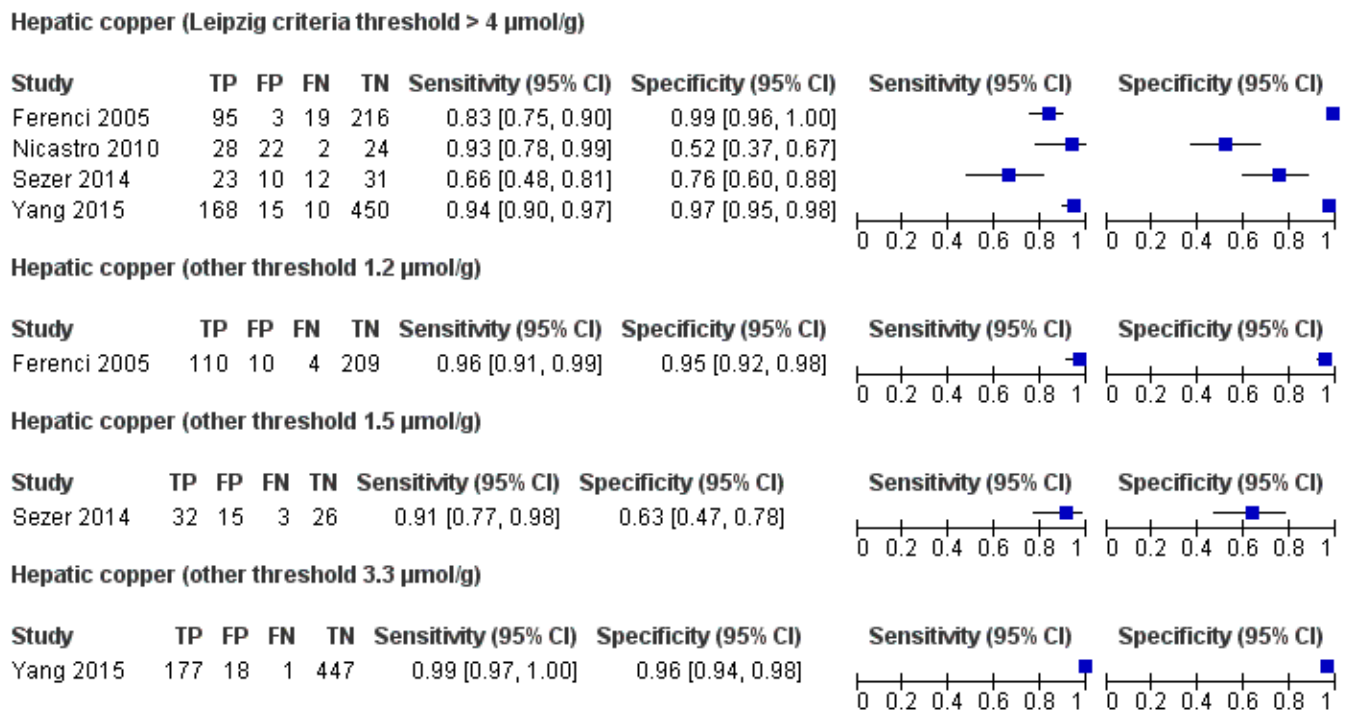
In the Xu study both children and adults with ceruloplasmin being measured by nephelometry in 297 with WD (hepatic, neurological, pre-symptomatic), 3751 with non-WD (chronic liver disease, nephrotic syndrome, movement disorders) (Xu 2018). The mean age of the participants at diagnosis was 21.8 years with a range of 2 years to 62 years. At the 0.2 g/L cut-off used in the Leipzig criteria (Table 1) (also presented by the study authors), sensitivity was 99% (95% CI 97.1% to 99.8%) and specificity was 80.9% (95% CI 79.6% to 82.2%) (Table 3). ROC curve analysis conducted by the authors showed that a cut-off of 0.15 g/L gave maximal sensitivity and specificity with an area under the ROC curve of 0.992 (95% CI 0.987 to 0.996) (Xu 2018). We constructed a 2 x 2 table based on this cut-off and calculated diagnostic test accuracy statistics (Table 3).

Index test - hepatic copper

Four studies evaluated various thresholds of hepatic copper (1150 participants, of whom 367 had WD) (Ferenci 2005; Nicastro 2010; Sezer 2014; Yang 2015). Three studies were conducted in Europe (one in Austria (Ferenci 2005), one in Italy (Nicastro 2010) and one in Turkey (Sezer 2014)) and one study was conducted in China (Yang 2015). Two were conducted in children (Nicastro 2010, Sezer 2014) and two in children and adults (Ferenci 2005; Yang 2015).

The cut-off defined by the Leipzig criteria (> 4 µmol/g, see Table 1) was reported by all four studies (Ferenci 2005; Nicastro 2010; Sezer 2014; Yang 2015) and three of the four studies reported other thresholds ranging from 1.2 µmol/g to 3.3 µmol/g, mostly determined as the 'optimal' or 'most useful' threshold by ROC curve analysis (Ferenci 2005; Nicastro 2010; Sezer 2014; Yang 2015). Sensitivity and specificity results for all thresholds are presented in a figure (Figure 4); 2 x 2 tables, sensitivity, specificity, PPV, NPV, LR+ and LR- for all thresholds are presented in an additional table (Table 4) and the results for the cut-offs defined by the Leipzig criteria are presented in a summary of findings table (Summary of findings 2).

Figure 4. Forest plot of tests: 7 Hepatic copper (Leipzig criteria threshold >4 µmol/g), 8 Hepatic copper (other threshold 1.2 µmol/g), 9 Hepatic copper (other threshold 1.5 µmol/g), 10 Hepatic copper (other threshold 3.3 µmol/g).



In the Ferenci study, hepatic biopsy pre-treatment was undertaken in a mixed clinical population (83 hepatic, 34 neurological, 18 pre-symptomatic) of children and adults (Ferenci 2005). There were 114 people with WD confirmed by the Leipzig criteria in whom copper content was measured (Table 1); these people were compared with 26 normal controls and 219 people with chronic liver disease (including hepatitis C, non-alcoholic liver disease, alcoholic liver disease, autoimmune hepatitis, Budd-Chiari syndrome, haemochromatosis). Hepatic copper content was measured using flame atomic absorption spectroscopy (Kingston 1986) and did not differ on the basis of age or mode of clinical presentation. At the pre-specified cut-off for hepatic copper of 250 µg/g (4 µmol/g), which is the cut off used in the Leipzig criteria (Table 1), sensitivity was 83.3% (95% CI 75.2% to 89.7%), specificity was 98.6% (95%CI 96.1% to 99.0%) (Table 4). ROC curve analysis conducted by the study authors showed increased sensitivity for detecting WD by reducing cut-off to 1.2 µmol/g, with an area under the ROC curve of 0.98 (95% CI 0.979 to 0.996) (Ferenci 2005). We constructed a 2 x 2 table based on this cut-off and calculated diagnostic test accuracy statistics (Table 4).

As stated above, in the Nicastro study, 40 children with treatment-naive WD (elevated transaminases or family screening) and 58 age- and gender-matched controls with non-WD-related chronic liver disease were evaluated (Nicastro 2010). In the WD group there were 26 males and 14 females; age range was 1.1 years to 20.9 years, with a median of 6.1 years. Hepatic copper was measured by atomic absorption spectroscopy. The authors did not undertake ROC analysis for hepatic copper, however, based on figures provided in the paper at the cut-off of > 4 µmol/g (as used in the Leipzig criteria) (Table 1), we constructed a 2 x 2 table and calculated diagnostic test accuracy statistics: sensitivity 93.3% (95% CI 77.9% to 99.2%); and specificity 52.2% (95% CI 37.0% to 67.1%) (Table 4).

The Sezer study was conducted in 35 children with WD (treatment-naive) as defined by the Leipzig criteria (Table 1) and 41 age- and gender-matched controls with non-WD-related chronic liver disease (Sezer 2014). In the WD group, mean age was 10.2 years with a range of 8 years to 16.5 years and 57.1% (20) male, 42.9% (15) female. Hepatic copper was measured by atomic absorption spectroscopy. At the cut-off of 4 µmol/g for hepatic copper used in the Leipzig criteria (Table 1), sensitivity was 65.7% (95% CI 47.8% to 80.9%) and specificity was 75.6% (95% CI 59.7% to 87.6%). ROC curve analysis conducted by the authors showed that a cut-off of 1.5 µmol/g gave maximal sensitivity and specificity with an area under the ROC curve of 0.838 (95% CI 0.749 to 0.927) (Sezer 2014). We constructed a 2 x 2 table based on this cut-off and calculated diagnostic test accuracy statistics (Table 4).

In the Yang study, 691 children and adults with chronic liver disease including 178 with WD (treatment-naive) as assessed by Leipzig criteria underwent liver biopsy pre-treatment with copper measurement by atomic absorption spectroscopy (Yang 2015). In the WD group 104 were male, 74 female and mean age was 19.7 years. The authors gave diagnostic accuracy figures at cut-offs in a group in which cholestatic diseases such as primary biliary cirrhosis, primary sclerosing cholangitis were excluded with total of 465 in the chronic liver disease group. The traditional Leipzig criteria cut-off

of 4 µmol/g gave a sensitivity of 94.4% (95% CI 89.9% to 97.3%) and a specificity of 96.8% (95% CI 94.75% to 98.2%) (Table 4). ROC curve analysis undertaken by the authors generated a cut-off of 3.3 µmol/g with an area under the ROC curve of 0.987 (Yang 2015). We constructed a 2 x 2 table based on this cut-off and calculated diagnostic test accuracy statistics (Table 4).

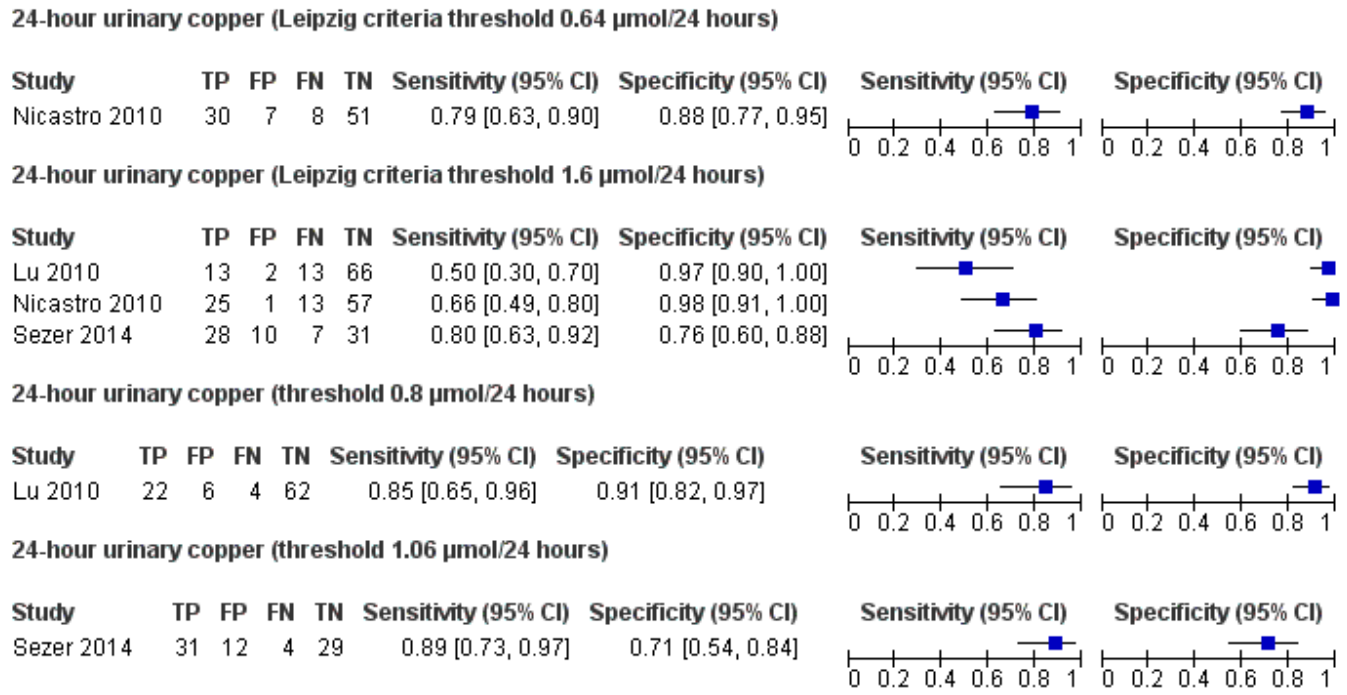
Index test - 24-hour urinary copper

Three studies evaluated various thresholds of 24-hour urinary copper (268 participants, of whom 101 had WD) (Lu 2010; Nicastro 2010; Sezer 2014). Two studies were conducted in Europe (one in Italy (Nicastro 2010) and one in Turkey (Sezer 2014)) and one study was conducted in China (Lu 2010). All three studies were conducted in children (Lu 2010; Nicastro 2010; Sezer 2014).

In the Leipzig criteria a score for 24-hour urinary copper is assigned based on number times ULN (Table 1) with greater than 1.6 µmol/24 hours and 0.64 µmol/24 hours indicative of WD in adults and children respectively. Although all three studies are undertaken in children, only the Nicastro study gives diagnostic accuracy figures approximating the paediatric cut-off used in the Leipzig criteria (Nicastro 2010), with the remaining two studies giving diagnostic accuracy figures for the adult cut-off used in the Leipzig criteria (Table 1) (Lu 2010; Sezer 2014).

The cut-offs defined by the Leipzig criteria (0.64 µmol/24 hours and 1.6 µmol/24 hours) was reported by all three studies (Lu 2010; Nicastro 2010; Sezer 2014) and two of the three studies reported other thresholds of 0.8 µmol/24 hours and 1.06 µmol/24 hours, determined as the 'optimal' or 'best fitting' threshold by ROC curve analysis (Lu 2010; Sezer 2014). Sensitivity and specificity results for all thresholds are presented in a figure (Figure 5); 2 x 2 tables, sensitivity, specificity, PPV, NPV, LR+ and LR- for all thresholds are presented in an additional table (Table 5) and the results for the cut-offs defined by the Leipzig criteria are presented in a summary of findings table (Summary of findings 3).

Figure 5. Forest plot of tests: 11 24-hour urinary copper (Leipzig criteria threshold 0.64 µmol/24 hours), 12 24-hour urinary copper (Leipzig criteria threshold 1.6 µmol/24 hours), 13 24-hour urinary copper (threshold 0.8 µmol/24 hours), 14 24-hour urinary copper (threshold 1.06 µmol/24 hours).



The Lu study was conducted in 26 children with WD and 68 with non-WD-related chronic liver disease (Lu 2010). 24-hour urinary copper was the analyte of interest and was measured using ICP mass spectrometry. The traditional Leipzig criteria cut-off of 1.6 µmol/24 hours gave a sensitivity of 50% (95% CI 29.9% to 70.1%) and a specificity of 97.1% (95% CI 89.8% to 99.6%) (Table 5). The study determined that the optimum cut-off for 24-hour urinary copper was 0.8 µmol/24 hours, with an area under the curve of 0.909 (95% CI 0.839 to 0.979) (Lu 2010). We constructed a 2 x 2 table based on this cut-off and calculated diagnostic test accuracy statistics (Table 5).

As stated above, in the Nicastro study, 40 children with treatment-naive WD (elevated transaminases or family screening) and 58 age- and gender-matched controls with non-WD-related chronic liver disease were evaluated (Nicastro 2010). In the WD group there were 26 males and 14 females, age range 1.1 years to 20.9 years, with a median of 6.1 years. Urine copper was measured by atomic absorption spectroscopy. ROC curve analysis conducted

by the study authors showed that a cut-off of 0.64 µmol/24 hours gave maximal sensitivity and specificity with an area under the ROC curve of 0.91 (95% CI 0.85 to 0.97) (Nicastro 2010). We constructed a 2 x 2 table based on this cut-off and calculated diagnostic test accuracy statistics: sensitivity 78.9% (95% CI 62.7% to 90.5%) and specificity 87.9% (95% CI 76.7% to 95.0%) (Table 5). Nicastro also presented the other traditional Leipzig criteria cut-off of 1.6 µmol/24 hours gave a sensitivity of 65.8% (95% CI 48.7% to 80.4%) and a specificity of 96.3% (95% CI 90.8% to 100%) (Table 5).

As stated above, the Sezer study was conducted in 35 children with WD (treatment-naive) as defined by the Leipzig criteria (Table 1) and 41 age- and gender-matched controls with non-WD-related chronic liver disease (Sezer 2014). In the WD group mean age was 10.2 years with a range of 8 years to 16.5 years and 57.1% (20) male, 42.9% (15) female. 24-hour urine copper was measured by atomic absorption spectroscopy. The traditional Leipzig criteria cut-off of 1.6 µmol/24 hours gave a sensitivity of 80.0% (95% CI 63.1% to 91.6%) and a specificity of 75.6% (95% CI 59.7% to 87.6%) (Table 5). ROC curve analysis was conducted by the authors and showed that a cut-off of 1.06 µmol/24 hours gave maximal sensitivity and specificity with an area under the ROC curve of 0.843 (95% CI 0.752 to 0.934) (Sezer 2014). We constructed a 2 x 2 table based on this cut-off and calculated diagnostic test accuracy statistics (Table 5).

DISCUSSION

The diagnosis of WD is a heterogenous process involving a combination of clinical, biochemical, immunological and genetic test results (Table 1). The limited evidence provided by the studies in this review suggests that at the optimum cut-offs quoted above based on ROC curve analysis, vary in terms of the method used for each analyte and the age of participant groups. The EASL guidelines on WD provided a narrative, expert-based review on the evidence supporting the original Leipzig criteria (EASL 2012).

Three main studies are provided to support caeruloplasmin cut-offs used (Merle 2009; Perlman 1979; Steindl 1997). In the Perlman study, laboratory measures of copper metabolism were assessed for 25 people with WD (diagnosis based on KF ring presence in absence of cholestasis) and 20 people with chronic active hepatitis (Perlman 1979). Caeruloplasmin was measured by both oxidase and an immunological method, however, only one reference range is quoted without a reference (0.22 g/L to 0.49 g/L) and although the authors describe the participants as children, ages ranged from 4 years to 20 years. The cut-off evaluated was < 22 mg/dL which the authors concluded could not accurately differentiate WD from chronic active hepatitis as 28% with WD were above this and 25% with chronic active hepatitis were below this cut-off. In the Steindl study, caeruloplasmin was measured by radial immunodiffusion in a mixed hepatic and neurological group of 55 children and adults with WD (Steindl 1997). Diagnosis was based on Sternlieb's criteria and those with KF rings, neurological features had a significantly lower caeruloplasmin compared with those without any of these features (Sternlieb 1990). No controls were used and although the method was referenced and a range quoted (0.20 to 0.60 g/L), how this was derived was unclear. Heterozygotes have been shown to have low caeruloplasmin (< 0.2 g/L) in 20% of cases but having a normal allele will have none of the clinical phenotype (Gromadzka 2010). Caeruloplasmin is an acute-phase protein and so may be elevated in inflammation as well as in those with increased endogenous oestrogen (pregnancy) or exogenous oestrogen (oral contraceptive pill).

For discussion of cut-off values for 24-hour urinary copper that support the Leipzig criteria, the EASL guidelines (EASL 2012) makes reference to a number of WD case-series that do not have control data and have limited detail regarding laboratory method use to measure the analyte or how the reference range used was derived (Ferenci 2007; Giacchino 1997; Sanchez-Albisua 1999; Steindl 1997). The EASL guidelines, however, do highlight some important points, that the cut-off for adults is higher than in children and that levels may be lower in asymptomatic siblings and that false positives for 24-hour urinary copper can occur in the presence of non-WD cholestatic liver disease and nephrotic syndrome (EASL 2012).

Previous commentary has suggested that measurement of hepatic copper content is the gold standard for diagnosing WD (EASL 2012); however, even the authors of one of the largest studies to date on this subject highlights studies of where hepatic copper may give false negative results, due to sampling error or possible differences in hepatic copper distribution (Ferenci 2005).

Summary of main results

We have tabulated a summary of the main results from this review that quote diagnostic accuracy figures for relevant Leipzig criteria (Table 1) for each index test in summary of findings tables and forest plots (Summary of findings 1, Summary of findings 2, Summary of findings 3, Figure 3, Figure 4, Figure 5).

In the five studies eligible for this review that provided a primary outcome for reporting for caeruloplasmin, the optimum cut-off varies between 0.14 g/L to 0.2 g/L (Mak 2008; Merle 2009; Nicastro 2010; Sezer 2014; Xu 2018). In general, these studies were well-designed, providing a clear WD disease definition, details of the laboratory method used, use of appropriate controls and evaluation of cut-offs using ROC curve analysis. The differences in the suggested optimum cut-offs may reflect the different methods used for caeruloplasmin evaluation, clinical presentation, possible age of the participants involved and their ethnicity. Some of these studies evaluated caeruloplasmin in those with mild hepatic WD (Nicastro 2010, Sezer 2014), whilst others were undertaken in those with hepatic of variable severity and neurological WD (Mak 2008; Merle 2009; Xu 2018). Caeruloplasmin levels are lower in neonates rising to adult range in two to three years, however, it rises during pregnancy, in women on the oral contraceptive pill and in those experiencing acute inflammation. Two of the studies were undertaken in children (Nicastro 2010, Sezer 2014), two in a mixed children and adult population (Mak 2008, Xu 2018) and one in an adult-only population (Merle 2009) (Table 2).

As it is synthesised in the liver, caeruloplasmin is lower in other causes of chronic liver disease apart from WD, but it is also lowering in protein-losing states such as enteropathies and nephrotic syndrome (Cox 1966; EASL 2012). Previous work has suggested that those on chelation therapy have lower caeruloplasmin levels (Grazyna 2014); however, this does not appear to be the case in the Merle study, where 65% where on penicillamine therapy but differences in assay methodology may also have contributed to this effect (Merle 2009). Different methods for a particular analyte, take caeruloplasmin in this instance, may vary in terms of reference range, bias and precision. Therefore, where cut-off values are used the importance of considering method is paramount as the same

sample run for the same analyte with a different method is likely to generate a different result for the reasons given above.

For 24-hour urinary copper there is limited evidence presented for the adult cut-off of > 1.6 µmol/24 hours (EASL 2012); however, our review included three studies evaluating 24-hour urinary copper cut-offs in children with WD-related chronic liver disease (Lu 2010; Nicastro 2010; Sezer 2014) (Table 2). The Lu study used inductively coupled mass spectrometry evaluating a Chinese population and Sezer, Nicastro used atomic absorption spectroscopy in a Turkish and Italian population, respectively (Lu 2010; Nicastro 2010; Sezer 2014). Both studies were well-designed, with clear criteria for WD diagnosis (Table 1), appropriate age, gender matched controls. The difference in terms of the optimum cut-offs may be due to different methods and ethnicity (Table 2).

For cut-offs for hepatic copper, four studies were eligible for this review, with different age groups, ethnicity and method used for index test analysis (Table 2). The Ferenci study was well designed, with clearly defined WD criteria carried out on a mixed children and adult population. Appropriate laboratory methods evaluating hepatic copper using atomic absorption spectroscopy were utilised (Ferenci 2005). The optimum cut-off based on ROC curve analysis was 1.2 µmol/g giving a sensitivity of 96.4% with a specificity of 95.4% (Ferenci 2005). This contrasts with the higher cut-off of 4 µmol/g used in the Leipzig criteria, however, as Ferenci discusses, this original cut-off was based on a sample of only seven people with WD (Ferenci 2005). Interestingly, the Chinese-based Yang study, also in children and adults, offered a robust defence of the 4 µmol/g cut-off for hepatic copper and showed that in their population this resulted in a sensitivity of 99.4% and a specificity of 96.1% (Yang 2015). This has a larger participant sample than the former European-based study, had a clearly defined WD disease criteria, clear methodology in an adult population. Again, such differences may be due to the differences in the method but may reflect ethnicity and different distribution in disease causing alleles in each population.

Both Nicastro and Sezer evaluated hepatic copper using two different methods of flame atomic absorption spectroscopy in children with hepatic disease in Italy and Turkey, respectively (Nicastro 2010; Sezer 2014). Nicastro did not undertake ROC analysis however at the cut-off of 4µmol/g used in the Leipzig criteria (Table 1) had a sensitivity of 87.5% and specificity of 96.6% (Nicastro 2010). Sezer found that the cut-off of 4µmol/g had a sensitivity of 65%, specificity of 77% and that decreasing the threshold to 1.5µmol/g increased sensitivity to 91.4% but with specificity of 65.8% (Sezer 2014). The differences between the two studies may be due to differences in: disease severity, genotypes between the two populations, and method of measuring hepatic copper.

Strengths and weaknesses of the review

The strength of this review is the adoption of high methodological standards, in particular, screening of 10,638 studies and detailed scrutiny of 20 studies allowed us to identify important implications for future research about the diagnostic accuracy of biochemical testing for WD in both children and adults. These implications relate to both the methodological conduct of future diagnostic studies but also the optimum method for monitoring chelation therapy for WD.

Weaknesses of the review include the small number of included eligible studies, data from which were unsuitable for pooling in a meta-analysis, due to differences in methods used for index tests. Key differences between studies include differences in assay methodology, different cut-off values for diagnostic thresholds, different age and ethnicity groups. Key biases of note were in participant selection and reference standard that may limit external validity of the study findings.

Applicability of findings to the review question

The findings of this review are applicable to the review question, although all included studies used the Leipzig criteria for WD disease definition, some of the included studies had limited detail on other methodologies used to calculate the score apart from the relevant index test being evaluated. A high proportion of the studies identified by our search were excluded from the review because of lack of disease definition and poorly defined index test methodology or reference range source. These methodological weakness could have resulted in bias or over estimation of the accuracy of the index test.

AUTHORS' CONCLUSIONS

Implications for practice

Any disease cut-off for biochemical testing for Wilson's disease (WD) is method dependent and the specificity of the results should be interpreted with caution in the presence of other causes of chronic liver disease. Clinicians should be aware that there is no gold standard test for diagnosing WD and that a combination of clinical and laboratory testing may be required for diagnosis.

Implications for research

Well-designed studies are needed to evaluate whether cut-offs for biochemical testing for WD are affected by age, clinical presentation and ethnicity. The invasiveness of liver biopsy has led to a decline in its routine use; however, use of non-invasive serum fibrosis markers may offer an opportunity to evaluate target organ damage particularly given the protean manifestation of hepatic disease in WD. Further diagnostic marker testing with standardised methodologies, with appropriately validated cut-offs and disease severity score, is only likely to succeed in a large multinational WD registry setting if this is to have impact on clinical practice. Such a facility would also allow for evaluation of biochemical tests to monitor chelation therapy.

ACKNOWLEDGEMENTS

This project was supported by the National Institute for Health Research, via Cochrane Infrastructure funding to the Cochrane Cystic Fibrosis and Genetic Disorders Group. The views and opinions expressed therein are those of the authors and do not necessarily reflect those of the Systematic Reviews Programme, NIHR, NHS or the Department of Health.

Translation assistance - Professor Liliya Ziganshina (Head of Basic and Clinical Pharmacology, Kazan Federal University), Liang Ning (PhD Candidate, Centre for Evidence Based Medicine, Beijing University of Chinese Medicine).

REFERENCES

References to studies included in this review

Ferenci 2005 *{published data only}*

Ferenci P, Steindl-Munda P, Vogel W, Jessner W, Gschwantler M, Stauber R, et al. Diagnostic value of quantitative hepatic copper determination in patients with Wilson's Disease. *Clinical Gastroenterology & Hepatology* 2005;**3**(8):811-8.

Lu 2010 *{published data only}*

Lu Y, Liu XQ, Wang XH, Wang JS. The reassessment of the diagnostic value of 24-hour urinary copper excretion in children with Wilson's disease. *Chinese Journal of Hepatology* 2010;**18**(1):49-53.

Mak 2008 *{published data only}*

Mak CM, Lam CW, Tam S. Diagnostic accuracy of serum ceruloplasmin in Wilson disease: determination of sensitivity and specificity by ROC curve analysis among ATP7B-genotyped subjects. *Clinical Chemistry* 2008;**54**(8):1356-62.

Merle 2009 *{published data only}*

Merle U, Eisenbach C, Weiss KH. Serum ceruloplasmin oxidase activity is a sensitive and highly specific diagnostic marker for Wilson's disease. *Journal of Hepatology* 2009;**51**(5):925-30.

Nicastro 2010 *{published data only}*

Nicastro E, Ranucci G, Vajro P, Vegnente A, Iorio R. Re-evaluation of the diagnostic criteria for Wilson disease in children with mild liver disease. *Hepatology* 2010;**52**(6):1948-56.

Sezer 2014 *{published data only}*

Sezer OB, Perk P, Hoşnut FÖ, Köse SK, Özcay F. Is it necessary to re-evaluate diagnostic criteria for Wilson disease in children?. *Turkish Journal of Gastroenterology* 2014;**25**(6):690-5.

Xu 2018 *{published data only}*

Xu R, Jiang YF, Zhang YH, Yang X. The optimal threshold of serum ceruloplasmin in the diagnosis of Wilson's disease: A large hospital-based study. *PLoS One* 2018;**13**(1):e0190887.

Yang 2015 *{published data only}*

Yang X, Tang XP, Zhang YH, Luo KZ, Jiang YF. Prospective evaluation of the diagnostic accuracy of hepatic copper content, as determined using the entire core of a liver biopsy sample. *Hepatology* 2015;**62**(6):1731-41.

References to studies excluded from this review

Causa 1997 *{published data only}*

Cauza E, Maier-Dobersberger T, Polli C, Kaserer K, Kramer L, Ferenci P. Screening for Wilson's disease in patients with liver diseases by serum ceruloplasmin. *Journal of Hepatology* 1997;**27**(2):358-62.

Frommer 1981 *{published data only}*

Frommer DJ. Urinary copper excretion and hepatic copper concentrations in liver disease. *Digestion* 1981;**21**(4):169-78.

Gibbs 1979 *{published data only}*

Gibbs K, Walshe JM. A study of the caeruloplasmin concentrations found in 75 patients with Wilson's disease, their kinships and various control groups. *Quarterly Journal of Medicine* 1979;**48**(191):477-63.

Gnanou 2006 *{published data only}*

Gnanou JV, Thykadavil VG, Thuppil V. Pros and cons of immunochemical and enzymatic method in the diagnosis of Wilson's disease. *Indian Journal of Medical Sciences* 2006;**60**(9):371-5.

Lech 2007 *{published data only}*

Lech T, Sadlik JK. Contribution to the data on copper concentration in blood and urine in patients with Wilson's disease and in normal subjects. *Biological Trace Element Research* 2007;**118**(1):16-20.

Li 1983 *{published data only}*

Li NZ. The value of biochemical determination of copper in clinical diagnosis of hepatolenticular degeneration. *Zhonghua Yi Xue Za Zhi* 1983;**63**(6):342-5.

Liggi 2013 *{published data only}*

Liggi M, Mais C, Demurtas M, Sorbello O, Demelia E, Civolani A, et al. Uneven distribution of hepatic copper concentration and diagnostic value of double-sample biopsy in Wilson's disease. *Scandinavaian Journal of Gasteroenterology* 2013;**48**(12):1452-8.

Mahjoub 2012 *{published data only}*

Mahjoub F, Fereiduni R, Jahanzad I, Farahmand F, Monajemzadeh M, Najafi M. Atomic absorption spectrometry in Wilson's disease and its comparison with other laboratory tests and paraclinical findings. *Iranian Journal of Paediatrics* 2012;**22**(1):52-6.

Markowitz 1955 *{published data only}*

Markowitz H, Gubler CJ, Mahoney JP, Cartwright GE, Wintrobe MM. Studies on copper metabolism. XIV. Copper, ceruloplasmin and oxidase activity in sera of normal human subjects, pregnant women, and patients with infection, hepatolenticular degeneration and the nephrotic syndrome. *Journal of Clinical Investigation* 1955;**34**(10):1498-508.

Mzhel'skaia 1994 *{published data only}*

Mzhel'skaia TI, Ivanova-Smolenskaia IA, Korobova NV. Factors affecting the concentration of serum ceruloplasmin in Wilson-Konovalov disease. *Voprosy Meditsinskoi Khimii* 1994;**40**(6):48-53.

Prasad 1998 *{published data only}*

Prasad R, Kaur G, Walia BN. A critical evaluation of copper metabolism in Indian Wilson's disease children with special reference to their phenotypes and relatives. *Biological Trace Element Research* 1998;**65**(2):153-65.

Siotto 2014 *{published data only}*

Siotto M, Pasqualetti P, Marano M, Squitti R. Automation of o-dianisidine assay for ceruloplasmin activity analyses:

usefulness of investigation in Wilson's disease and in hepatic encephalopathy. *Journal of Neural Transmission (Vienna)* 2014;**121**(10):1281-6.

References to studies awaiting assessment

Aksu 2018 *{published data only}*

Aksu AÜ, Sarı S, Gürkan ÖE, Dalgıç B. Urinary 24-hour copper excretion at the time of diagnosis in children with Wilson's disease. *Acta Gastro-Enterologica Belgica* 2018;**81**(3):410-414.

Zhou 2019 *{published data only}*

Zhou XX, He RX, Pu XY, Li XH, Qin HL, Huang HW, Liang XL. Clinical characteristics of the Wilson disease carrier [肝豆状核变性携带者的临床特点及治疗策略] . *Zhonghua Yi Xue Za Zhi* 2019;**99**(11):806-11.

Additional references

Berman 1999

Berman DH, Leventhal RI, Gavaler JS, Cadoff EM, Van Thiel DH. Clinical differentiation of fulminant wilsonian hepatitis from other causes of hepatic failure. *Gastroenterology* 1991;**100**(4):1129-34.

Bull 1993

Bull PC, Thomas GR, Rommens JM, Forbes JR, Cox DW. The Wilson disease gene is a putative copper transporting P-type ATPase similar to the Menkes gene. *Nature Genetics* 1993;**5**(4):327-37.

Coffey 2013

Coffey AJ, Durkie M, Hague S, McLay K, Emmerson J, Lo C, et al. A genetic study of Wilson's disease in the United Kingdom. *Brain* 2013;**136**(Pt 5):1476-87.

Compston 2009

Compston A. Progressive lenticular degeneration: a familial nervous disease associated with cirrhosis of the liver, by S. A. Kinnier Wilson (From the National Hospital, and the Laboratory of the National Hospital, Queen Square, London) Brain 1912: 34; 295-509. *Brain* 2009;**132**(Pt 8):295-509.

Cox 1966

Cox DW. Factors influencing serum ceruloplasmin levels in normal individuals. *Journal of Laboratory and Clinical Medicine* 1966;**68**(6):893-904.

Czlonkowska 2009

Czlonkowska A, Gromadzka G, Chabik G. Monozygotic female twins discordant for phenotype of Wilson's disease. *Movement Disorders* 2009;**24**(7):1066-9.

Członkowska 2018

Członkowska A, Litwin T, Dusek P, Ferenci P, Lutsenko S, Medici V, et al. Wilson Disease. *Nature Reviews. Disease Primers* 2018;**4**(1):21.

Davis 1996

Davis W, Chowrimootoo GF, Seymour CA. Defective biliary copper excretion in Wilson's disease: the role of caeruloplasmin. *European Journal of Clinical Investigation* 1996;**26**(10):893-901.

Duncan 2016

Duncan A, Yacoubian C, Beetham R, Catchpole A, Bullock D. The role of calculated non-caeruloplasmin-bound copper in Wilson's disease. *Annals of Clinical Biochemistry* 2016;**54**(6):649-54.

EASL 2012

European Association for Study of Liver. EASL Clinical Practice Guidelines: Wilson's disease. *Journal of Hepatology* 2012;**56**(3):671-85.

El Balkhi 2011

El Balkhi S, Trocello JM, Poupon J, Chappuis P, Massicot F, Girardot-Tinant N, et al. Relative exchangeable copper: a new highly sensitive and highly specific biomarker for Wilson's disease diagnosis. *Clinica Chimica Acta* 2011;**412**(23-4):2254-60.

Ferenci 2003

Ferenci P, Caca K, Loudianos G, Mieli-Vergani G, Tanner S, Sternlieb I, et al. Diagnosis and phenotypic classification of Wilson disease. *Liver International* 2003;**23**(3):139-42.

Ferenci 2007

Ferenci P, Członkowska A, Merle U, Ferenc S, Gromadzka G, Yurdaydin C. Late-onset Wilson's disease. *Gastroenterology* 2007;**132**(4):1294-8.

Ferenci 2019

Ferenci P, Stremmel W, Czlonkowska A, Szalay F, Viveiros A, Slattermayer AF, et al. Age and sex but not ATP7B genotype effectively influence the clinical phenotype of Wilson disease. *Hepatology* 2019;**69**(4):1464-76.

Giacchino 1997

Giacchino R, Marazzi MG, Barabino A, Fasce L, Ciravegna B, Famularo L. Syndromic variability of Wilson's disease in children. Clinical study of 44 cases. *Italian Journal of Gastroenterology and Hepatology* 1997;**29**(2):155-61.

Grazyna 2014

Gromadzka G, Karpińska A, Przybyłkowski A, Litwin T, Wierzchowska-Ciok A, Dzieżyc K. Treatment with D-penicillamine or zinc sulphate affects copper metabolism and improves but not normalizes antioxidant capacity parameters in Wilson disease. *Biometals* 2014;**27**(1):207-15.

Gromadzka 2010

Gromadzka G, Chabik G, Mendel T, Wierzchowska A, Rudnicka M, Czlonkowska A. Middle-aged heterozygous carriers of Wilson's disease do not present with significant phenotypic deviations related to copper metabolism. *Journal of Genetics* 2010;**89**(4):463-7.

Hellman 2002

Hellman NE, Gitlin JD. Ceruloplasmin metabolism and function. *Annual Review of Nutrition* 2002;**22**:439-58.

Huster 2012

Huster D, Kühne A, Bhattacharjee A, Raines L, Jantsch V, Noe J, et al. Diverse functional properties of Wilson disease ATP7B variants. *Gastroenterology* 2012;**142**(4):947-56.

Kelson 1978

Kelson JR, Shamberger RJ. Methods compared for determining zinc in serum by flame atomic absorption spectroscopy. *Clinical Chemistry* 1978;**24**(2):240-4.

Kenney 2007

Kenney SM, Cox DW. Sequence variation database for the Wilson disease copper transporter, ATP7B. *Human Mutation* 2007;**28**(12):1171-7.

Kingston 1986

Kingston HM, Jassie LB. Microwave energy for acid decomposition at elevated temperatures and pressures using biological and botanical samples. *Analytical Chemistry* 1986;**58**(12):2534-41.

Korman 2008

Korman JD, Volenberg I, Balko J, Webster J, Schiodt FV, Squires RH Jr, et al. Screening for Wilson disease in acute liver failure: a comparison of currently available diagnostic tests. *Hepatology* 2008;**48**(4):1167-74.

LaRusso 1976

LaRusso NF, Summerskill WH, McCall JT. Abnormalities of chemical tests for copper metabolism in chronic active liver disease: differentiation from Wilson's disease. *Gasteroenterology* 1976;**70**:653-5.

Lee 1998

Lee JJ, Kim HJ, Chung IJ, Kook H, Byun JR, Kwon SY, et al. Acutehemolytic crisis with fulminant hepatic failure as the first manifestation of Wilson's disease: a case report. *Journal of Korean Medical Science* 1998;**13**(5):548-50.

Merle 2007

Merle U, Schaefer M, Ferenci P, Stremmel W. Clinical presentation, diagnosis and long-term outcome of Wilson's disease: a cohort study. *Gut* 2007;**56**(1):115-20.

Ostapowicz 2002

Ostapowicz G, Fontana RJ, Schiodt FV, Larson A, Davern TJ, Han SH, et al. Results of a prospective study of acute liver failure at 17 tertiary care centers in the United States. *Annals of Internal Medicine* 2002;**137**(12):947-54.

Perlman 1979

Perman JA, Werlin SL, Grand RJ, Watkins JB. Laboratory measures of copper metabolism in the differentiation of chronic active hepatitis and Wilson disease in children. *Journal of Pediatrics* 1979;**94**(4):564-8.

Pfeiffenberger 2019

Pfeiffenberger J, Lohse CM, Gotthardt D, Rupp C, Weiler M, Teufel U, et al. Long-term evaluation of urinary copper excretion and non-caeruloplasmin associated copper in Wilson disease patients under medical treatment. *Journal of Inherited Metabolic Disease* 2019;**42**(2):371-80.

Poujois 2017

Poujois A, Trocello JM, Djebrani-Oussedik N, Poupon J, Collet C, Girardot-Tinant N, et al. Exchangeable copper: a reflection of the neurological severity in Wilson's disease. *European Journal of Neurology* 2017;**24**(1):154-60.

PRISMA 2009

Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analysis: the PRISMA statement.. *PLOS Medicine* 2009;**6**(7):1-6.

Roberts 2008

Roberts EA, Schilsky ML. American Association for Study of Liver Diseases (AASLD). Diagnosis and treatment of Wilson disease: an update. *Hepatology* 2008;**47**(6):2089-111.

Rutter 2001

Rutter CA, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in Medicine* 2001;**20**(19):2865-84.

Sanchez-Albisua 1999

Sánchez-Albisua I, Garde T, Hierro L, Camarena C, Frauca E, de la Vega A. A high index of suspicion: the key to an early diagnosis of Wilson's disease in childhood. *Journal of Pediatric Gastroenterology and Nutrition* 1999;**28**(2):186-90.

Steindl 1997

Steindl P, Ferenci P, Dienes HP, Grimm G, Pabinger I, Madl C, et al. Wilson's disease in patients presenting with liver disease: a diagnostic challenge. *Gasteroenterology* 1997;**113**(1):212-8.

Sternlieb 1990

Sternlieb I. Perspectives on Wilson's disease. *Hepatology* 1990;**12**(5):1234-39.

Taly 2007

Taly AB, Meenakshi-Sundaram S, Sinha S, Swarmy HS, Arunodaya GR. Wilson disease: description of 282 patients evaluated over 3 decades. *Medicine* 2007;**86**(2):112-21.

Thomas 1995

Thomas GR, Forbes JR, Roberts EA, Walshe JM, Cox DW. The Wilson disease gene: spectrum of mutations and their consequences. *Nature Genetics* 1995;**9**(2):210-7.

Tietz 2012a

Hortin GL. Chapter 21 Amino acids, Peptides and Proteins. In: Burtis CA, Ashwood ER, Burns DE editor(s). Tietz Textbook of Clinical Chemistry and Molecular Diagnostics. 5th Edition. Elsevier, 2012:533-5.

Tietz 2012b

Kricka LJ, Park JY. Chapter 10 Optical Techniques. In: Burtis CA, Ashwood ER, Burns DE editor(s). Tietz Textbook of Clinical Chemistry and Molecular Diagnostics. 5th Edition. Elsevier, 2012:243-4.

Tietz 2012c

Shenkin A, Roberts NB. Chapter 31 Vitamins and Trace Elements. In: Burtis CA, Ashwood ER, Burns DE editor(s). Tietz Textbook of Clinical Chemistry and Molecular Diagnostics. 5th Edition. Elsevier, 2012:942-3.

Twomey 2005

Twomey PJ, Viljoen A, House IM, Reynolds TM, Wierzbicki AS. Relationship between serum copper, ceruloplasmin, and non-ceruloplasmin-bound copper in routine clinical practice. *Clinical Chemistry* 2005;;**51**(8):1558-9.

UK NEQAS 2018

UK NEQAS for trace elements. External quality assurance scheme for trace elements. EQA SCHEME REPORT (accessed 06 October 2018).

Walshe 2003

Walshe JM. Wilson's disease: the importance of measuring serum caeruloplasmin non-immunologically. *Annals of Clinical Biochemistry* 2003;**40**(Pt 2):115-21.

Walshe 2011

Walshe JM. The eye in Wilson disease. *QJM* 2011;**104**(5):451-3.

Whiting 2011

Whiting PF, Rutjes AWS, Westwood ME, Mallet S, Deeks JJ, Reitsma JB. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine* 2011;**155**(8):529-36.

Wilson 1987

Willson RA, Clayson KJ, Leon S. Unmeasurable serum alkaline phosphatase activity in Wilson's disease associated with fulminant hepatic failure and hemolysis. *Hepatology* 1987;**7**(3):613-5.

Zegers 2013

Zegers I, Beetham R, Keller T, Sheldon J, Bullock D, MacKenzie F, et al. The importance of commutability of reference materials used as calibrators: the example of ceruloplasmin. *Clinical Chemistry* 2013;**59**(9):1322-9.

References to other published versions of this review

Ryan 2016

Ryan A, Nolan SJ, Cook P. Biomarkers for diagnosis of Wilson's disease. *Cochrane Database of Systematic Reviews* 2016, Issue 6. [DOI: 10.1002/14651858.CD012267]

CHARACTERISTICS OF STUDIES

Characteristics of included studies *[ordered by study ID]*

Ferenci 2005

Study characteristics			
Patient sampling	Case-control study		
Patient characteristics and setting	People with WD (hepatic, neuropsychiatric or asymptomatic), normal controls and those with alternative confirmed hepatic pathology		
Index tests	Hepatic copper measured using atomic absorption spectroscopy		
Target condition and reference standard(s)	WD and the Leipzig criteria		
Flow and timing	Participant flow not clearly delineated		
Comparative			
Notes	Given case-control design, selection bias is the main potential cause for bias		
Methodological quality			
Item	**Authors' judgement**	**Risk of bias**	**Applicability concerns**
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	No		

Ferenci 2005 *(Continued)*

Was a case-control design avoided?	No		
Did the study avoid inappropriate exclusions?	Yes		
		High	**Low**
DOMAIN 2: Index Test All tests			
Were the index test results interpreted without knowledge of the results of the reference standard?	No		
If a threshold was used, was it pre-specified?	Yes		
		Low	**Low**
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	No		
Does prior knowledge of a single index test result affect reference standard conduct?	No		
		High	**Low**
DOMAIN 4: Flow and Timing			
Was there an appropriate interval between index test and reference standard?	No		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
		Low	

Lu 2010

Study characteristics	
Patient sampling	Case-control study
Patient characteristics and setting	Hepatic disease unknown
Index tests	24-hour urinary copper measured by ICP mass spectrometry
Target condition and reference standard(s)	WD and Leipzig criteria
Flow and timing	Flow of participants not clearly delineated
Comparative	

Lu 2010 *(Continued)*

Notes	Given the case-control design, selection bias is likely to be main source of bias

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	No		
Was a case-control design avoided?	No		
Did the study avoid inappropriate exclusions?	Yes		
		High	**Low**
DOMAIN 2: Index Test All tests			
Were the index test results interpreted without knowledge of the results of the reference standard?	No		
If a threshold was used, was it pre-specified?	Yes		
		Low	**Low**
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	No		
Does prior knowledge of a single index test result affect reference standard conduct?	No		
		High	**Low**
DOMAIN 4: Flow and Timing			
Was there an appropriate interval between index test and reference standard?	No		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
		Low	

Mak 2008

Study characteristics	
Patient sampling	Case-control study

Mak 2008 *(Continued)*

Patient characteristics and setting	Those with WD (hepatic, neuropsychiatric or asymptomatic), normal controls, undiagnosed hepatic or neurological deficit
Index tests	Caeruloplasmin using nephelometry Beckman Coulter IMMAGE
Target condition and reference standard(s)	WD and the Leipzig criteria
Flow and timing	Participants tested at presentation prior to commencement of chelation therapy
Comparative	
Notes	Due to the nature of the clinical reference standard potential for bias is a possibility

Methodological quality

Item	**Authors' judgement**	**Risk of bias**	**Applicability concerns**
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	No		
Was a case-control design avoided?	No		
Did the study avoid inappropriate exclusions?	Yes		
		High	**Low**
DOMAIN 2: Index Test All tests			
Were the index test results interpreted without knowledge of the results of the reference standard?	No		
If a threshold was used, was it pre-specified?	Yes		
		Low	**Low**
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	No		
Does prior knowledge of a single index test result affect reference standard conduct?	No		
		High	**Low**
DOMAIN 4: Flow and Timing			
Was there an appropriate interval between index test and reference standard?	No		

Mak 2008 *(Continued)*

Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
		Low	

Merle 2009

Study characteristics			
Patient sampling	Case-control study		
Patient characteristics and setting	People with WD (hepatic or neurological), normal controls and those with alternative confirmed hepatic pathology		
Index tests	Caeruloplasmin nephelometry Dade Behring		
Target condition and reference standard(s)	WD and the Leipzig criteria		
Flow and timing	Flow of participants not clearly delineated		
Comparative			
Notes	Given the case-control design, selection is likely to be the main source of bias		

Methodological quality

Item	**Authors' judgement**	**Risk of bias**	**Applicability concerns**
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		
Was a case-control design avoided?	No		
Did the study avoid inappropriate exclusions?	Yes		
		High	**Low**
DOMAIN 2: Index Test All tests			
Were the index test results interpreted without knowledge of the results of the reference standard?	No		
If a threshold was used, was it pre-specified?	No		
		High	**Low**
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Yes		

Merle 2009 *(Continued)*

Were the reference standard results interpreted without knowledge of the results of the index tests?	No		
Does prior knowledge of a single index test result affect reference standard conduct?	No		
		High	**Low**
DOMAIN 4: Flow and Timing			
Was there an appropriate interval between index test and reference standard?	No		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
		Low	

Nicastro 2010

Study characteristics			
Patient sampling	Case-control study		
Patient characteristics and setting	People with WD (hepatic or asymptomatic family screening), normal controls and those with alternative confirmed hepatic pathology		
Index tests	Caeruloplasmin by radial immunodiffusion NOR-Partigen Behring Urine copper by flame absorption spectrophotometry Hepatic copper by flame absorption spectrophotometry		
Target condition and reference standard(s)	WD and the Leipzig criteria		
Flow and timing	Flow of participants and timing of tests clearly delineated		
Comparative			
Notes	Given the case-control study design, selection bias is likely to be the main source of bias		
Methodological quality			
Item	**Authors' judgement**	**Risk of bias**	**Applicability concerns**
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	No		

Nicastro 2010 *(Continued)*

Was a case-control design avoided?	No		
Did the study avoid inappropriate exclusions?	Yes		
		High	**Low**
DOMAIN 2: Index Test All tests			
Were the index test results interpreted without knowledge of the results of the reference standard?	No		
If a threshold was used, was it pre-specified?	Yes		
		Low	**Low**
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	No		
Does prior knowledge of a single index test result affect reference standard conduct?	No		
		High	**Low**
DOMAIN 4: Flow and Timing			
Was there an appropriate interval between index test and reference standard?	No		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
		Low	

Sezer 2014

Study characteristics	
Patient sampling	Case-control study
Patient characteristics and setting	People with WD (hepatic), normal controls and those with alternative confirmed hepatic pathology
Index tests	Hepatic copper by atomic absorption spectrophotometry AA-6701F Shimadzu; urine copper by atomic absorption spectrophotometry AA-6701F Shimadzu; caeruloplasmin by immunoturbidimetry Roche Modular
Target condition and reference standard(s)	WD and the Leipzig criteria

Sezer 2014 *(Continued)*

Flow and timing	Flow of participants and timing clearly delineated
Comparative	
Notes	Given the case-control study design, selection bias is likely to be the main source of bias

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	No		
Was a case-control design avoided?	No		
Did the study avoid inappropriate exclusions?	Yes		
		High	**Low**
DOMAIN 2: Index Test All tests			
Were the index test results interpreted without knowledge of the results of the reference standard?	No		
If a threshold was used, was it pre-specified?	Yes		
		Low	**Low**
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	No		
Does prior knowledge of a single index test result affect reference standard conduct?	No		
		High	**Low**
DOMAIN 4: Flow and Timing			
Was there an appropriate interval between index test and reference standard?	No		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
		Low	

Xu 2018

Study characteristics			
Patient sampling	Cross-sectional study		
Patient characteristics and setting	All participants that had ceruloplasmin analysed were eligible. Testing and records were accessed in a university hospital centre		
Index tests	Ceruloplasmin by nephelometry, Beckman Coulter Immage		
Target condition and reference standard(s)	WD and the Leipzig criteria		
Flow and timing	Flow of participants clearly delineated		
Comparative			
Notes	Due to the cross-sectional study design, selection bias is likely to be the main source of bias		

Methodological quality			
Item	**Authors' judgement**	**Risk of bias**	**Applicability concerns**
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	No		
Was a case-control design avoided?	Yes		
Did the study avoid inappropriate exclusions?	Yes		
		High	**Low**
DOMAIN 2: Index Test All tests			
Were the index test results interpreted without knowledge of the results of the reference standard?	No		
If a threshold was used, was it pre-specified?	Yes		
		Low	**Low**
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	No		
Does prior knowledge of a single index test result affect reference standard conduct?	No		
		High	**Low**

Xu 2018 *(Continued)*

DOMAIN 4: Flow and Timing			
Was there an appropriate interval between index test and reference standard?	Unclear		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
		Low	

Yang 2015

Study characteristics			
Patient sampling	Prospective cohort		
Patient characteristics and setting	People with suspected WD (hepatic), relatives of people with WD and those with alternative confirmed hepatic pathology		
Index tests	Hepatic copper by atomic absorption spectrophotometry Bejing Purkinje General Instruments		
Target condition and reference standard(s)	WD and the Leipzig criteria		
Flow and timing	Flow of participants clearly delineated		
Comparative			
Notes	Design and conduct of study minimised risk of bias leading to inaccurate conclusion with population selected reflecting clinical practice		

Methodological quality			
Item	**Authors' judgement**	**Risk of bias**	**Applicability concerns**
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	No		
Was a case-control design avoided?	Yes		
Did the study avoid inappropriate exclusions?	Yes		
		High	**Low**
DOMAIN 2: Index Test All tests			
Were the index test results interpreted without knowledge of the results of the reference standard?	No		
If a threshold was used, was it pre-specified?	Yes		

Yang 2015 *(Continued)*

		Low	**Low**
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	No		
Does prior knowledge of a single index test result affect reference standard conduct?	No		
		High	**Low**
DOMAIN 4: Flow and Timing			
Was there an appropriate interval between index test and reference standard?	No		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
		Low	

ICP: inductively coupled plasma; WD: Wilson's disease

Characteristics of excluded studies *[ordered by study ID]*

Study	Reason for exclusion
Causa 1997	Did not use Leipzig criteria for diagnosis of WD
Frommer 1981	Did not use Leipzig criteria for diagnosis of WD
Gibbs 1979	Did not use Leipzig criteria for diagnosis of WD
Gnanou 2006	Method comparison study undiagnosed cohort
Lech 2007	Did not use Leipzig criteria for diagnosis of WD
Li 1983	Did not use Leipzig criteria for diagnosis of WD
Liggi 2013	Case-series of people with WD, no controls
Mahjoub 2012	Lack of WD definition and clarity of methods used
Markowitz 1955	Lack of WD definition and clarity of methods used
Mzhel'skaia 1994	Did not use Leipzig criteria for diagnosis of WD
Prasad 1998	Caeruloplasmin oxidase method used which is not used for Leipzig criteria. Also unclear if children are or were on any chelation therapy
Siotto 2014	Caeruloplasmin oxidase method used which is not used for Leipzig criteria

WD: Wilson's disease

Characteristics of studies awaiting classification *[ordered by study ID]*

Aksu 2018

Study characteristics	
Patient sampling	Case-control, convenience
Patient characteristics and setting	66 children with confirmed WD and 88 children without WD
Index tests	24-hour urinary copper levels
Target condition and reference standard(s)	WD and unclear reference standard, abstract only
Flow and timing	Flow of participants not clearly delineated, abstract only
Comparative	Unclear abstract only
Notes	

Zhou 2019

Study characteristics	
Patient sampling	Case-control, convenience
Patient characteristics and setting	40 people with WD, 40 carriers and 20 normal controls
Index tests	Caeruloplasmin and 24-hour urinary copper
Target condition and reference standard(s)	WD, unclear reference standard, abstract only
Flow and timing	Flow of participants not clearly delineated, abstract only
Comparative	Unclear, abstract only
Notes	

Abbreviations: WD: Wilson's disease

DATA

Presented below are all the data for all of the tests entered into the review.

Table Tests. Data tables by test

Test	No. of studies	No. of participants
1 Caeruloplasmin (Leipzig criteria threshold 0.1 g/L)	2	291

Test	No. of studies	No. of participants
2 Caeruloplasmin (Leipzig criteria threshold 0.2 g/L)	4	4415
3 Caeruloplasmin (other threshold 0.14 g/L)	2	291
4 Caeruloplasmin (other threshold 0.15 g/L)	1	4048
5 Caeruloplasmin (other threshold 0.18 g/L)	1	98
6 Caeruloplasmin (other threshold 0.19 g/L)	1	161
7 Hepatic copper (Leipzig criteria threshold > 4 µmol/g)	4	1128
8 Hepatic copper (other threshold 1.2 µmol/g)	1	333
9 Hepatic copper (other threshold 1.5 µmol/g)	1	76
10 Hepatic copper (other threshold 3.3 µmol/g)	1	643
11 24-hour urinary copper (Leipzig criteria threshold 0.64 µmol/24 hours)	1	96
12 24-hour urinary copper (Leipzig criteria threshold 1.6 µmol/24 hours)	3	266
13 24-hour urinary copper (threshold 0.8 µmol/24 hours)	1	94
14 24-hour urinary copper (threshold 1.06 µmol/24 hours)	1	76

Test 1. Caeruloplasmin (Leipzig criteria threshold 0.1 g/L).

Test 2. Caeruloplasmin (Leipzig criteria threshold 0.2 g/L).

Test 3. Caeruloplasmin (other threshold 0.14 g/L).

Test 4. Caeruloplasmin (other threshold 0.15 g/L).

Test 5. Caeruloplasmin (other threshold 0.18 g/L).

Test 6. Caeruloplasmin (other threshold 0.19 g/L).

Test 7. Hepatic copper (Leipzig criteria threshold > 4 µmol/g).

Test 8. Hepatic copper (other threshold 1.2 µmol/g).

Test 9. Hepatic copper (other threshold 1.5 µmol/g).

Test 10. Hepatic copper (other threshold 3.3 µmol/g).

Test 11. 24-hour urinary copper (Leipzig criteria threshold 0.64 µmol/24 hours).

Test 12. 24-hour urinary copper (Leipzig criteria threshold 1.6 µmol/24 hours).

Test 13. 24-hour urinary copper (threshold 0.8 µmol/24 hours).

Test 14. 24-hour urinary copper (threshold 1.06 µmol/24 hours).

ADDITIONAL TABLES

Table 1. Leipzig criteria

Criteria	Score*
KF rings	
Present	2
Absent	0
Neurology	
Severe	2
Mild	1
Absent	0
Caeruloplasmin	

Table 1. Leipzig criteria *(Continued)*

> 0.2 g/L	0
0.1 - 0.2 g/L	1
< 0.1 g/L	2
Coombs negative	
Present	1
Absent	0
Liver copper	
> 4 µmol/g	2
0.8 - 4.0 µmol/g	1
< 0.8 µmol/g	0
Rhodaine positive	1
Urinary copper	
Normal	0
1 - 2 x upper limit normal	1
> 2 x upper limit normal	2
Normal but > 5 x upper limit normal after D-penicillamine	2
Mutational analysis	
2 chromosomes affected	4
1 chromosome affected	1
No mutation detected	0

***Total score evaluation:** 4 or more: diagnosis established; 3: more tests needed; 2 or less: diagnosis very unlikely
Abbreviations: KF: Kayser-Fleischer

Table 2. Characteristics of included studies

Study ID	Index test assay	Threshold	Mean, median or age range of WD patients	WD clinical presentation	Number of patients
Ferenci 2005	**Hepatic copper**	**Leipzig criteria**: > 4 µmol/g	Median age	Hepatic, neuropsychiatric	*Participants included in DTA analyses of this review* • 114 children and adults with WD

Table 2. Characteristics of included studies *(Continued)*

	Flame atomic absorption spectroscopy (Kingston 1986)	**Other** ('most useful'): 1.2 µmol/g	18 years for hepatic, 25 years for neuropsychiatric, 18 years for sibling	or asymptomatic	• 219 chronic liver disease controls *Participants not included in DTA analyses of this review* • 26 normal controls (included for reference interval of hepatic copper content)
Lu 2010	**24-hour urinary copper** Inductively coupled plasma mass spectrometry	**Other** ('best fitting'): 0.8 µmol/24 hours **Other** (comparator threshold): 1.54 µmol/24 hours	Mean age 7.4 years	Hepatic	*Participants included in DTA analyses of this review* • 26 children with WD • 68 with non-WD-related chronic liver disease
Mak 2008	**Caeruloplasmin** Nephelometry Beckman Coulter IMMAGE	**Leipzig criteria**: 0.2 g/L and 0.1 g/L **Other** ('most useful'): 0.14 g/L	Age range 4 - 50 years	Hepatic, neuropsychiatric or asymptomatic	*Participants included in DTA analyses of this review* • 59 children and adults with WD • 71 family members (49 heterozygotes, 22 wild type homozygotes), • Validation group of 25 with suspected WD • 40 normal controls with serum ceruloplasmin concentrations ≤ 0.20 g/L *Participants not included in DTA analyses of this review* • 650 normal controls with serum ceruloplasmin concentrations > 0.20 g/L (included for reference interval of serum ceruloplasmin)
Merle 2009	**Caeruloplasmin** Nephelometry Dade Behring Germany	**Other** ('greatest sum on sensitivity and specificity'): 0.19 g/L	Median age 37 years	Hepatic, neurological or asymptomatic	*Participants included in DTA analyses of this review* • 110 adults with WD • 51 with cirrhosis not due to WD *Participants not included in DTA analyses of this review* • 52 healthy controls (DTA analyses of serum ceruloplasmin oxidase activity of WD participants and healthy controls were conducted in the original study at the same threshold)
Nicastro 2010	**Caeruloplasmin** Radial Immunodiffusion NOR-Partigen Coeruloplasmin	*Caeruloplasmin* **Leipzig criteria**: 0.2 g/L and 0.1 g/L **Other** (from Mak 2008): 0.14 g/L and 0.18 g/L	Median age 6.1 years	Hepatic, neurological or asymptomatic family screening	*Participants included in DTA analyses of this review* • 40 children with WD • 58 age, sex-matched with non-WD-related chronic liver disease

Table 2. Characteristics of included studies *(Continued)*

	Marburg Behring, Germany **Hepatic and 24-hour urinary copper** Flame atomic absorption spectroscopy (Kelson 1978)	*Hepatic copper* **Leipzig criteria**: > 4 µmol/g *24-hour urinary copper* 0.6 µmol/24 hours 1.6 µmol/24 hours			
Sezer 2014	**Caeruloplasmin,** **Hepatic and 24-hour urinary copper** Atomic absorption spectroscopy AA6701F, Shimadzu, Japan	*Caeruloplasmin* **Leipzig criteria**: 0.2 g/L *Hepatic copper* **Leipzig criteria**: > 4 µmol/g **Other** ('optimal value') 1.5 µmol/g for hepatic copper *24-hour urinary copper* 1.06 µmol/24 hours 1.6 µmol/24 hours	Mean age 10.2 years	Hepatic	*Participants included in DTA analyses of this review* • 35 children with WD • 41 with non-WD-related chronic liver disease
Yang 2015	**Hepatic copper** Flame atomic absorption spectroscopy Model TAS-986, Beijing Purkinje General Instruments, Beijing China	**Leipzig criteria**: > 4 µmol/g **Other** ('most useful'): 3.3 µmol/g	Mean age 19.7 years	Hepatic, neurological or presymptomatic	*Participants included in DTA analyses of this review* • 178 children and adults with WD • 465 with non-WD-related chronic liver disease (in the absence of primary biliary cirrhosis)
Xu 2018	**Ceruloplasmin** Nephelometry Beckman Coulter Image Immunchemistry System, Brea Ca USA	**Leipzig criteria**: 0.2 g/L **Other** ('most useful'): 0.15 g/L	Mean age 21.8 years	Hepatic, neurological or presymptomatic	*Participants included in DTA analyses of this review* • 297 children and adults with WD • 3751 with non-WD-related chronic liver disease and nephrotic syndrome

Abbreviations: CI: Confidence interval; DTA: diagnostic test accuracy; USA: United States of America; WD: Wilson's disease

Table 3. Diagnostic test accuracy of index tests - caeruloplasmin *(Continued)*

Index test: caeruloplasmin		Wilson's disease		Sensitivity (95% CI)[a]	Specificity (95% CI)[a]	PPV (95% CI)[a]	NPV (95% CI)[a]	LR+ (95% CI)[a]	LR- (95% CI)
		Positive	Negative						
Threshold(s) according to the Leipzig criteria									
Study ID: Mak 2008 **Threshold**: 0.1 g/L	Positive	45	0	78.9% (66.1% to 88.6%)	100% (97.3% to 100.0%)	100% (NE)[b]	91.9% (87.3% to 94.9%)	NE[b]	0.21 (0.13 to 0.35)
	Negative	12	136						
Study ID: Nicastro 2010 **Threshold**: 0.1 g/L	Positive	26	2	65.0% (48.3% to 79.4%)	96.6% (88.1% to 99.6%)	92.9% (76.6 to 98.1%)	80.0% (72.3% to 86.0%)	18.9 (4.74 to 75.0)	0.36 (0.24 to 0.55)
	Negative	14	56						
Study ID: Mak 2008 **Threshold**: 0.2 g/L	Positive	56	60	98.3% (90.6% to 100.0%)	55.9% (47.1% to 64.4%)	48.3% (43.5% to 53.1%)	98.7% (91.6% to 99.8%)	2.23 (1.84 to 2.70)	0.03 (0.00 to 0.22)
	Negative	1	76						
Study ID: Nicastro 2010 **Threshold**: 0.2 g/L	Positive	38	10	95.0% (83.1% to 99.4%)	82.8% (70.6% to 91.4%)	79.2% (68.3% to 87.0%)	96.0% (86.1% to 98.9%)	5.51 (3.12 to 9.73)	0.06 (0.02 to 0.23)
	Negative	2	48						
Study ID: Sezer 2014 **Threshold**: 0.2 g/L	Positive	27	14	77.1% (59.9% to 89.6%)	65.9% (49.4% to 79.9%)	65.9% (54.9% to 75.4%)	77.1% (63.9% to 86.6%)	2.26 (1.42 to 3.59)	0.35 (0.18 to 0.66)
	Negative	8	27						
Study ID: Xu 2018 **Threshold**: 0.2 g/L	Positive	294	716	99.0% (97.1% to 99.8%)	80.9% (79.6% to 82.2%)	29.1% (27.8% to 30.5%)	99.9% (99.7% to 100.0%)	5.19 (4.85 to 5.54)	0.01 (0.00 to 0.04)
	Negative	3	3035						
Other threshold(s) reported[c]									

Table 3. Diagnostic test accuracy of index tests - caeruloplasmin *(Continued)*

Study ID: Mak 2008 **Threshold**: 0.14 g/L	Positive	53	0	93.0% (83.0% to 98.1%)	100% (97.3% to 100.0%)	100% (NE)[b]	97.1% (93.0 to 98.9%)	NE[b]	0.07 (0.03 to 0.18)
	Negative	4	136						
Study ID: Nicastro 2010 **Threshold**: 0.14 g/L	Positive	28	4	70.0% (53.5% to 83.4%)	93.1% (83.3% to 98.1%)	87.5% (72.7% to 94.9%)	81.8% (73.6% to 87.9%)	10.2 (3.86 to 26.7)	0.32 (0.20 to 0.52)
	Negative	12	54						
Study ID: Xu 2018 **Threshold**: 0.15 g/L	Positive	284	169	95.6% (92.6% to 97.7%)	95.5% (94.8% to 96.1%)	62.7% (59.1% to 66.1%)	99.6% (99.3% to 99.8%)	21.2 (18.3 to 24.6)	0.05 (0.03 to 0.08)
	Negative	13	3582						
Study ID: Nicastro 2010 **Threshold**: 0.18 g/L	Positive	32	5	80.0% (64.4% to 91.0%	91.4% (81.0% to 97.1%)	86.5% (73.2% to 93.8%)	86.9% (78.0% to 92.5%)	9.28 (3.96 to 21.8)	0.22 (0.12 to 0.41)
	Negative	8	53						
Study ID: Merle 2009 **Threshold**: 0.19 g/L	Positive	103	21	93.6% (87.3% to 97.4%)	58.8% (44.2% to 72.4%)	83.1% (77.9% to 87.2%)	81.0% (66.9% to 90.1%)	2.27 (1.63 to 3.17)	0.11 (0.05 to 0.23)
	Negative	7	30						

Abbreviations: CI: confidence interval; PPV: positive predictive value; NPV: negative predictive value; LR+: positive-likelihood ratio; LR-: negative-likelihood ratio; NA: not available; NE: not estimable.

a. Due to rounding to the appropriate number of decimal places, CIs presented may not be symmetric around the point estimate

b. Not estimable: No false positives (FP = 0) results in a LR+ of infinity and CIs of 100% to 100% for PPV

c. 'Other' threshold reported in the studies were defined as the 'most useful cut-off value' (Mak 2008, Xu 2018), 'the greatest sum of sensitivity and specificity of immunoreactive ceruloplasmin concentrations' (Merle 2009) and the other thresholds examined in Nicastro 2010 were informed by Mak 2008

Table 4. Diagnostic test accuracy of index tests - hepatic copper *(Continued)*

Index test: hepatic hopper	**Wilson's disease**		**Sensitivity**	**Specificity**	**PPV**	**NPV**	**LR+**	**LR-**
	Positive	**Negative**	**(95% CI)[a]**	**(95% CI)[a]**	**(95% CI)[a]**	**(95% CI)[a]**	**(95% CI)[a]**	**(95% CI)[a]**

Table 4. Diagnostic test accuracy of index tests - hepatic copper *(Continued)*

Threshold according to the Leipzig criteria									
Study ID: Ferenci 2005 **Threshold**: >4 µmol/g	Postive	95	3	83.3% (75.2% to 89.7%)	98.6% (96.1% to 99.7%)	96.9% (91.1% to 99.0%)	91.9% (88.3% to 94.5%)	60.8 (19.7 to 187)	0.17 (0.11 to 0.25)
	Negative	19	216						
Study ID: Nicastro 2010 **Threshold**: >4 µmol/g	Positive	28	22	93.3% (77.9% to 99.2%)	52.2% (37.0% to 67.1%)	56.0% (48.1% to 63.6%)	92.3% (75.4% to 97.9%)	1.95 (1.42 to 2.68)	0.13 (0.03 to 0.50)
	Negative	2	24						
Study ID: Sezer 2014 **Threshold**: > 4 µmol/g	Positive	23	10	65.7% (47.8% to 80.9%)	75.6% (59.7% to 87.6%)	69.7% (56.1% to 80.6%)	72.1% (61.3% to 80.8%)	2.69 (1.49 to 4.86)	0.45 (0.28 to 0.74)
	Negative	12	31						
Study ID: Yang 2015 **Threshold**: > 4 µmol/g	Positive	168	15	94.4% (89.9% to 97.3%)	96.8% (94.7% to 98.2%)	91.8% (87.2% to 94.9%)	97.8% (96.1% to 97.5%)	29.3 (17.8.8 to 48.2)	0.06 (0.03 to 0.11)
	Negative	10	450						
Other thresholds reported[b]									
Study ID: Ferenci 2005 **Threshold**: 1.2 µmol/g	Positive	110	10	96.5% (91.3% to 99.0%)	95.4% (91.8% to 97.8%)	91.7% (85.7% to 95.3%)	98.1% (95.2% to 99.3%)	21.1 (11.5 to 38.8)	0.04 (0.01 to 0.10)
	Negative	4	209						
Study ID: Sezer 2014 **Threshold**: 1.5 µmol/g	Positive	32	15	91.4% (76.9% to 98.2%)	63.4% (46.9% to 77.9%)	68.1% (58.5% to 76.4%)	89.7% (74.1% to 96.3%)	2.50 (1.65 to 3.79)	0.14 (0.04 to 0.41)
	Negative	3	26						
Study ID: Yang 2015 **Threshold**: 3.3 µmol/g	Positive	177	18	99.4% (96.9% to 100.0%)	96.1% (94.0% to 97.7%)	90.8% (86.2% to 93.9%)	99.8% (98.4% to 100.0%)	25.7 (16.3 to 40.4)	0.01 (0.00 to 0.04)
	Negative	1	447						

Abbreviations: CI: confidence interval; PPV: positive-predictive value; NPV: negative-predictive value; LR+: positive-likelihood ratio; LR-:negative-likelihood ratio; NA: not available; NE: not estimable; WD: Wilson's disease

a. Due to rounding to the appropriate number of decimal places, CIs presented may not be symmetric around the point estimate

b. 'Other' threshold reported in the studies were defined as the 'most useful cut-off value' (Ferenci 2005; Yang 2015) or the 'optimum' cut-off value (Sezer 2014)

Table 5. Diagnostic test accuracy of index tests - 24-hour urinary copper *(Continued)*

Index Test: 24-hour urinary copper		Wilson's disease		Sensitivity (95% CI)[a]	Specificity (95% CI)[a]	PPV (95% CI)[a]	NPV (95% CI)[a]	LR+ (95% CI)[a]	LR- (95% CI)[a]
		Positive	Negative						
Threshold (s) according to the Leipzig criteria									
Study ID: Nicastro 2010 **Threshold**: 0.64 µmol/24 hours	Positive	30	7	78.9% (62.7% to 90.5%)	87.9% (76.7% to 95.0%)	81.1% (67.7% to 89.7%)	86.4% (77.4% to 92.2%)	6.54 (3.2 to 13.3)	0.24 (0.13 to 0.45)
	Negative	8	51						
Study ID: Lu 2010 **Threshold**: 1.6 µmol/24 hours	Positive	13	2	50.0% (29.9% to 70.1%)	97.1% (89.8% to 99.6%)	86.7% (61.1% to 96.4%)	83.5% (77.5% to 88.2%)	17.0 (4.12 to 70.2)	0.52 (0.35 to 0.76)
	Negative	13	66						
Study ID: Nicastro 2010 **Threshold**: 1.6 µmol/24 hours	Positive	25	1	65.8% (48.7% to 80.4%)	98.3% (90.8% to 100.0%)	96.2% (77.9% to 99.4%)	81.4% (73.8% to 87.2%)	38.2 (5.39 to 269)	0.35 (0.22 to 0.54)
	Negative	13	57						
Study ID: Sezer 2014 **Threshold**: 1.6 µmol/24 hours	Positive	28	10	80.0% (63.1% to 91.6%)	75.6% (59.7% to 87.6%)	73.7% (61.4% to 83.1%)	81.6% (69.1% to 86.4%)	3.28 (1.87 to 5.76)	0.26 (0.13 to 0.52)
	Negative	7	31						
Other thresholds reported[b]									
Study ID: Lu 2010 **Threshold**: 0.8 µmol/24 hours	Positive	22	6	84.6% (65.1% to 95.6%)	91.2% (81.8% to 96.7%)	78.6% (62.7% to 88.9%)	93.9% (86.3% to 97.5%)	9.59 (4.39 to 21.0)	0.17 (0.07 to 0.42)
	Negative	4	62						
Study ID: Sezer 2014	Positive	31	12	88.6%	70.7%	72.1%	87.9%	3.03	0.16
	Negative	4	29						

Table 5. Diagnostic test accuracy of index tests - 24-hour urinary copper *(Continued)*

Threshold: 1.06 µmol/24 hours	(73.3% to 96.8%)	(54.5% to 83.8%)	(61.3% to 80.8%)	(73.8% to 94.9%)	(1.85 to 4.94)	(0.06 to 0.41)

Abbreviations: CI: confidence interval; PPV: positive-predictive value; NPV: negative-predictive value; LR+: positive-likelihood ratio; LR-: negative-likelihood ratio; hrs: hours; NA: not available; NE: not estimable

a. Due to rounding to the appropriate number of decimal places, CIs presented may not be symmetric around the point estimate

b. In Lu 2010, 0.8 µmol/24 hours is the 'best fitting' cut-off value (compared to the Leipzig defined cut-off for adults 1.6 µmol/24 hours) and in Sezer 2014, the thresholds were based on the 'optimum' cut-off value

APPENDICES

Appendix 1. Electronic search strategy

Database/ Resource	Strategy
Cochrane Central Register of Controlled Trials (CENTRAL) and Cochrane Database of Systematic Reviews (CDSR)	*[Advanced Search – Search Builder]* #1 MeSH descriptor: [Hepatolenticular Degeneration] this term only #2 (hepatolenticular degeneration) or (hepatic lenticular degeneration) #3 "*wilson* disease" #4 #1 or #2 or #3 #5 MeSH descriptor: [Ceruloplasmin] explode all trees #6 caeruloplasmin or ceruloplasmin or apocaeruloplasmin or apoceruloplasmin or holocaeruloplasmin or holoceruloplasmin or holo-caeruloplasmin or holo-ceruloplasmin #7 "urinary copper" or "urine copper" or "hepatic copper" or "liver copper" #8 #5 or #6 or #7 #9 #4 and #8
PubMed (1946 onwards)	("Hepatolenticular Degeneration"[Mesh] OR hepatolenticular degeneration OR hepatic lenticular degeneration OR “wilson disease” OR “wilsons disease” OR “wilson’s disease”) AND (caeruloplasmin OR apocaeruloplasmin OR apoceruloplasmin OR holocaeruloplasmin OR holoceruloplasmin OR holo-caeruloplasmin OR holo-ceruloplasmin OR “urinary copper” OR “urine copper” OR “hepatic copper” OR “liver copper”)
CINAHL EBSCO (1982 onwards)	((MH "Hepatolenticular Degeneration") OR TX (hepatolenticular degeneration OR hepatic lenticular degeneration OR “wilson disease” OR “wilsons disease” OR “wilson’s disease”)) AND TX ((caeruloplasmin OR ceruloplasmin OR apocaeruloplasmin OR apoceruloplasmin OR holocaeruloplasmin OR holoceruloplasmin OR holo-caeruloplasmin OR holo-ceruloplasmin OR “urinary copper” OR “urine copper” OR “hepatic copper” OR “liver copper”))
Embase Ovid (1982 onwards)	1 WILSON DISEASE/ 2 ((hepatolenticular AND degeneration) OR "hepatic lenticular degeneration" OR "wilson disease" OR "wilsons disease" OR "wilson’s disease").ti,ab 3 1 OR 2 4 CERULOPLASMIN/ 5 (caeruloplasmin OR ceruloplasmin OR apocaeruloplasmin OR apoceruloplasmin OR holocaeruloplasmin OR holoceruloplasmin OR holo-caeruloplasmin OR holo-ceruloplasmin).ti,ab 6 ("urinary copper" OR "urine copper" OR "hepatic copper" OR "liver copper").ti,ab 7 4 OR 5 OR 6 8 3 AND 7
Web of Science’s Conference Proceedings Citation Index (CPCI) (1990 onwards)	((hepatolenticular degeneration) OR (hepatic lenticular degeneration) OR "wilson* disease") [TOPIC] AND

(Continued)

	(caeruloplasmin OR ceruloplasmin OR apocaeruloplasmin OR apoceruloplasmin OR holocaeruloplasmin OR holoceruloplasmin OR holo-caeruloplasmin OR holo-ceruloplasmin OR “urinary copper” OR “urine copper” OR “hepatic copper” OR “liver copper”) [TOPIC]
British Library’s ZETOC (1993 onwards) for conference abstracts	*[Conference Search]* Search 1: “Wilson* disease” [all fields] Search 2: “hepatolenticular degeneration” [all fields]
PROSPERO	#1 MeSH DESCRIPTOR Hepatolenticular Degeneration EXPLODE ALL TREES #2 "Hepatolenticular Degeneration" #3 "hepatic lenticular degeneration" #4 "wilson* disease" #5 #1 OR #2 OR #3 OR #4
Clinicaltrials.gov	Condition/ Disease: Wilson disease Other terms: caeruloplasmin OR ceruloplasmin OR apocaeruloplasmin OR apoceruloplasmin OR holocaeruloplasmin OR holoceruloplasmin OR holo-caeruloplasmin OR holo-ceruloplasmin OR “urinary copper” OR “urine copper” OR “hepatic copper” OR “liver copper”
WHO ICTRP	Wilson disease

Appendix 2. QUADAS-2 methodological assessment tool

QUADAS-2 is structured so that four key domains are each rated in terms of the risk of bias and the concern regarding applicability to the research question (as defined above). Each key domain has a set of signalling questions to help reach the judgments regarding bias and applicability.

Domain 1: Participant selection

Risk of bias - could the selection of participants have introduced bias?

1. Was a consecutive or random sample of patients enrolled?

Assessment	Description
Yes	if the articles clearly states that a consecutive or random samples have been enrolled
No	if it is clear that this was not the case (e.g. if a study included participants 'at the discretion of the clinician')
Unclear	in other cases where it is not clear if consecutive or random samples have been enrolled

2. Was a case-control design avoided?

Assessment	Description

(Continued)

Yes	if the enrolled sample was a random or consecutive enrolment of participants with suspected WD and not separate samples from WD-positive participants and healthy controls
No	if the enrolled samples consist of WD-confirmed cases and healthy controls
Unclear	if the sampling regarding case-control design was not clear

3. Did the study avoid inappropriate exclusions?

Inappropriate exclusions include participants with tremor and chronic liver disease of unknown cause etc.

Assessment	Description
Yes	if inappropriate exclusions were not found in the included study
No	if reasons for inappropriate exclusion were found
Unclear	if there was no description of the inclusion and exclusion criteria and inappropriate exclusion could not be ascertained

4. Could the selection of patients have introduced bias?

Assessment	Description
Low risk	if all questions were scored 'Yes', or a maximum of one question with unclear
High risk	If at least one question was scored as 'No'
Unclear risk	If at least two questions were scored as unclear and one as 'No'

5. Concerns regarding applicability

Is there concern that the included participants do not match the review question?

Assessment	Description
Low concern	if all included participants according to our definition and if they were suspected of WD
High concern	If at least 10% of the included participants were suspected of WD
Unclear concern	if it is unclear whether the study fulfilled either the criteria for low concern or for high concern

Domain 2: Index test(s)

Describe the index test and how it was conducted and interpreted - this will vary with each test method. If more than one index test was used, please complete for each test.

Risk of bias - could the conduct or interpretation of the test have introduced bias?

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

Assessment	Description
Yes	if the index test is conducted and interpreted without the knowledge of the results of the reference standard
No	if the index test is interpreted with the knowledge of the results of the reference standard
Unclear	if it is not clear whether the index test was interpreted without the knowledge of the results of the reference standard

2. If a threshold was used, was it pre-specified?

Assessment	Description
Yes	thresholds were used and were clearly defined
No	thresholds were not used or were not clearly defined
Unclear	unclear whether thresholds were used or predefined

3 Could the conduct or interpretation of the index test have introduced bias?

Assessment	Description
Low risk	if 'yes' classification for both questions above
High risk	if 'no' classification for any of the above two questions
Unclear risk	if 'unclear' classification for any of the above two questions, but without a 'no' classification for any of the above two questions

4. Concerns regarding applicability

Are there concerns that the index test, its conduct, or interpretation differ from the review question?

Assessment	Description

(Continued)

Low concern	if the index test used for the diagnosis of Wilsons was a molecular assay as defined in our protocol
High concern	if the index test used for the diagnosis of Wilsons varies from what was defined in the protocol
Unclear concern	if it is unclear whether the study fulfils criteria for low concern or high concern or if the study provided limited information regarding the conduct and interpretation of the index test

Domain 3: Reference standard

Describe the reference standard and how it was conducted and interpreted.

Risk of bias - could the reference standard, its conduct, or its interpretation have introduced bias?

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

Assessment	Description
Yes	if the reference standard used was consistent with the Leipzig criteria in the diagnosis of WD
No	if the test used as reference standard was a test other than those listed in the Leipzig criteria
Unclear	if there was no description of the reference standard

2. Could prior knowledge of a single index test result affect reference standard conduct?

No being a clinical reference standard, for the Leipzig criteria score the index tests are incorporated into the Leipzig score and a single index test of any of the biochemical tests being evaluated in this review, on it's own will not affect the reference standard conduct.

3. Could the reference standard, its conduct, or its interpretation have introduced bias?

Not applicable as the Leipzig criteria is a clinical reference standard and its application therefore will depend on application of results of clinical, biochemical and genetic results, including the index tests used within the review.

4. Concerns regarding applicability

Are there concerns that the target condition as defined by the reference standard does not match the question?

Assessment	Description
Low concern	if the reference standard was the Leipzig criteria and if the target condition was suspected Wilsons disease in a patient as defined in our protocol
High concern	if the reference standard was a test other than that specified in the Leipzig criteria and if the subjects were not suspected of WD
Unclear concern	if it was unclear whether the study fulfilled either the criteria for low concern or for high concern

Domain 4: Flow and timing

· Describe any participants who did not receive the index test(s) and/or reference standard or who were excluded from the 2 x 2 table (refer to flow diagram).

· Describe the time interval and any interventions between index test(s) and reference standard - again, not sure this applies as it is a clinical reference standard.

Risk of bias - could the participant flow have introduced bias?

1. Was there an appropriate interval between index test(s) and reference standard?

Does not apply - clinical reference standard.

2. Did all participants receive a reference standard?

Assessment	Description
Yes	if all participants underwent testing according to the Leipzig criteria
No	if at least one participant did not have the reference standard performed
Unclear	if the study does not describe clearly which participants received the reference standard and which ones did not

3. Did participants receive the same reference standard?

Assessment	Description
Yes	if all participants underwent testing according to the Leipzig criteria
No	if a different reference standard other than the Leipzig criteria
Unclear	if the study does not describe clearly what type of reference standard was used to diagnose a participant with WD

4. Were all participants included in the analysis?

Assessment	Description
Yes	if all enrolled participants with the target condition that underwent testing using the index test and reference standard were included in the analysis
No	if all enrolled participants were not accounted for in the analysis
Unclear	if it is unclear from the study about the inclusion of all enrolled participants in the analysis

5. Could the participant flow have introduced bias?

Assessment	Description

(Continued)

Low concern	if the answers to above questions were all YES which means that all participants enrolled in the study were subjected to the same reference standard and index test, and all participants were included in the final analysis
High concern	if at least two questions had a 'No' answer
Unclear concern	If at least one question had a 'No' answer or it was unclear whether the study fulfilled either the criteria for low concern or for high concern

Abbreviations: WIlson's disease

Appendix 3. Table used for assessing tests

(Continued)

		Target condition (reference standard)	
		Present	**Absent**
Index test	**Positive (+)**	True positives (TP)	False positives (FP)
	Negative (-)	False negatives (FN)	True negatives (TN)

CONTRIBUTIONS OF AUTHORS

AR and OT validated the search. AR and SN analysed data extraction. AR and OT drafted the final review with input from SN, PC. The final version of the review was agreed by all authors.

DECLARATIONS OF INTEREST

Aidan Ryan: none known.
Sarah Nevitt: none known.
Orla Tuohy: none known.
Paul Cook: none known.

SOURCES OF SUPPORT

Internal sources

- No sources of support supplied

External sources

- National Institute for Health Research, UK.

  This systematic review was supported by the National Institute for Health Research, via Cochrane Infrastructure funding to the Cochrane Cystic Fibrosis and Genetic Disorders Group.

DIFFERENCES BETWEEN PROTOCOL AND REVIEW

Due to being unable to access the Science Citation Index via the Web of Science (1898 to 02 August 2017), this database was not included in the search. However given the other databases searched we do not believe that this has impacted on the conclusions of this review.

In addition to sensitivity and specificity of the tests, we also calculated the positive likelihood ratios, negative likelihood ratios, positive predictive values and negative predictive values of each test, as in the absence of meta-analysis, such statistics may be useful for readers to interpret the narrative results.