PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Fatty liver predicts the risk for cardiovascular events in middle-aged population: a population-based cohort study
AUTHORS	Pisto, Pauliina; Santaniemi, Merja; Bloigu, Risto; Ukkola, Olavi; Kesäniemi, Antero

VERSION 1 - REVIEW

REVIEWER	Dr Matthew Armstrong Wellcome Trust Research Fellow in Hepatology NIHR centre for liver research University of Birmingham
REVIEW RETURNED	17-May-2013

GENERAL COMMENTS	Very topical and relevant to the field. Overall, I commend the authors on a well designed study and analysis. The real strengths of this study are the long follow-up (the longest in the literature) and the detailed metabolic characterization at baseline (ie QUICKI, waist measurements, alcohol, smoking – which previous studies have not uniformly mentioned). Prior to publication however the following points need addressing. Abstract: 'Liver adiposity' change to either liver fat or hepatic steatosis. Please change throughout the manuscript. Strength: 1. This is not the first large prospective population based study addressing this question – see Haring et al (2009), Wong et al (2011), Lazo et al (2011), Zhou et al (2012) and Treeprasertsuk (2012). This is the first northern european study to address this question. Its uniqueness is the duration of follow-up, which far exceeds the previous studies. Introduction: 1. For the prevalence of NAFLD use an original article rather than a review, also specifically use a population-based study – in keeping with your own work (ie Wong GUT et al, Armstrong J Hepatol et al) 2. Last sentence of 1st paragraph is wordy – break up into two sentences. 3. There is no mention in the introduction of the association between NAFLD and stroke – if the literature is lacking please state or otherwise provide a reference 4. There also needs to be a lot more mention of previous CVD studies and NAFLD ie Haring et al (2009), Wong et al (2011), Lazo

	et al (2011), Zhou et al (2012) and Treeprasertsuk (2012). Stating what they showed – which was independent risk, but also pointing out their limitations which was duration of follow-up was less than 10 years in most cases. Also most of these studies don't discuss stroke. Methods: 1. Characterise the radiologists experience (ie approximately how many scans per year etc) 2. Has there severity staging of liver fat accumulation on USS been validated locally with biopsy or MRS in other studies? If so please mention, as this is crucial. 3. Were other causes of Liver disease/steatosis ruled out at baseline (i.e. HBV, HCV). This again is very important. If not please state what the prevalences are in Finland, esp with regards to HBV/HCV (this can placed in the discussion) 4. A huge strength of this study is the characterization of insulin sensitivity (this should be highlighted in the strengths) 5. I presume the bloods were, but please mention 'fasting' if this were so. Results: 1. For model 4 please provide actual p-value rather than NS Discussion: 1. The discussion needs to be a lot more transparent of what data is already in the literature. To the best of my knowledge (if you include Targhers hospital-based studies, which are briefly mentioned) there are 9 prospective studies that have used imaging to define NAFLD and assess the association of CVD development. Please discuss the pros and cons of these studies more, and really emphasize that your patients have extensive follow-up duration, insulin sensitivity and central adiposity (ie waist measurements, which are missing throughout the literature) characterization 2. Would the others be willing to look for how much incident type 2 diabetes occurred in their study, as very few studies have successfully ruled out T2DM at baseline when looking at NAFLD as an independent predictor of T2DM. 3. The limitation is not a 'may be,' it is a fact. USS is limited at detecting hepatic steatosis < 30%, and ideally MRS should be used, but that doesn't come without significant cost. To strengthen the accuracy of the liver USS for detecting fat, the authors could calculate the kotronen score (validated by MRS) and see if this correlates at baseline with their USS scoring. They have all of the components at baseline to calculate this. 4. You need to mention MRS, but state that this would have been costly and at the time of study design was likely not available. 5. Adding more weight to your radiologist experience would be useful here. Very topical and relevant to the field. Overall, I commend the authors on a well designed study and analysis. The real strengths of this study are the long follow-up (the longest in the literature) and the detailed metabolic characterization at baseline (ie QUICKI, waist measurements, alcohol, smoking – which previous studies have not uniformly mentioned). Prior to publication however the above points need addressing, but overall this is a very useful addition to the field. Please ensure 'NAFLD' or 'nonalcoholic fatty liver disease' is in your key words to aid with identifying this study on pubmed.

REVIEWER	Giulio Marchesini, University of Bologna The undersigned person declare no conflict of interest in relation to this review process
REVIEW RETURNED	20-Jul-2013

GENERAL COMMENTS	Discuss the negative findings Limits of the US imaging technique The study addresses a debated question, i.e. the role of liver fat in the determination of cardiovascular risk, by measuring the events occurring in a 18-year follow-up of the OPERA cohort in Finland. The severity of adiposity was measured by ultrasounds and graded as 0 (absent, 1 (mild-moderate) and 2 (severe). The cohort included subjects with hypertension and matched controls. The CV risk associated with the presence of fatty liver was highly significant, after correction for multiple confounders, but disappeared after adjustment for insulin resistance. The authors correctly identify the strengths (long follow-up, well-defined cohort, the use of ICD-9 codes) and the limits (ultrasonographic assessment of fatty liver) of the study. Problems 1. Ultrasonographic assessment remains the crucial problem. Quantitative measurement off liver fat by US is subject to several limitations dictated by old technology (1991) and scarce reproducibility also in expert hands. The 2-level classification is probably more sensitive than a 3-level classification (mild, moderate, severe), but also in this case a considerable uncertainty is expected. We eagerly need studies based on more solid measurements (NMR). 2. The authors should report the inter-/intra-operator reproducibility of the US assessment, based on the techniques and instruments available in 1990. No data are reported in ref 16-17 they quote. 3. Although there is evidence that fatty liver may exacerbate the risk of CV outcomes, not all studies are along this line. The authors are asked to discuss the negative findings (see, Ghouri, HEPATOLOGY 2010;52:1156-1161) 4. Surprisingly, study group did not predict outcome. This means that being hypertensive does not increase the risk. A comment is needed. It would also be important to see whether the prediction is maintained in the two different cohorts of hypertensive and control subjects. 5. CV events in a 19-year follow up are predicted by QUICKI, a very crude assessment of insulin sensitivity. This means that insulin levels were available. It would be interesting to know whether liver enzymes were also present in the dataset and their predictive ability. 6. It is nonetheless surprising that such a crude measure may predict (or cancel the prediction) of CV events. Minor problems 1. I would suggest change the term “accumulation” with “content”. Accumulation refers to a dynamic process, whereas in this case we only have a static measurement of an imaging surrogate of the liver fat “content”.

REVIEWER	Masahide Hamaguchi, MD, PhD, Department of Experimental Immunology, World Premier International Immunology Frontier Research Center,

REVIEW RETURNED	24-Jul-2013

GENERAL COMMENTS	First of all, authors should describe fetal and non-fetal events in each group. Authors mentioned how they counted CVD using ICD-8/9 or 10 at the method section. However, they didn't describe the results. Authors diagnose fatty liver and graded it into 3 grades (0, 1, 2). They referred reference 16 and 17. However, neither reference 16 nor 17 is paper which validate this grading system. In the current paper, only severe fatty liver was a statistically significant factor. The association of moderate fatty liver and CVD was not statistically significant. Thus, the validity of grading system is a critical factor for this study. In method section, authors mentioned how they followed study participants. However, I felt a difficulty to understand. For example, how many people were suffered fetal events and non-fetal events? Were all of participants who were free from CVD followed until December 31, 2009. Or some participants might be dropped out from the follow ups. I guess flow chart might be useful to understand this study. This manuscript has provided important evidence that fatty liver predicted future cardiovascular event, but it is depend on traditional metabolic risk factors. This study had actually long follow up time. However, this study included some critical problems as listed below. Thus, I guess this manuscript might be required a major revision. 1. First of all, authors should describe the details of fetal and non-fetal events in each group. Authors mentioned how they counted CVD using ICD-8/9 or 10 at the method section. However, they didn't describe the results. 2. Authors diagnosed fatty liver and graded them into 3 grades (0, 1, 2). They referred reference 16 and 17. However, neither reference 16 nor 17 is the paper that validates this grading system directly. In the current paper, only severe fatty liver was a statistically significant factor. The association of moderate fatty liver and CVD was not statistically significant. Thus, the validity of grading system for fatty liver is a critical factor for this study. 3. In the paper, authors didn't mention about the metabolic syndrome. It's so curious. The association between fatty liver and metabolic syndrome, or CVD and metabolic syndrome has been well known. Authors should discuss these relationships. Specific comments 4. In introduction section, authors mentioned about NAFLD and NASH. However, the study subjects included participants who consume alcohol regularly. I'm afraid this introduction could mislead readers. 5. In method section, authors mentioned how they followed study participants. However, I felt a difficulty to understand. For example, how many people were suffered fetal events and non-fetal events? Were all of participants who were free from CVD

	followed until December 31, 2009. Or some participants might be dropped out from the follow ups. I guess flow chart might be useful to understand this study. 6. Author used “present tense” at conclusion. I think it is too strong. I guess “past tense” might be suitable for a conclusion of original paper.

VERSION 1 – AUTHOR RESPONSE

Reviewer: Dr Matthew Armstrong

Abstract:

‘Liver adiposity’ change to either liver fat or hepatic steatosis. Please change throughout the manuscript.

Answer: We have now changed “liver adiposity” into “hepatic steatosis”(see rows 31, 125, 126 and 128)

Strength:

1. This is not the first large prospective population based study addressing this question – see Haring et al (2009), Wong et al (2011), Lazo et al (2011), Zhou et al (2012) and Treeprasertsuk (2012). This is the first northern european study to address this question. Its uniqueness is the duration of follow-up, which far exceeds the previous studies.

Answer: We have now changed “Study seems to be the first follow-up study with a large population-based study group and a very long follow-up time”

into

“This is a follow-up study with a large population-based study group and a very long follow-up time”(see row 59)

Introduction:

1. For the prevalence of NAFLD use an original article rather than a review, also specifically use a population-based study – in keeping with your own work (ie Wong GUT et al, Armstrong J Hepatol et al)

Answer: We have now cited “Armstrong et al. Presence and severity of non-alcoholic fatty liver disease in a large prospective primary care cohort J Hepatol 2012” (see row 71)

2. Last sentence of 1st paragraph is wordy – break up into two sentences.

Answer: This sentence is now divided into 2 sentences (see row 84)

3. There is no mention in the introduction of the association between NAFLD and stroke – if the literature is lacking please state or otherwise provide a reference

Answer: Sentence concerning this issue has now been added (see row 101-102)

4. There also needs to be a lot more mention of previous CVD studies and NAFLD ie Haring et al (2009), Wong et al (2011), Lazo et al (2011), Zhou et al (2012) and Treeprasertsuk (2012). Stating what they showed – which was independent risk, but also pointing out their limitations which was duration of follow-up was less than 10 years in most cases. Also most of these studies don't discuss stroke.

Answer: New chapter has been added to introduction (Please see rows 87-92, See also discussion rows 284-285)

Methods:

1. Characterise the radiologists experience (ie approximately how many scans per year etc)

Answer: When OPERA-cohort was designed, Markku Päivänsalo had already had 10 years' experience in abdominal ultrasound examination (see row 127)

2. Has there severity staging of liver fat accumulation on USS been validated locally with biopsy or MRS in other studies? If so please mention, as this is crucial.

Answer: It is true that in previous publications US-diagnosed fatty liver has usually been graded into 2 groups (i.e Lazo et al 2012) Originally, our patients were graded into 3 goups according to liver brightness (non-fatty liver, moderate fat accumulation, severe fat accumulation) by one trained radiologist with 10 years' experience in abdominal ultrasound examinations. We have performed all our statistical analyses when our study subjects were graded into 2 groups ("non-fatty liver" vs "fatty liver", where we joined "moderate liver fat accumulation" and "severe liver fat accumulation" into" fatty liver") and when our subjects were graded into 3 groups ("non-fatty liver"," moderate fat accumulation"," severe fat accumulation"). We noticed, that when the liver brightness was graded into 2 groups, "fatty liver" described the results of subjects with "severe liver brightness". Therefore we decided to use 3 groups in the final analyses. We think that it is more informative to the reader that "severe liver fat accumulation" predicts the future risk for cardiovascular event instead of "fatty liver" in general.

3. Were other causes of Liver disease/steatosis ruled out at baseline (i.e. HBV, HCV). This again is very important. If not please state what the prevalences are in Finland, esp with regards to HBV/HCV (this can placed in the discussion)

Answer: At the baseline, there were 15 patients with hepatic disease (viral, toxins etc.) Excluding these subjects did not have any effect on the results (please see row 254)

4. A huge strength of this study is the characterization of insulin sensitivity (this should be highlighted in the strengths)

Answer: Quantitative insulin sensitivity check index (QUICKI) has been used to determine insulin sensitivity and it is true that QUICKI is a good method (Katz et al. 2000)

5. I presume the bloods were, but please mention 'fasting' if this were so.

Answer: Yes, all samples were obtained after an overnight fast. This statement has been added. (See row 163)

Results:
1. For model 4 please provide actual p-value rather than NS

Answer: Actual p-value has now been added (see row 227)

Discussion:

1. The discussion needs to be a lot more transparent of what data is already in the literature. To the best of my knowledge (if you include Targhers hospital-based studies, which are briefly mentioned) there are 9 prospective studies that have used imaging to define NAFLD and assess the association of CVD development. Please discuss the pros and cons of these studies more, and really emphasize that your patients have extensive follow-up duration, insulin sensitivity and central adiposity (ie waist measurements, which are missing throughout the literature) characterization

Answer: A new chapter has been written into discussion (see row 267-273)

2. Would the others be willing to look for how much incident type 2 diabetes occurred in their study, as very few studies have successfully ruled out T2DM at baseline when looking at NAFLD as an independent predictor of T2DM.

Answer: Thank you for this relevant comment. At the baseline, there were 86 subjects with T2DM. We will have also follow-up data about the prevalence of T2DM and this would be very interesting issue in future.

3. The limitation is not a 'may be,' it is a fact. USS is limited at detecting hepatic steatosis < 30%, and ideally MRS should be used, but that doesn't come without significant cost. To strengthen the accuracy of the liver USS for detecting fat, the authors could calculate the kotronen score (validated by MRS) and see if this correlates at baseline with their USS scoring. They have all of the components at baseline to calculate this.

Answer: We have now changed "may be" into "is" (see row 317). Unfortunately we do not have baseline fS-aspartate aminotransferase (AST) measurements available and therefore we are not able to calculate this index (Kotronen et al 2009)

4. You need to mention MRS, but state that this would have been costly and at the time of study design was likely not available.

Answer: Statement has now been added into text (see rows 322 and 325)

5. Adding more weight to your radiologist experience would be useful here.

Answer: Please see row 127: "one trained radiologist with 10 years' experience in abdominal ultrasound examinations"has been added into methods-section.

Very topical and relevant to the field. Overall, I commend the authors on a well designed study and analysis. The real strengths of this study are the long follow-up (the longest in the literature) and the detailed metabolic characterization at baseline (ie QUICKI, waist measurements, alcohol, smoking – which previous studies have not uniformly mentioned). Prior to publication however the above points need addressing, but overall this is a very useful addition to the field.

Please ensure 'NAFLD' or 'nonalcoholic fatty liver disease' is in your key words to aid with identifying this study on pubmed.

Answer: “fatty liver” has now been added into keywords.

Reviewer: Giulio Marchesini

Discuss the negative findings

Limits of the US imaging technique

The study addresses a debated question, i.e. the role of liver fat in the determination of cardiovascular risk, by measuring the events occurring in a 18-year follow-up of the OPERA cohort in Finland. The severity of adiposity was measured by ultrasounds and graded as 0 (absent, 1 (mild-moderate) and 2 (severe). The cohort included subjects with hypertension and matched controls. The CV risk associated with the presence of fatty liver was highly significant, after correction for multiple confounders, but disappeared after adjustment for insulin resistance.
The authors correctly identify the strengths (long follow-up, well-defined cohort, the use of ICD-9 codes) and the limits (ultrasonographic assessment of fatty liver) of the study.

Problems

1. Ultrasonographic assessment remains the crucial problem. Quantitative measurement off liver fat by US is subject to several limitations dictated by old technology (1991) and scarce reproducibility also in expert hands. The 2-level classification is probably more sensitive than a 3-level classification (mild, moderate, severe), but also in this case a considerable uncertainty is expected. We eagerly need studies based on more solid measurements (NMR).

Answer: Thank you for the relevant comment. It is true that in previous publications US-diagnosed fatty liver has usually been graded into 2 groups (i.e Lazo et al 2012) Originally, our patients were graded into 3 goups according to liver brightness (non-fatty liver, moderate fat accumulation, severe fat accumulation) by one trained radiologist with 10 years’ experience in abdominal ultrasound examinations. We have performed all our statistical analyses when our study subjects were graded into 2 groups (“non-fatty liver” vs “fatty liver”, where we joined “moderate liver fat accumulation” and “severe liver fat accumulation” into” fatty liver”) and when our subjects were graded into 3 groups (“non-fatty liver”,” moderate fat accumulation”,” severe fat accumulation”). We noticed, that when the liver brightness was graded into 2 groups, “fatty liver” described the results of subjects with “severe liver brightness”. Therefore we decided to use 3 groups in the final analyses. We think that it is more informative to the reader that “severe liver fat accumulation” predicts the future risk for cardiovascular event instead of “fatty liver” in general.

It is true that today magnetic resonance spectroscopy is regarded as the best method for the quantification of liver fat. Magnetic resonance spectroscopy was not available at the baseline (please see rows 321-325).

2. The authors should report the inter-/intra-operator reproducibility of the US assessment, based on the techniques and instruments available in 1990. No data are reported in ref 16-17 they quote.

Answer: Our analyses were performed by one trained radiologist with 10 years’ experience in abdominal ultrasound examination, which means that inter-/intra-operator reproducibility can be regarded as good as possible.

3. Although there is evidence that fatty liver may exacerbate the risk of CV outcomes, not all studies are along this line. The authors are asked to discuss the negative findings (see, Ghouri,

HEPATOLOGY 2010;52:1156-1161)

Answer: Discussion has now been widened. A statement concerning this issue has been added (see row 268 and 300).

4. Surprisingly, study group did not predict outcome. This means that being hypertensive does not increase the risk. A comment is needed. It would also be important to see whether the prediction is maintained in the two different cohorts of hypertensive and control subjects.

Answer: This is a very relevant comment. It is known that hypertension is risk factor for cardiovascular disease. It would have been good idea to perform analyses in both groups separately but the number of events in these groups would have been too small to perform all analyses separately. Larger number of subjects and longer follow-up time may be needed to validate these results.

5. CV events in a 19-year follow up are predicted by QUICKI, a very crude assessment of insulin sensitivity. This means that insulin levels were available. It would be interesting to know whether liver enzymes were also present in the dataset and their predictive ability.

Answer: Liver enzymes (ALT and GGT) were available at the baseline. GGT and ALT concentrations at baseline were divided into tertiles (mean concentrations GGT [IU/L]: tertile 1: 18.4, tertile 2: 30.7, tertile 3: 91.0, ALT [IU/L]: tertile 1: 17.3, tertile 2: 27.0, tertile 3: 53.5) Higher concentration of GGT or ALT predicted the future risk for cardiovascular event. After adjusting for age, gender and study group, subjects with highest level of GGT or ALT had higher risk for future CVD event compared to those with lower levels (data not shown). Ultrasound-diagnosed fatty liver was stronger predictor of future risk for CVD event compared to GGT or ALT level when adjusted for age, sex and study group (data not shown).We decided not to publish these results to keep article easier to the reader.

6. It is nonetheless surprising that such a crude measure may predict (or cancel the prediction) of CV events.

Answer:The significance disappeared after adjustment of almost any marker of insulin resistance (2h OGTT, fasting glucose, HOMA etc…)

Minor problems

1. I would suggest change the term “accumulation” with “content”. Accumulation refers to a dynamic process, whereas in this case we only have a static measurement of an imaging surrogate of the liver fat “content”.

Answer: We have now changed “accumulation” into”content” through the manuscript

Reviewer: Masahide Hamaguchi

First of all, authors should describe fetal and non-fetal events in each group. Authors mentioned how they counted CVD using ICD-8/9 or 10 at the method section. However, they didn’t describe the results.

Authors diagnose fatty liver and graded it into 3 grades (0, 1, 2). They referred reference 16 and 17. However, neither reference 16 nor 17 is paper which validate this grading system. In the current paper, only severe fatty liver was a statistically significant factor. The association of moderate fatty liver and CVD was not statistically significant. Thus, the validity of grading system is a critical factor for this study.

In method section, authors mentioned how they followed study participants. However, I felt a difficulty to understand. For example, how many people were suffered fetal events and non-fetal events? Were all of participants who were free from CVD followed until December 31, 2009. Or some participants might be dropped out from the follow ups. I guess flow chart might be useful to understand this study.

This manuscript has provided important evidence that fatty liver predicted future cardiovascular events, but it is depend on traditional metabolic risk factors. This study had actually long follow up time. However, this study included some critical problems as listed below. Thus, I guess this manuscript might be required a major revision.

1. First of all, authors should describe the details of fetal and non-fetal events in each group. Authors mentioned how they counted CVD using ICD-8/9 or 10 at the method section. However, they didn't describe the results.

Answer: This data is in the results-section, where we added some information according to this comment (Please see rows 210-215).This data is available also in Table 2 and Figure legend.

2. Authors diagnosed fatty liver and graded them into 3 grades (0, 1, 2). They referred reference 16 and 17. However, neither reference 16 nor 17 is the paper that validates this grading system directly. In the current paper, only severe fatty liver was a statistically significant factor. The association of moderate fatty liver and CVD was not statistically significant. Thus, the validity of grading system for fatty liver is a critical factor for this study.

Answer: Thank you for the relevant comment. It is true that in previous publications US-diagnosed fatty liver has usually been graded into 2 groups (i.e Lazo et al 2012) Originally, our patients were graded into 3 goups according to liver brightness (non-fatty liver, moderate fat accumulation, severe fat accumulation) by one trained radiologist with 10 years' experience in abdominal ultrasound examinations. We have performed all our statistical analyses when our study subjects were graded into 2 groups ("non-fatty liver" vs "fatty liver", where we joined "moderate liver fat accumulation" and "severe liver fat accumulation" into" fatty liver") and when our subjects were graded into 3 groups ("non-fatty liver"," moderate fat accumulation"," severe fat accumulation"). We noticed, that when the liver brightness was graded into 2 groups, "fatty liver" described the results of subjects with "severe liver brightness". Therefore we decided to use 3 groups in the final analyses. We think that it is more informative to the reader that "severe liver fat accumulation" predicts the future risk for cardiovascular event instead of "fatty liver" in general.

3. In the paper, authors didn't mention about the metabolic syndrome. It's so curious. The association between fatty liver and metabolic syndrome, or CVD and metabolic syndrome has been well known. Authors should discuss these relationships.

Answer: The role of metabolic syndrome in predicting cardiovascular events in this OPERA-data is published very recently. (Santaniemi et al. Metabolic syndrome in the prediction of cardiovascular events: The potential additive role of hsCRP and adiponectin. Eur J Prev Cardiol. 2013 Jun 20)

Specific comments

4. In introduction section, authors mentioned about NAFLD and NASH. However, the study subjects included participants who consume alcohol regularly. I'm afraid this introduction could mislead readers.

Answer: We have tried to take account this issue. When we speak about our study (subjects, results

etc.) we always use term “fatty liver” instead of “non-alcoholic fatty liver disease” to avoid misleading readers.

5. In method section, authors mentioned how they followed study participants. However, I felt a difficulty to understand. For example, how many people were suffered fetal events and non-fetal events? Were all of participants who were free from CVD followed until December 31, 2009. Or some participants might be dropped out from the follow ups. I guess flow chart might be useful to understand this study.

Answer: The exact data about non-fatal and fatal events are in the results-section (please see rows 210-215). In the Cox regression analyses the follow-up time ended December 31, 2009 or whenever the first event occurred. It means that if the subject was free from CVD event the, follow-up ended December 31, 2009.

6. Author used “present tense” at conclusion. I think it is too strong. I guess “past tense” might be suitable for a conclusion of original paper.

Answer: Conclusion-section is now modified into “past tense” (please see rows 332-337).

VERSION 2 – REVIEW

REVIEWER	Dr Matthew Armstrong MBChB MRCP Wellcome Trust Clinical Research Fellow Registrar in Hepatology University of Birmingham
REVIEW RETURNED	12-Aug-2013

GENERAL COMMENTS	The authors may consider a line specifically to highlight the flaws with the Lazo paper. In that, they characterised 'normal' as people with no liver fat or mild liver fat. i.e. the lazo paper may have been comparing like-with-like to a certain degree, and hence NO increased CVS risk in their cohort. Otherwise v.good paper

REVIEWER	Giulio Marchesini, Head Unit of Metabolic Diseases & Clinical Dietetics University of Bologna, Italy NO competing interests in relation to the material presented in the article
REVIEW RETURNED	18-Aug-2013

GENERAL COMMENTS	I remain not at all satisfied from the answer to one question, i.e., the reproducibility of US technique with the equipments available in the early '90s. No data of intra-observer reproducibility presented.

REVIEWER	Masahide Hamaguchi, Department of Experimental Immunology, World Premier International Immunology Frontier Research
REVIEW RETURNED	29-Aug-2013

GENERAL COMMENTS	It is a critical point for epidemiological study to use suitable,

	standardized and validated method including diagnostic criteria and diagnostic procedure. In this study, the authors separated the subject according to the grade of liver brightness. In this case, it is a critical point how the authors diagnosed liver brightness. However, the authors describe neither the diagnostic criteria they used nor the accuracy of their ultrasonographic diagnosis for fatty liver. At least, the authors must use a standardized method for ultrasonographic diagnosis of NAFLD like as Huang RC et al used (hepatology 2013). When they didn't use a standardized method, the authors must present the accuracy of their methods. For example, they should compare their ultrasonographic diagnosis to the results of liver biopsy. At least, the authors must present intra-observer reliability. However, they performed none of above mentioned validation for their methods. In addition to that, the results of present study are different from previous reported evidences. In the present study, the authors claimed that severe fatty liver could predict future CVD but mild fatty liver was not able to predict it. However, previous studies reported that fatty liver including mild fatty liver could predict future CVD like as Hamaguchi M et al reported (World Journal of Gastroenterology 2007). This discrepancy might be based on the difference of diagnostic accuracy for NAFLD. Thus, the authors must present the reliability and accuracy of their diagnostic method for fatty liver. The results and conclusion are not reliable because the authors didn't use reliable methods. I wrote the reason at above section. The authors used ICD and counted the number of fetal and non-fetal coronary heart disease and strokes. However, they listed only total number of cardio vascular events. I believe that it is not enough. It is important to list the detail of main outcomes. This information helps general reader to evaluate the generalizability and validity of the study. When other researcher will use the outcome of this follow up study for meta-analysis, the information is necessary. The authors should describe the types and number of cardio vascular events like as follows; Grade of liver brightness 0 1 2 P (0 vs. 1) P (0 vs. 2) P (0 vs. 3) Number 720 124 144 Non-fetal CVD Total %, Number %, Number %, Number CHD %, Number %, Number %, Number CABG %, Number %, Number %, Number Angioplasty %, Number %, Number %, Number Stroke %, Number %, Number %, Number Fetal CVD %, Number %, Number %, Number Total %, Number %, Number %, Number CHD %, Number %, Number %, Number Stroke %, Number %, Number %, Number

VERSION 2 – AUTHOR RESPONSE

Reviewer: Dr Matthew Armstrong

The authors may consider a line specifically to highlight the flaws with the Lazo paper. In that, they characterised 'normal' as people with no liver fat or mild liver fat. i.e. the lazo paper may have been comparing like-with-like to a certain degree, and hence NO increased CVS risk in their cohort.

Response: Thank you for the relevant comment. One line has now been added (please see rows 268-271) “A previous large population-based prospective cohort study found no association between NAFLD and CVD, however they categorized the degree of steatosis as a two level variable: none to mild and moderate to severe.”

Reviewer: Giulio Marchesini,

NO competing interests in relation to the material presented in the article

Response: Authors report no conflict of interests. This is mentioned in the disclosure summary.

I remain not at all satisfied from the answer to one question, i.e., the reproducibility of US technique with the equipments available in the early '90s. No data of intra-observer reproducibility presented.

Response: Ultrasonography for investigation of the liver echogenicity was performed by one radiologist with a ten year experience among abdominal ulrasound examinations. According to previous study of Yajima et al. in the early 80s (which was used as a reference in our study), when fatty change of over 30% in the hepatic lobule was adopted as the definition of fatty liver, the satisfaction of both liver-kidney contrast and vascular blurring presented an ultrasound diagnostic criterion for fatty liver, with sensitivity of 83%, specificity of 100%, and an accuracy of 96% (1). Recently published study of 235 ultrasound-diagnosed NAFLD patients found a high specificity of 96% but a low sensitivity of 67%. When patients with less than 30% hepatic steatosis were excluded from the study, specificity increased to 93% and sensitivity to 91% (2).

Yajima et al used 3.5 Mhz Toshiba transducer in the early 80s. Palmentieri et al used 3-6Mhz Toshiba transducer recently. In our study, the abdominal ultrasound examination was carried out using a Toshiba SSA 270 ultrasound system with a scanning frequency of 5 Mhz. It can be stated that the specificity and sensitivity of the US technique with the equipment available in the early '90s were rather good and have not significantly improved.

Normal liver parenchyma should be slightly more echogenic (brighter) than the kidney parenchyma. In a case of increased liver echogenicity an ultrasound diagnosis of bright liver was settled (1). With a ten year experience the radiologist classified increased echogenicity subjectively as a slight or a clear bright liver finding. A reference (no 22) and sentence concerning this issue has now been added (please see rows 126-130).

Among OPERA study subjects, the prevalence of moderate liver fat accumulation was 13.1% and that of severe fat accumulation 14.2%, which is in line with the overall prevalence of NAFLD in the Western world (3).

Intra-observer reproducibility was not investigated and this may be one shortcoming of this study. It should be noted that the liver brightness classification was performed by one single radiologist with a then year experience among ultrasound examination.

1) Yajima, Y., Ohta, K., Narui, T., Abe, R., Suzuki, H. and Ohtsuki, M. (1983) Ultrasonographical

Diagnosis of Fatty Liver; Significance of the Liver-Kidney Contrast. Tohoku J. exp. Med. 139 (1), 43-50
2) Palmentieri B, de Sio I, La Mura V, Masarone M, Vecchione R, Bruno S, Torella R &Persico M (2006) The role of bright liver echo pattern on ultrasound B-mode examination in the diagnosis of liver steatosis. Dig Liver Dis 38(7): 485–489.
3) Vernon G, Baranova A & Younossi ZM (2011) Systematic review: the epidemiology and natural history of non-alcoholic fatty liver disease and non-alcoholic steatohepatitis in adults. Aliment Pharmacol Ther 34(3): 274–285.

Reviewer: Masahide Hamaguchi

It is a critical point for epidemiological study to use suitable, standardized and validated method including diagnostic criteria and diagnostic procedure. In this study, the authors separated the subject according to the grade of liver brightness. In this case, it is a critical point how the authors diagnosed liver brightness. However, the authors describe neither the diagnostic criteria they used nor the accuracy of their ultrasonographic diagnosis for fatty liver. At least, the authors must use a standardized method for ultrasonographic diagnosis of NAFLD like as Huang RC et al used (hepatology 2013). When they didn't use a standardized method, the authors must present the accuracy of their methods. For example, they should compare their ultrasonographic diagnosis to the results of liver biopsy. At least, the authors must present intra-observer reliability. However, they performed none of above mentioned validation for their methods.
In addition to that, the results of present study are different from previous reported evidences.
In the present study, the authors claimed that severe fatty liver could predict future CVD but mild fatty liver was not able to predict it. However, previous studies reported that fatty liver including mild fatty liver could predict future CVD like as Hamaguchi M et al reported (World Journal of Gastroenterology 2007). This discrepancy might be based on the difference of diagnostic accuracy for NAFLD.
Thus, the authors must present the reliability and accuracy of their diagnostic method for fatty liver.

Response: Ultrasonography for investigation of the liver echogenicity was performed by one radiologist with a ten year experience among abdominal ulrasound examinations. According to previous study of Yajima et al. in the early 80s (which was used as a reference in our study), when fatty change of over 30% in the hepatic lobule was adopted as the definition of fatty liver, the satisfaction of both liver-kidney contrast and vascular blurring presented an ultrasound diagnostic criterion for fatty liver, with sensitivity of 83%, specificity of 100%, and an accuracy of 96% (1). Recently published study of 235 ultrasound-diagnosed NAFLD patients found a high specificity of 96% but a low sensitivity of 67%. When patients with less than 30% hepatic steatosis were excluded from the study, specificity increased to 93% and sensitivity to 91% (2).

Yajima et al used 3.5 Mhz Toshiba transducer in the early 80s. Palmentieri et al used 3-6Mhz Toshiba transducer recently. In our study, the abdominal ultrasound examination was carried out using a Toshiba SSA 270 ultrasound system with a scanning frequency of 5 Mhz. It can be stated that the specificity and sensitivity of the US technique with the equipment available in the early '90s were rather good and have not significantly improved.

Normal liver parenchyma should be slightly more echogenic (brighter) than the kidney parenchyma. In a case of increased liver echogenicity an ultrasound diagnosis of bright liver was settled (1). With a ten year experience the radiologist classified increased echogenicity subjectively as a slight or a clear bright liver finding. A reference (no 22) and sentence concerning this issue has now been added (please see rows 126-130).

Among OPERA study subjects, the prevalence of moderate liver fat accumulation was 13.1% and that

of severe fat accumulation 14.2%, which is in line with the overall prevalence of NAFLD in the Western world (3).

Intra-observer reproducibility was not investigated and this may be one shortcoming of this study. It should be noted that the liver brightness classification was performed by one single radiologist with a then year experience among ultrasound examination.

1) Yajima, Y., Ohta, K., Narui, T., Abe, R., Suzuki, H. and Ohtsuki, M. (1983) Ultrasonographical Diagnosis of Fatty Liver; Significance of the Liver-Kidney Contrast. Tohoku J. exp. Med. 139 (1), 43-50
2) Palmentieri B, de Sio I, La Mura V, Masarone M, Vecchione R, Bruno S, Torella R &Persico M (2006) The role of bright liver echo pattern on ultrasound B-mode examination in the diagnosis of liver steatosis. Dig Liver Dis 38(7): 485–489.
3) Vernon G, Baranova A & Younossi ZM (2011) Systematic review: the epidemiology and natural history of non-alcoholic fatty liver disease and non-alcoholic steatohepatitis in adults. Aliment Pharmacol Ther 34(3): 274–285.

The authors used ICD and counted the number of fetal and non-fetal coronary heart disease and strokes. However, they listed only total number of cardio vascular events. I believe that it is not enough. It is important to list the detail of main outcomes. This information helps general reader to evaluate the generalizability and validity of the study. When other researcher will use the outcome of this follow up study for meta-analysis, the information is necessary. The authors should describe the types and number of cardio vascular events.

Response: Thank you for the relevant comment. New table has now been added (please see page 27, Table 3).

VERSION 3 - REVIEW

REVIEWER	Masahide Hamaguchi Osaka University, Japan
REVIEW RETURNED	12-Feb-2014

GENERAL COMMENTS	The manuscript has been improved, but the authors addressed my comments partially. So some problems has been still remains. 1. The method for diagnosing NAFLD by ultrasonography. The authors diagnosed NAFLD by ultrasonography and graded into 3 levels. The authors referred their previous paper (ref 24). However, they didn't validate their method for diagnosing NAFLD in the previous paper, as well as in the current paper. Thus, I couldn't consider that the authors used a validated and a standardized method for diagnosing NAFLD. Authors should write it as the limitation of the study. The authors claimed the generalities of ultrasonographic diagnosis for NAFLD in the rebuttal letter. However, the accuracy of their method was not confirmed by the general accuracy of ultrasonographic diagnosis of NAFLD. Moreover, the intra-observer reproducibility was not investigated totally. They should also mention it as the limitation of the study. If trained radiologist with 10 years' experience diagnosed NAFLD, it was not a direct answer for the intra-observer reproducibility.

	2. The authors added a new table (page 27, Table 3). I believe that it improved this paper. However, they didn’t explain their statistic method in this table. Please write the explanation of statistic method in the table legend.

VERSION 3 – AUTHOR RESPONSE

Reviewer Name Masahide Hamaguchi
Institution and Country Osaka University, Japan

1. The method for diagnosing NAFLD by ultrasonography.
The authors diagnosed NAFLD by ultrasonography and graded into 3 levels. The authors referred their previous paper (ref 24). However, they didn't validate their method for diagnosing NAFLD in the previous paper, as well as in the current paper. Thus, I couldn't consider that the authors used a validated and a standardized method for diagnosing NAFLD. Authors should write it as the limitation of the study.
The authors claimed the generalities of ultrasonographic diagnosis for NAFLD in the rebuttal letter. However, the accuracy of their method was not confirmed by the general accuracy of ultrasonographic diagnosis of NAFLD.
Moreover, the intra-observer reproducibility was not investigated totally. They should also mention it as the limitation of the study. If trained radiologist with 10 years' experience diagnosed NAFLD, it was not a direct answer for the intra-observer reproducibility.

Response: Thank you for the relevant comment. New sentence has now been added (Please see rows 327-330)

2. The authors added a new table (page 27, Table 3). I believe that it improved this paper. However, they didn't explain their statistic method in this table. Please write the explanation of statistic method in the table legend.

Response: New sentence concerning this issue has been added (Please see Table 3 legend)