The FLIP partners’ consortium

Vlad Ratziu, Thierry Poynard, Assistance Publique Hôpitaux de Paris, Groupe Hospitalier Pitie Salpetriere, Universite Pierre et Marie Curie

Biopredictive SAS, Jean-Marie Castille, jean-marie.castille@biopredictive.com, Yen Ngo

Alma Consulting Group SAS, Tania Langon, tlangon@almacg.com

University of Newcastle, Chris Day, c.p.day@newcastle.ac.uk

University of Bristol, Debbie Lawlor, d.a.lawlor@bristol.ac.uk

Università di Bologna, Giulio Marchesini, giulio.marchesini@unibo.it

University of Firenze, Fabio Marra, f.marra@DMI.unifi.it

Università degli studi di torino, Elisabetta Bugianesi, ebugianesi@yahoo.it

Universita di Modena e Reggio Emilia, Stefano Bellentani, bellentanistefano@gmail.com

Université de Bern, Jean-François Dufour, jf.dufour@ikp.unibe.ch

Servicio Andaluz de Salud, Manuel Romero Gomez, manual.romero.sspa@juntadeandalucia.es

Bispebjerg hospital, Region Hovedstaden, Thorkild Sorensen, TIAS@ipm.regionh.dk

Fondazione Italiana Fegato - Fonlus, Claudio Tribelli, etliver@csf.units.it

Universita di Ancona, Samuele De Minicis, s.de.minicis@yahoo.it

Medizinischen Universitaet Wien, Michael Trauner, michael.trauner@meduniwien.ac.at

The FLIP Pathology Consortium:

Pierre Bedossa, Assistance Publique-Hopitaux de Paris, hopital Beaujon, University Paris-Diderot, Paris, France. Alastair D. Burt, School of Medicine, University of Adelaide, Adelaide, Australia. Annette S.H. Gouw, Dept. of Pathology & Medical Biology, University Medical Center Groningen, Groningen, The Netherlands. Carolin Lackner, Institute of Pathology, Medical University of Graz, Graz, Austria. Peter Schirmacher, Institut Fur Pathologie, Universitätsklinikum Heidelberg, Heidelberg, Germany. Luigi Terracciano, Institute of Pathology, University Hospital Basel, Switzerland. D. Tiniakos, Medical School, National and Kapodistrian University of Athens, Greece and Medical School, Newcastle University, Newcastle-upon-Tyne, United Kingdom. J. Brain, Medical School, Newcastle University, Newcastle-upon-Tyne, United Kingdom. Yvonne Bury, Royal Victoria Infirmary, Hospitals Foundation Trust, Newcastle upon Tyne, United Kingdom. Daniela Cabibi, Universita di Palermo, Palermo, Italy. Frederic Charlotte, Assistance Publique-Hopitaux de Paris, Groupe Hospitalier Pitie-Salpetriere, Universite Pierre et Marie Curie, Paris, France. Ezio David, University of Torino, Ospedale Molinette, Torino, Italy. Luisa Losi, Modena University Hospital, Modena, Italy. Matteo Montani, Institute of Pathology, University of Bern, CH-3010 Bern, Switzerland. Maria Jesus Pareja, Hospital Universitario de Valme, Sevilla, Spain. Dominique Wendam, Assistance Publique-Hopitaux de Paris, Hopital St Antoine Universite Pierre et Marie Curie. Fritz Wrba, Medical University of Vienna, Vienna, Austria. Marianne Ziol, Assistance Publique-Hopitaux de Paris, Hopital Jean Verdier, Universit_e Paris 13, Bobigny, France. Vlad Ratziu, Assistance Publique Hopitaux de Paris, Hopital Pitie Salpetriere, Universite Pierre et Marie Curie, Paris, France.

The FibroFrance-CPAM group:


The FibroFrance-Obese group:

Thierry Poynard, Hugo Perazzo, Frederic Charlotte and Vlad Ratziu, APHP UPMC Liver Center, Paris, France. Guillaume Lassailly, Robert Cauazzo, Francois Pattou and Philippe Mathurin, Hepatology Department, Centre Hospitalier Lille, Lille, France.
In ActiTest, the possible choices of predetermined cutoffs were those previously [19] and based on 3 levels, as follows: Level 1 = 0.52 METAVIR A2, the Standard for chronic hepatitis C and B; Level 2 = 0.29 METAVIR A1; and Level 3 = 0.17 METAVIR A0-A1. For SteatoTest, the four possible choices were predetermined cutoffs for SAF-S1, previously published [11,12], as follows: Level 1 = 0.57 S1 ≥ 5%; Level 2 = 0.48 S1 > 0%; Level 3 = 0.38 S0-S1 ≥ 5%; and Level 4 = 0.30 S0-S1 > 0%.

The Liver-FibroSTARD checklist summarizes the important information that must be present in the manuscripts of diagnostic studies on non-invasive tools for liver fibrosis evaluation. Compared to STARD, the Liver-FibroSTARD checklist includes 2 additional items (#12 and #26) and 44 sub-items. The sub-items correspond to those proposals that clearly depicted, within the items, each of the particular features of diagnostic studies on liver fibrosis tests. Finally, Liver-FibroSTARD presents as a complementary module of the STARD checklist.

Some items or sub/items include several criteria; major criteria are indicated by an asterisk (*). Example: item #3: “The study population: The inclusion and exclusion criteria*, setting, and locations* where data were collected”. If a major item is missing, the corresponding criterion has to be rated absent. Some items/sub-items (#12.1 and #23.1, #13.10 and #22.2) are redundant since they can be found in different locations of the article.

---

| TITLE/ABSTRACT /KEYWORDS | 1. Identify the article as a study of diagnostic accuracy (recommend MeSH heading “sensitivity and specificity”).
  | 1.1. Identify the article, especially in the title, as a study of the diagnostic performance of liver fibrosis/cirrhosis biomarker(s)/test(s).
  | 1.2. Recommended key words (choose the most appropriate): “liver fibrosis”, “cirrhosis”, “diagnosis”, “biomarker”, “diagnostic test”, “noninvasive diagnosis”.
  |
| INTRODUCTION | 2. State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups.
  | In study aims, specify:
  | 2.1. If the aim is to identify new marker(s)/develop new test(s), or to evaluate published marker(s)/test(s).
  | 2.2. Whether the study is performed in a single or multiple cause(s) of chronic liver disease.
  | 2.3. The reference used for fibrosis diagnosis in the study.
  | 2.4. The diagnostic target used as the primary aim of the study and, if appropriate, other diagnostic targets used as secondary aims.
  |
| METHODS | Describe:
  |
| Participants | 3. The study population: The inclusion and exclusion criteria*, setting, and locations* where data were collected.
  |
| 4. Participant recruitment: Was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?
  | 4.1. State if healthy subjects without chronic liver disease are included or not in the study.
  | 4.2. State if patients were selected by one abnormal or several discordant fibrosis test(s).
  | 4.3. State if patients were selected according to the availability of reference or index test(s) result(s).
  |
| 5. Participant sampling: Was the study population a consecutive series of participants defined by the selection criteria in item 3 and 4? If not, specify how participants were further selected.
  |
| 6. Data collection: Was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?
  | 6.1. The chronology between patient inclusion*, data collection (reference/index tests)*, and data analysis is well described.
  | 6.2. Has the study population been previously used/published for the evaluation of the studied fibrosis test(s)?
  |
7. **The reference standard and its rationale.**

8. **Technical specifications of material and methods involved including how and when measurements were taken, and/or cite references for index tests and reference standard.**

   For the reference and index test(s), specify characteristics with sufficient detail to permit exact reoperation, when appropriate:

   - **8.1. Center:** standardization of procedures across centers.
   - **8.2. Patient:** fasting conditions*, time, posture, etc. (give information about the influence of conditions on the intra-individual variability).
   - **8.3. Delay:** time interval between reference and index test(s).
   - **8.4. Material:** technical specifications (name, generation, manufacturer, instrument), method of measurement, applicability (failure/reliability criteria)*. Specifically for liver biopsy, indicate material used per center, i.e. percutaneous/transjugular/other, needle diameter.
   - **8.5. Biological samples:** description of method of collection, transport, storage*.
   - **8.6. Specify how the index tests were calculated.**
   - **8.7. Specify how the risk for false negative/positive results was taken into account.**
   - **8.8. How sample bias was limited:** minimal biopsy size (length)*, number of portal tracts required, number of fragments.
   - **8.10. Scoring system used (Metavir, Ishak, Scheuer, etc.).**

9. **Definition of and rationale for the units, cut-offs*, and/or categories of the results of the index tests and the reference standard.**

10. **The number*, training and expertise* of the persons executing and reading the index tests and the reference standard.**

11. **Whether or not the readers of the index tests and reference standard were blind (masked) to the results of the other test and describe any other clinical information available to the readers.**

12. **State if the study is conducted on an intention-to-diagnose basis or if the analysis is per-protocol (i.e. with exclusion of failed/unreliable fibrosis test(s)/reference measurements).**

   - **12.1. If intention-to-diagnose analysis, specify how failure and unreliable test(s)/reference are taken into account in the analysis.**

13. **Methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g. 95% confidence intervals).**

   - **Specify:**
     - **13.1. Detailed sample size calculation.**
     - **13.2. Statistical methods used to quantify uncertainty (e.g. 95% confidence intervals).**
     - **13.3. Control of multiple comparisons that increases type I error: multiple comparisons of tests (e.g. Bonferroni correction, etc.), multiple diagnostic targets.**
     - **13.4. Method for calculation of fibrosis test(s) diagnostic cut-offs.**
     - **13.5. Method for validation of new test(s) or new calculated diagnostic cut-off(s) (e.g. external validation set, internal validation by bootstrapping, etc.).**
     - **13.6. Method for control of center/operator effect.**
     - **13.7. Method for control of spectrum effect if unrepresentative prevalence of fibrosis stages (e.g. Obuchowski index, DANA, etc.).**
     - **13.8. Method for control of misclassification errors by the reference test.**
     - **13.9. Use of a reference without gold standard.**
     - **13.10. Analysis of discordances between reference/index test(s).**
### RESULTS

**Report:**

#### Participants

15. When study was performed, including beginning and end dates of recruitment.

#### Supplementary-Table-S3

16. Clinical and demographic characteristics of the study population (e.g. age*, sex*, spectrum of presenting symptoms, comorbidity, current treatments, recruitment centers).

16.1. For liver biopsy: size (length)*, number of portal tracts, number of fragments.

16.2. For index test(s): confounding factors that potentially influence the test(s) results (flare-up, inflammation, other liver lesions, intrinsic characteristics, etc.).

#### Supplementary-Table-S14

17. The number of participants satisfying the criteria for inclusion who did or did not undergo the index tests and/or the reference standard*; describe why participants failed to undergo either test (a flow diagram is strongly recommended).

17.1. If per-protocol analysis, report comparisons between patients excluded due to failed/unreliable test(s)/reference and patients with reliable fibrosis test(s)/reference.

#### Test results

18. Time-interval* between the index tests and the reference standard, and any treatment administered between.

#### Supplementary-Table-S3

19. Distribution of severity of disease (define criteria) in those with the target condition*; other diagnoses in participants without the target condition.

19.1. Specify the prevalence* of the diagnostic condition (spectrum effect).

#### Page 12

20. A cross tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard; for continuous results, the distribution of the test results by the results of the reference standard.

20.1. Presentation of contingency tables, box/scatter plots.

21. Any adverse events from performing the index tests or the reference standard.

### Estimates

#### Page 12

22. Estimates of diagnostic accuracy* and measures of statistical uncertainty (e.g. 95% confidence intervals).

22.1. Specify sensitivity* and specificity* with 95% confidence intervals; ROC analysis.

22.2. Analyzing discordances between fibrosis tests(s)/reference.

#### Table S10

23. How indeterminate results, missing data and outliers of the index tests were handled.

23.1. How missing/failure/unreliable results of index test(s)/reference were handled (intention-to-diagnose/per-protocol analysis).

23.2. How outliers of the index tests were handled.

#### Table S11

24. Estimates of variability of diagnostic accuracy between subgroups of participants, readers or centers, if done.

25. Estimates of test reproducibility, if done.


### DISCUSSION

#### Page 14 to 17

27. Discuss the clinical applicability of the study findings.

27.1. Discuss the representativeness of the study sample and recruiting centers (i.e. spectrum effect, etc.).

27.2. Discuss the interpretation of fibrosis test(s) results in clinical practice.

27.3. Discuss the clinical relevance of the study results.

---

* Items 12.1 and 23.1 are redundant but retained since they can be located in different paragraphs within an article

* Items 13.10 and 22.2 are redundant but retained since they can be located in different paragraphs within an article

---

This file is the proprietary of AFEF and can be reproduced without authorization.

**Explanations:** see glossary

**Authors:** ARDENT group (see details in glossary) and AFEF (French Association for the Study of the Liver)

Version: February 2015
The prevalence of NASH using standard definition was 50.8% (47.8-53.8) (549/1081), and using simplified definition, 54.4% (51.4-57.4) (588/1081).

*These 39 NASH cases defined by ballooning+ lobular inflammation stages ≥2, (3.6%;2.6-4.9) that were missed by the FLIP-algorithm included 15 cases with significant fibrosis (6 F2, 5 F3 and 4 cirrhosis).
Figure, Supplemental-Digital-Content-5
Performance of NIT-NASHs versus non-patented tests.

NIT-NASHs had a significantly higher AUROC (0.671;0.614-0.721), than NAFLD-score AUROC (0.570;0.510-0.626;P=0.006), FIB4 (0.528;0.467-0.584;P=0.0003) and BARD index (0.541;0.476-0.599;P=0.003).
Figure, Supplemental-Digital-Content-6:
Performance of NIT-A2orF2 versus non-patented tests.

NIT-A2orF2 had a significantly higher AUROC (0.671;0.613-0.721) than the NAFLD-score (0.570;0.510-0.626; P=0.006), FIB4 (0.528;0.467-0.584;P=0.0003) and BARD (0.541;0.476-0.599;P=0.003)
Table, Supplemental-Digital-Content-7: Characteristics of patients included in the construction populations, with missing non-patented NITs (NAFLD-score, FIB4 and BARD), in comparison with cases with both patented and not-patented NITs.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Patented NITs missing</th>
<th>Patented NITs assessed</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>507 (100%)</td>
<td>574 (100%)</td>
<td>1</td>
</tr>
<tr>
<td>Presumed NAFLD</td>
<td>507 (100%)</td>
<td>574 (100%)</td>
<td>1</td>
</tr>
<tr>
<td>Gender male</td>
<td>128 (25.2%)</td>
<td>359 (62.5%)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Diabetes treated or glucose6.1mmol/L</td>
<td>145 (28.6%)</td>
<td>209 (36.4%)</td>
<td>0.006</td>
</tr>
<tr>
<td>Age (year)</td>
<td>43.2 (41.3-44.7)</td>
<td>53.0 (51.1-54.0)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>BMI (weight/height^2) &gt;= 30</td>
<td>491 (96.8%)</td>
<td>268 (46.7%)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td><strong>Biopsy number</strong></td>
<td><strong>507 (100%)</strong></td>
<td><strong>574 (100%)</strong></td>
<td></td>
</tr>
<tr>
<td>Biopsy length (mm)</td>
<td>12 (11-12)</td>
<td>25 (22-25)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Biopsy-test days</td>
<td>0.0 (0)</td>
<td>0.0 (0.0-0.1)</td>
<td>1</td>
</tr>
<tr>
<td><strong>Stage of fibrosis (SAF F biopsy)</strong></td>
<td></td>
<td>&lt;0.0001</td>
<td></td>
</tr>
<tr>
<td>F0 no fibrosis</td>
<td>237 (46.7%)</td>
<td>117 (20.4%)</td>
<td></td>
</tr>
<tr>
<td>F1 perisinusoidal or portal</td>
<td>214 (42.2%)</td>
<td>173 (30.1%)</td>
<td></td>
</tr>
<tr>
<td>F2 sinusoidal or perportal without bridging</td>
<td>33 (6.5%)</td>
<td>135 (23.5%)</td>
<td></td>
</tr>
<tr>
<td>F3 bridging fibrosis</td>
<td>15 (3.0%)</td>
<td>118 (20.6%)</td>
<td></td>
</tr>
<tr>
<td>F4 cirrhosis</td>
<td>8 (1.6%)</td>
<td>31 (5.4%)</td>
<td></td>
</tr>
<tr>
<td><strong>Ballooning</strong></td>
<td>&lt;0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 0</td>
<td>313 (61.7%)</td>
<td>108 (18.8%)</td>
<td></td>
</tr>
<tr>
<td>Grade 1</td>
<td>124 (24.5%)</td>
<td>240 (41.8%)</td>
<td></td>
</tr>
<tr>
<td>Grade 2</td>
<td>70 (13.8%)</td>
<td>228 (39.4%)</td>
<td></td>
</tr>
<tr>
<td><strong>Lobular inflammation</strong></td>
<td>&lt;0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 0</td>
<td>307 (60.6%)</td>
<td>112 (19.5%)</td>
<td></td>
</tr>
<tr>
<td>Grade 1</td>
<td>170 (33.5%)</td>
<td>317 (55.2%)</td>
<td></td>
</tr>
<tr>
<td>Grade 2</td>
<td>30 (5.9%)</td>
<td>145 (25.3%)</td>
<td></td>
</tr>
<tr>
<td><strong>Grade of activity (SAF A biopsy)</strong></td>
<td>&lt;0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A0 no activity</td>
<td>252 (49.7%)</td>
<td>61 (10.6%)</td>
<td></td>
</tr>
<tr>
<td>A1 mild</td>
<td>101 (19.9%)</td>
<td>79 (13.8%)</td>
<td></td>
</tr>
<tr>
<td>A2 moderate</td>
<td>86 (17.0%)</td>
<td>181 (31.5%)</td>
<td></td>
</tr>
<tr>
<td>A3 severe or A4 very severe</td>
<td>68 (13.4%)</td>
<td>253 (44.1%)</td>
<td></td>
</tr>
<tr>
<td><strong>Grade of steatosis (sensitive)</strong></td>
<td>&lt;0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S0 no steatosis 0%</td>
<td>32 (6.3%)</td>
<td>19 (3.3%)</td>
<td></td>
</tr>
<tr>
<td>S0 1-4%</td>
<td>18 (7.5%)</td>
<td>1 (0.2%)</td>
<td></td>
</tr>
<tr>
<td>S1 mild 5%-100%</td>
<td>437 (86.2%)</td>
<td>554 (96.5%)</td>
<td></td>
</tr>
<tr>
<td>FLIP-algo Steatosis ≥5%</td>
<td></td>
<td>&lt;0.0001</td>
<td></td>
</tr>
<tr>
<td>No-steatosis (&quot;No-NAFLD&quot;)</td>
<td>70 (13.8%)</td>
<td>20 (3.5%)</td>
<td></td>
</tr>
<tr>
<td>Steatosis only</td>
<td>299 (59.0%)</td>
<td>143 (24.9%)</td>
<td></td>
</tr>
<tr>
<td>NASH</td>
<td>138 (27.2%)</td>
<td>411 (71.6%)</td>
<td></td>
</tr>
</tbody>
</table>

1 Cases with histological steatosis (5%) or activity (A>0) were excluded. 2 One case had steatosis 2% and Ballooning and Lobular inflammation grade 1 and therefore classified NASH with FLIP algorithm using 0% cutoff and "no steatosis" using 5% cutoff
In order to describe the construction of NITs, and the impact of definitions on their accuracy, we review the literature to clarify the main definitions of the population of interest (the appropriate context of use was defined as carriers of metabolic risk factor), the definition of the disease of interest (metabolic liver diseases included steatosis, activity and fibrosis [SAF], in the absence of other known liver disease). We screened PUBMED with the following tags: "NAFLD metabolic liver disease biopsy human" (January 5th 2017). The criteria of inclusion were studies in adults, with 500 or more biopsies and giving the definition of histological steatosis.

<table>
<thead>
<tr>
<th>Context of use</th>
<th>Author, year</th>
<th>Number</th>
<th>Prevalence S0 0%-4%</th>
<th>0% Prevalence A0</th>
<th>Prevalence A0S0 S0&lt;5%*</th>
<th>Prevalence S0 S0&lt;5%</th>
<th>Cirrhosis F4</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAFLD</td>
<td>Kleiner, 2005</td>
<td>576</td>
<td>58 (10%)</td>
<td>NA</td>
<td>NA (14%)*</td>
<td>13 (2.2%)</td>
<td>NA</td>
</tr>
<tr>
<td>NAFLD</td>
<td>Brunt, 2011</td>
<td>934</td>
<td>37 (4%)</td>
<td>NA</td>
<td>3 (0.3%)</td>
<td>3 (0.3%)</td>
<td>3 (0.4%)</td>
</tr>
<tr>
<td>NAFLD Obese</td>
<td>Bedossa, 2012</td>
<td>679</td>
<td>158 (23%)</td>
<td>NA</td>
<td>248 (36.5%)</td>
<td>147 (21.6%)</td>
<td>147 (21.6%)</td>
</tr>
<tr>
<td>NAFLD</td>
<td>Kessoku, 2014</td>
<td>1,048</td>
<td>0 (0%)</td>
<td>NA</td>
<td>157 (25.4%)</td>
<td>0 (0%)</td>
<td>NA</td>
</tr>
<tr>
<td>NAFLD</td>
<td>Angulo, 2015</td>
<td>619</td>
<td>0 (0%)</td>
<td>NA</td>
<td>157 (25.4%)</td>
<td>0 (0%)</td>
<td>NA</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>3856</td>
<td>253 (6.6%)</td>
<td>NA</td>
<td>163 (4.2%)</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

NA=not available.

* Lobular inflammation taken if no details overall activity as most sensitive than ballooning for grade 1. 13 out of 575 (2.2%) cases only were A0S0 (see Figure 4 in the Kleiner article). Only 163 out of 3856 (4.2%) were A0S0, which should be the appropriate controls for assessing NITs performance for NASH prediction.

No study detailed the full spectrum of steatosis, including cases without any steatosis (0%). One study included presumed NAFLD cases with steatosis 1-4% without excluding any cases, but did not specify the prevalence of 0% versus 1-4%. One study included presumed NAFLD, only with severe obesity, with steatosis 1-4% without excluding any cases, but did not specify the prevalence of 0% steatosis versus 1-4%. One study included presumed NAFLD cases with steatosis 1-4%, did not specify the prevalence of 0% versus 1-4%, although cases with cirrhosis were excluded. The two last studies excluded cases with steatosis 1-4%. Finally, no large study estimated the prevalence of minimal steatosis (1-4%) among presumed NAFLD and therefore the estimated prevalence of absence of any steatosis.

Biopsy length’s medians were not given in these 5 studies, but mentioned in Angulo study as a confounding factor analyzed in the prognostic analysis.