ONLINE-ONLY SUPPLEMENTARY MATERIAL

METHODS

Data sources and searches
We performed a systematic review and meta-analysis following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (http://www.prisma-statement.org). We systematically searched in PubMed, Scopus, and Web of Science, from the inception date to 1 September 2023, using the terms “NAFLD” OR “non-alcoholic fatty liver disease” OR “non-alcoholic steatohepatitis” OR “NASH” AND “FGF-21 analogues” OR “efruxifermin” OR “pegbelfermin” OR “pegzafermin” IN “humans”, to identify all published phase-2 randomized placebo-controlled trials (RCTs) examining the efficacy of FGF-21 analogues on the histological resolution of NASH without worsening of fibrosis or the improvement in liver fibrosis of one stage or more without worsening of NASH (i.e., the two FDA defined histological endpoints for conditional approval of new drugs for NASH). No language restrictions were imposed. In addition, we reviewed references from relevant original papers and reviews to identify further eligible RCTs not covered by the three large electronic databases.

Study selection
Studies were included if they met the following criteria: (1) phase-2 placebo-controlled RCTs examining the efficacy of FGF-21 analogues on histological resolution of NASH without worsening of liver fibrosis, or ≥1-stage fibrosis improvement without worsening of NASH in adults with biopsy-confirmed NASH and different stages of liver fibrosis; and (2) phase 2 RCTs reporting odds ratios (OR) with 95% confidence intervals for the two histological endpoints of interest. Phase-2 placebo-controlled RCTs examining the efficacy of FGF-21 analogues on hepatic fat content assessed by MRI-proton density fat fraction and non-invasive biomarkers of fibrosis that did not also have histological liver endpoints were excluded from the meta-analysis. Congress abstracts, theses, case reports, reviews, commentaries, editorials, practice guidelines, and observational studies were also excluded from the meta-analysis.

Data extraction and quality assessment
Two investigators (AM and GT) independently examined all titles and abstracts and obtained full texts of potentially eligible studies. Working independently and in duplicate, we analyzed whether these studies met the inclusion criteria mentioned above. Discrepancies were resolved by consensus. For all phase 2 RCTs, we extracted information on study design, sample size, and the number of successes and failures for each histological liver endpoint for both the active-comparator group (using FGF-21 analogues) and the placebo group. For the active-comparator group, in each eligible RCT, we combined treatment effects of variable weekly dosages of FGF-21 analogues into a single group (to avoid including subjects in the placebo group several times in the analysis). We did not contact any corresponding author of the eligible RCTs to obtain further information for the meta-analysis.

Two investigators (AM and GT) independently evaluated the risk of bias for each eligible RCT. For this purpose, we used the Cochrane Collaboration’s tool, which assesses the following seven potential sources of bias: random sequence generation (selection bias), allocation concealment (selection bias), blinding of participants and personnel (performance bias), blinding of outcome assessment (detection bias), incomplete outcome data (attrition bias), selective reporting (reporting bias), and other biases. For each of these domains, we categorized each eligible RCT into three categories: low, unclear, or high risk of bias (Cumpston M et al. Updated guidance for trusted systematic reviews: a

**Data synthesis and analysis**
The odds ratios with their 95% CIs were considered as the effect size for each eligible study. An overall effect size estimate was calculated using the common-effects model (i.e., the Mantel-Haenszel method) or the random-effects model (i.e., the DerSimonian-Laird method). Visual inspection of the forest plots was used to evaluate the possibility of statistical heterogeneity. Statistical heterogeneity was assessed by the $I^2$-statistics, which estimates the percentage of variability across the eligible RCTs due to heterogeneity rather than chance alone. $I^2$ values of about 25% represent low heterogeneity, about 50% represent medium heterogeneity, and about 75% represent high heterogeneity. We also performed multivariate meta-regression models adjusted for the different type and the different weekly dosage of FGF-21 analogue used in each single RCT.

The fragility index (FI) for meta-analysis was also calculated. In particular, the FI was defined as the minimum number of participants from trials included in the meta-analysis for which a modification of the event status (i.e., changing events to non-events or non-events to events) can change the statistical significance of the pooled treatment effect (see Atal I et al. The statistical significance of meta-analyses is frequently fragile: definition of a fragility index for meta-analyses. J Clin Epidemiol. 2019;111:32-40). The larger the FI, the more robust are the trial’s data. Given the small number of RCTs in the meta-analysis (n=5 studies), a formal assessment of the publication bias using the funnel plot was not performable according to the Cochrane guidance for systematic reviews (Cumpston M et al. Updated guidance for trusted systematic reviews: a new edition of the Cochrane Handbook for Systematic Reviews of Interventions. Cochrane Database Syst Rev. 2019;10:ED000142).

All two-sided statistical tests used a significance level of P-value <0.05. We used STATA version 17.0 (StataCorp, College Station, Texas, USA) and its meta-analysis package and R version 4.2.2 for all statistical analyses.
Supplementary Figure 1. PRISMA flow chart of the study selection.

**PRISMA 2009 Flow Diagram**

Records identified through database searching on three large electronic databases, including articles published until September 1st, 2023 (n = 59)

Additional records identified through other sources (n = 0)

Records after duplicates removed (n = 59)

Records screened (n = 59)

Records excluded (n = 51)

Full-text articles assessed for eligibility (n = 8)

Full-text articles excluded, with reasons (n = 3)

Studies included in qualitative synthesis (n = 5)

Studies included in quantitative synthesis (meta-analysis) (n = 5)