

STROBE-MR checklist of recommended items to address in reports of Mendelian randomization studies^{1 2}

Item No.	Section	Checklist item	Page No.	Relevant text from manuscript
1	TITLE and ABSTRACT	Indicate Mendelian randomization (MR) as the study's design in the title and/or the abstract if that is a main purpose of the study	1	Mendelian randomization (MR) to determine causal relationships between DEGs and BA.
INTRODUCTION				
2	Background	Explain the scientific background and rationale for the reported study. What is the exposure? Is a potential causal relationship between exposure and outcome plausible? Justify why MR is a helpful method to address the study question	2	Mendelian randomization (MR) represents an innovative approach that uses genetic variants as proxies to evaluate how exposure factors influence outcomes. The key advantage of this method lies in its ability to reduce confounding and reverse causality bias, thereby strengthening causal inference.
3	Objectives	State specific objectives clearly, including pre-specified causal hypotheses (if any). State that MR is a method that, under specific assumptions, intends to estimate causal effects	2	Our study integrates expression quantitative trait loci (eQTL) and transcriptomic data, applying Mendelian randomization to estimate the causal relationship between differentially expressed genes (DEGs) and the pathogenesis of BA, with the aim of identifying new, reliable biomarkers for BA.
METHODS				
4	Study design and data sources	Present key elements of the study design early in the article. Consider including a table listing sources of data for all phases of the study. For each data source contributing to the analysis, describe the following:		
	a)	Setting: Describe the study design and the underlying population, if possible. Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection, when available.	N/A	N/A
	b)	Participants: Give the eligibility criteria, and the sources and methods of selection of participants. Report the sample size, and whether any power or sample size calculations were carried out prior to the main analysis	N/A	N/A

	c) Describe measurement, quality control and selection of genetic variants	4	The eQTL summary data used in this study were obtained from the GWAS Catalog website (https://gwas.mrcieu.ac.uk/). The BA data used in this study were derived from a GWAS conducted by Glessner JT et al., involving 811 BA cases and 4,654 genetically matched controls, available for download and query through the GWAS summary data (GCST90296484) The R package “TwoSampleMR” was used to identify strongly associated SNPs ($p < 5e-08$) as instrumental variables (IVs). Linkage disequilibrium parameters were set to $r^2 < 0.001$, ensuring that the selected SNPs were not in close genetic proximity (at least 10,000 kbp apart). The F-statistic was calculated using the formula $F = (R^2 (N-K-1)) / (K (1-R^2))$, where R^2 represents the proportion of variance in exposure explained by IVs, K denotes the number of SNPs, and N represents the sample size. SNPs with weak instrument strength ($F < 10$) were excluded from the analysis.
	d) For each exposure, outcome, and other relevant variables, describe methods of assessment and diagnostic criteria for diseases	N/A	N/A
	e) Provide details of ethics committee approval and participant informed consent, if relevant	4	All GWAS data used in this study are publicly accessible and were ethically approved through the original analyses.
5	Assumptions Explicitly state the three core IV assumptions for the main analysis (relevance, independence and exclusion restriction) as well assumptions for any additional or sensitivity analysis	4	The “TwoSampleMR” software package was used for MR analysis, and the three core assumptions of MR analysis (Relevance Assumption, Independence Assumption,

Exclusion Restriction Assumption) were followed. The

6	Statistical methods: main analysis	Describe statistical methods and statistics used		
	a)	Describe how quantitative variables were handled in the analyses (i.e., scale, units, model)	N/A	N/A
	b)	Describe how genetic variants were handled in the analyses and, if applicable, how their weights were selected	4	The R package “TwoSampleMR” was used to identify strongly associated SNPs ($p < 5e-08$) as instrumental variables (IVs). Linkage disequilibrium parameters were set to $r^2 < 0.001$, ensuring that the selected SNPs were not in close genetic proximity (at least 10,000 kbp apart). The F-statistic was calculated using the formula $F = (R^2 (N-K-1)) / (K (1-R^2))$, where R^2 represents the proportion of variance in exposure explained by IVs, K denotes the number of SNPs, and N represents the sample size. SNPs with weak instrument strength ($F < 10$) were excluded from the analysis.
	c)	Describe the MR estimator (e.g. two-stage least squares, Wald ratio) and related statistics. Detail the included covariates and, in case of two-sample MR, whether the same covariate set was used for adjustment in the two samples	4	The inverse variance weighted (IVW) method was employed to assess the relationship between genes and BA. Additional sensitivity analyses were performed using the MR-Egger, simple mode, weighted median, and weighted mode methods
	d)	Explain how missing data were addressed	N/A	N/A
	e)	If applicable, indicate how multiple testing was addressed	N/A	N/A

7	Assessment of assumptions	Describe any methods or prior knowledge used to assess the assumptions or justify their validity	4	The R package “TwoSampleMR” was used to identify strongly associated SNPs ($p < 5e-08$) as instrumental variables (IVs). Linkage disequilibrium parameters were set to $r^2 < 0.001$, ensuring that the selected SNPs were not in close genetic proximity (at least 10,000 kbp apart). The F-statistic was calculated using the formula $F = (R^2 (N-K-1)) / (K (1-R^2))$, where R^2 represents the proportion of variance in exposure explained by IVs, K denotes the number of SNPs, and N represents the sample size. SNPs with weak instrument strength ($F < 10$) were excluded from the analysis.
8	Sensitivity analyses and additional analyses	Describe any sensitivity analyses or additional analyses performed (e.g. comparison of effect estimates from different approaches, independent replication, bias analytic techniques, validation of instruments, simulations)	4, 5	Heterogeneity tests, pleiotropy tests, and leave-one-out sensitivity analysis were conducted to evaluate the robustness of the MR results. The results were visualized through scatter plots, forest plots, and funnel plots.
9	Software and pre-registration			
	a)	Name statistical software and package(s), including version and settings used	3, 4	R software (version 4.3.2). The R package “TwoSampleMR” was used to identify strongly associated SNPs ($p < 5e-08$) as instrumental variables (IVs). All GWAS data used in this study are publicly accessible and were ethically approved through the original analyses.
	b)	State whether the study protocol and details were pre-registered (as well as when and where)	N/A	N/A

RESULTS

10	Descriptive data		
	a) Report the numbers of individuals at each stage of included studies and reasons for exclusion. Consider use of a flow diagram	4	The BA data used in this study were derived from a GWAS conducted by Glessner JT et al., involving 811 BA cases and 4,654 genetically matched controls, available for download and query through the GWAS summary data (GCST90296484)
	b) Report summary statistics for phenotypic exposure(s), outcome(s), and other relevant variables (e.g. means, SDs, proportions)	N/A	N/A
	c) If the data sources include meta-analyses of previous studies, provide the assessments of heterogeneity across these studies	N/A	N/A
	d) For two-sample MR: i. Provide justification of the similarity of the genetic variant-exposure associations between the exposure and outcome samples ii. Provide information on the number of individuals who overlap between the exposure and outcome studies	5	i.3.2 Identification of Differentially Expressed Genes Based on the screening of differentially expressed genes in the GEO database, this study believes that the inclusion of associated EQTLs is relevant to disease outcomes. ii :N/A
11	Main results		
	a) Report the associations between genetic variant and exposure, and between genetic variant and outcome, preferably on an interpretable scale	6	After filtering, 26,152 SNPs were selected as instrumental variables (detailed in Supplementary Table S2). Based on the MR analysis results and filtering criteria, 222 BA-related genes were identified (Supplementary Table S3 for details).
	b) Report MR estimates of the relationship between exposure and outcome, and the measures of uncertainty from the MR analysis, on an interpretable scale, such as odds ratio or relative risk per SD difference	7	We then performed MR analysis on these 7 co-expressed genes to determine their causal effects on the disease. The results showed that all 4 upregulated co-expressed genes had a significant positive causal relationship with BA. Specifically: C12orf75 [OR = 1.312 (95% CI: 1.022–1.685), P = 0.033], PSD3 [OR = 2.208 (95% CI: 1.253–3.890), P = 0.003], CRIM1 [OR = 1.857 (95% CI: 1.186–2.908), P = 0.006], and CHIT1 [OR = 1.942

(95% CI: 1.218–3.098), $P = 0.005$]. In contrast, the 3 downregulated co-expressed genes showed a significant negative causal relationship with BA: SEC14L4 [OR = 0.496 (95% CI: 0.294–0.838), $P = 0.008$], MAPRE3 [OR = 0.632 (95% CI: 0.435–0.919), $P = 0.016$], and TCEA3 [OR = 0.636 (95% CI: 0.411–0.985), $P = 0.042$]. The results of all five MR methods for the co-expressed genes were consistent (Figure 6).

	c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	N/A	N/A
	d) Consider plots to visualize results (e.g. forest plot, scatterplot of associations between genetic variants and outcome versus between genetic variants and exposure)	7	The results of all five MR methods for the co-expressed genes were consistent (Figure 6).
12	Assessment of assumptions		
	a) Report the assessment of the validity of the assumptions	7	Additionally, the co-expressed genes passed heterogeneity and pleiotropy tests (Supplementary Table S4).
	b) Report any additional statistics (e.g., assessments of heterogeneity across genetic variants, such as I^2 , Q statistic or E-value)	7	Additionally, the co-expressed genes passed heterogeneity and pleiotropy tests (Supplementary Table S4).
13	Sensitivity analyses and additional analyses		
	a) Report any sensitivity analyses to assess the robustness of the main results to violations of the assumptions	7	Scatter plots, forest plots, funnel plots, and leave-one-out sensitivity analyses can be found in Supplementary Figure S1. To further clarify the chromosomal distribution of these genes, we visualized the co-expressed genes (Figure 7).

b)	Report results from other sensitivity analyses or additional analyses	7	Scatter plots, forest plots, funnel plots, and leave-one-out sensitivity analyses can be found in Supplementary Figure S1. To further clarify the chromosomal distribution of these genes, we visualized the co-expressed genes (Figure 7).
----	---	---	---

c)	Report any assessment of direction of causal relationship (e.g., bidirectional MR)	N/A	N/A
----	--	-----	-----

d)	When relevant, report and compare with estimates from non-MR analyses	N/A	N/A
----	---	-----	-----

e)	Consider additional plots to visualize results (e.g., leave-one-out analyses)	7	Scatter plots, forest plots, funnel plots, and leave-one-out sensitivity analyses can be found in Supplementary Figure S1. To further clarify the chromosomal distribution of these genes, we visualized the co-expressed genes (Figure 7).
----	---	---	---

DISCUSSION

14	Key results	Summarize key results with reference to study objectives	11	In this study, we integrated GWAS and transcriptomic data and used Mendelian randomization (MR) to assess the causal relationship between differentially expressed genes (DEGs) and BA pathogenesis. Our analysis identified 458 upregulated and 358 downregulated genes, which were intersected with BA-related genes identified through MR analysis. This led to the discovery of 7 key co-expressed genes, including 4 upregulated (C12orf75, PSD3, CRIM1, and CHIT1) and 3 downregulated genes (SEC14L4, MAPRE3, and TCEA3).
15	Limitations	Discuss limitations of the study, taking into account the validity of the IV assumptions, other sources of potential bias, and imprecision. Discuss both direction and magnitude of any potential bias and any efforts to address them	12, 13	However, several limitations should be acknowledged: (1) The small sample size may introduce bias, so larger and multi-center

studies are needed to confirm these findings. (2) Although correlations between co-expressed genes and BA were identified, the specific mechanisms by which these genes contribute to BA remain unclear, and further basic research is required. (3) The GWAS data used in this study were derived from European populations, which introduces potential racial limitations. Future studies should include samples from diverse ethnic backgrounds to increase the generalizability of the results.

16 Interpretation

- a) Meaning: Give a cautious overall interpretation of results in the context of their limitations and in comparison with other studies

13 However, several limitations should be acknowledged: (1) The small sample size may introduce bias, so larger and multi-center studies are needed to confirm these findings. (2) Although correlations between co-expressed genes and BA were identified, the specific mechanisms by which these genes contribute to BA remain unclear, and further basic research is required. (3) The GWAS data used in this study were derived from European populations, which introduces potential racial limitations.

- b) Mechanism: Discuss underlying biological mechanisms that could drive a potential causal relationship between the investigated exposure and the outcome, and whether the gene-environment equivalence assumption is reasonable. Use causal language carefully, clarifying that IV estimates may provide causal effects only under certain assumptions

11,12 In this study, the 4 upregulated genes (C12orf75, PSD3, CRIM1, and CHIT1) suggest their potential significance in the pathological mechanisms of BA. C12orf75, located on chromosome 12, is not widely studied in the literature, but our findings suggest its involvement in immune regulation, consistent with the conclusions of Cai G and Wu Y et al[21, 22]. PSD3, located on

chromosome 8, encodes a protein containing Pleckstrin and Sec7 domains, which are involved in lipid metabolism, intracellular signaling pathways, and the regulation of the cytoskeleton, cell proliferation, and migration[23, 24]. Moreover, PSD3 polymorphisms have been linked to autoimmune diseases such as systemic sclerosis and diabetes[25, 26]. Given BA's innate immune characteristics[27], PSD3 may play a critical role in the immune response, aligning with the findings of this study. CRIM1, located on chromosome 2, has been shown to be involved in organogenesis in the placenta and is associated with cell proliferation and angiogenesis[28]. In the context of BA, the upregulation of CRIM1 suggests its role in bile duct formation, tissue remodeling, and liver fibrosis. CHIT1, located on chromosome 1, is a marker protein of activated macrophages and is involved in tissue repair and inflammatory responses[29]. The upregulation of CHIT1 indicates the importance of macrophages in the immune microenvironment of BA, potentially promoting chronic inflammation and fibrosis of the bile ducts, consistent with the single-cell sequencing results of Abigail E et al. in BA[30].

The three downregulated genes identified in this study (SEC14L4, MAPRE3, and TCEA3) also play critical roles in BA pathogenesis. SEC14L4, located on chromosome 22, is involved in phospholipid metabolism and lipid transport[31]. While SEC14L4 has not previously been reported in the context of BA, its downregulation in this study suggests that lipid metabolism abnormalities may be associated with bile duct injury and bile flow

				obstruction, influencing disease progression in BA. MAPRE3, located on chromosome 2, primarily regulates microtubule stability within the cytoskeleton[32]. The downregulation of MAPRE3 in BA may lead to structural and functional damage in bile duct cells, affecting bile flow and further contributing to the development of biliary atresia. To date, there have been no studies linking MAPRE3 to BA, and our findings may open a new direction for research. TCEA3, located on chromosome 12, is a transcription elongation factor for RNA polymerase II, responsible for regulating gene transcription[33]. Its downregulation in BA may be closely related to reduced transcriptional activity in biliary epithelial cells, warranting further investigation into its role in BA.
	c)	Clinical relevance: Discuss whether the results have clinical or public policy relevance, and to what extent they inform effect sizes of possible interventions	12	This study not only reinforces the current body of knowledge but also lays the groundwork for developing precise immunotherapy strategies for BA, offering new directions for future exploration.
17	Generalizability	Discuss the generalizability of the study results (a) to other populations, (b) across other exposure periods/timings, and (c) across other levels of exposure	13	Future studies should include samples from diverse ethnic backgrounds to increase the generalizability of the results.
OTHER INFORMATION				
18	Funding	Describe sources of funding and the role of funders in the present study and, if applicable, sources of funding for the databases and original study or studies on which the present study is based	N/A	N/A
19	Data and data sharing	Provide the data used to perform all analyses or report where and how the data can be accessed, and reference these sources in the article. Provide the statistical code needed to reproduce the results in the article, or report whether the code is publicly accessible and if so, where	13	GWAS data can be accessed and downloaded from the GWAS Catalog website (https://gwas.mrcieu.ac.uk/).

20	Conflicts of Interest	All authors should declare all potential conflicts of interest	N/A	N/A
----	------------------------------	--	-----	-----

This checklist is copyrighted by the Equator Network under the Creative Commons Attribution 3.0 Unported (CC BY 3.0) license.

1. Skrivankova VW, Richmond RC, Woolf BAR, Yarmolinsky J, Davies NM, Swanson SA, et al. Strengthening the Reporting of Observational Studies in Epidemiology using Mendelian Randomization (STROBE-MR) Statement. JAMA. 2021;under review.
2. Skrivankova VW, Richmond RC, Woolf BAR, Davies NM, Swanson SA, VanderWeele TJ, et al. Strengthening the Reporting of Observational Studies in Epidemiology using Mendelian Randomisation (STROBE-MR): Explanation and Elaboration. BMJ. 2021;375:n2233.