

Liability threshold model-based disease risk prediction based on Electronic Health Record phenotypes

Corresponding Author: Dr Iuliana Ionita-Laza

Version 0:

Decision Letter:

16th February 2024

Dear Iuliana,

Your Article "Liability threshold model-based disease risk prediction based on Electronic Health Record phenotypes" has been seen by two referees. You will see from their comments below that, while they find your work of potential interest, they have raised several overlapping concerns that must be addressed. In light of these comments, we cannot accept the manuscript for publication at this time, but we would be interested in considering a substantially revised version that addresses the referees' concerns.

We hope you will find the referees' comments useful as you decide how to proceed. If you wish to submit a substantially revised manuscript, please bear in mind that we will be reluctant to approach the referees again in the absence of major revisions.

To guide the scope of the revisions, the editors discuss the referee reports in detail within the team, including with the chief editor, with a view to identifying key priorities that should be addressed in revision, and sometimes overruling referee requests that are deemed beyond the scope of the current study. In this case, we ask that you perform additional analyses to calibrate the performance of LTPI relative to other multi-trait and phenotype imputation methods, including careful assessment of the specificity of GWAS performed on LTPI-predicted liability, control of false positives, and other technical issues raised by the referees. We hope you will find this prioritized set of referee points to be useful when revising your study. Please do not hesitate to get in touch if you would like to discuss these issues further.

If you choose to revise your manuscript taking into account all reviewer and editor comments, please highlight all changes in the manuscript text file. At this stage, we will need you to upload a copy of the manuscript in MS Word .docx or similar editable format.

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

If revising your manuscript:

*1) Include a "Response to referees" document detailing, point-by-point, how you addressed each referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the referees along with the revised manuscript.

*2) If you have not done so already, please begin to revise your manuscript so that it conforms to our Article format instructions, available here. Refer also to any guidelines provided in this letter.

*3) Include a revised version of any required Reporting Summary: <https://www.nature.com/documents/nr-reporting-summary.pdf>

It will be available to referees (and, potentially, statisticians) to aid in their evaluation if the manuscript goes back for peer review.

A revised checklist is essential for re-review of the paper.

Please be aware of our guidelines on

digital image standards.

You may use the link below to submit your revised manuscript and related files:

Link Redacted

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

If you wish to submit a suitably revised manuscript, we hope to receive it within 3-6 months. If you cannot send it within this time, please let us know. We will be happy to consider your revision so long as nothing similar has been accepted for publication at Nature Genetics or published elsewhere. Should your manuscript be substantially delayed without notifying us in advance and your article is eventually published, the received date would be that of the revised, not the original, version.

Please do not hesitate to contact me if you have any questions or would like to discuss the required revisions further.

Nature Genetics is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information, please visit www.springernature.com/orcid.

Thank you for the opportunity to review your work.

Sincerely,
Kyle

Kyle Vogan, PhD
Senior Editor
Nature Genetics
<https://orcid.org/0000-0001-9565-9665>

Referee expertise:

Referee #1: Genetics, computational methods, electronic health records

Referee #2: Genetics, computational methods, electronic health records

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:

This paper aims to address an interesting challenge to predict continuous disease value (liability) underlying binary traits. It follows past work on Liability-threshold models such as Pedersen et al. 2022 AJHG, Hujoel et al. 2022 Cell Genomics, and Hujoel et al. 2020 Nature Genetics. The most important application presented in this work is improving GWAS power; the authors also tried to use imputed liability to predict binary labels, which is less convincing as the model intrinsically uses the binary label. The paper aligns with recent efforts on leveraging additional traits to increase GWAS power (Turley et al. 2018 Nature Genetics, Cortes et al. 2017/2020 Nature Genetics, Zhang et al. 2023 Cell Genomics, Dahl et al. 2023 Nature Genetics, and An et al. 2023 Nature Genetics), which could be included for comparison to better illustrate the contribution of this work.

There are several merits of this paper, including clear explanation of statistical methodology, comprehensive simulations, good separation of training and validation data sets (eMERGE and UKB), and statistically principled trait selection method (termed "ATSA" in the paper). However, I do have a few major comments/concerns that I think the authors need to address plus several minor comments/questions.

Main comments:

1. Additional methods need to be compared. LT-FH models (Pedersen et al. 2022 AJHG, Hujoel et al. 2022 Cell Genomics, Hujoel et al. 2020 Nature Genetics) are designed specifically for including family history using liability-threshold model; however, the proposed model uses a wide range of phenotypes to predict liability, which is closer to phenotype imputation models (Dahl et al. 2023 Nature Genetics, An et al. 2023 Nature Genetics, Bastarache et al. 2018 Science). It is unclear whether the proposed method, which uses non-target traits for imputation, has performance comparable to simpler methods

such as matrix factorization (e.g. SoftImpute used in Dahl et al. 2023). As the proposed model could use both Family-history and phenotype information, it would suffice to show that (1) the proposed model has comparable performance to state-of-art methods when only using traits and (2) when proposed model uses both FH and non-target traits, it improves GWAS power compared to models that use only one set of the information.

2. The specificity of GWAS on predicted liability needs to be evaluated. Specifically, in GWAS analyses on imputed traits (in this case, non-target traits span all PheCodes), GWAS signals could be spurious signals of other PheCodes that are used to impute/predict the liability. These effects should be evaluated (Dahl et al. 2023 Nature Genetics) and could be potentially corrected (Miao et al. 2024 MedRxiv <https://www.medrxiv.org/content/10.1101/2024.01.03.24300779v1>). Despite mentioning “controlling the false positive rate (FPR) driven by non-target traits” in the text, I did not find detailed methods for this. Figure 9BD shows that a significant proportion of variants are not replicated in the GWASs, suggesting potential false positives.

3. It is unclear what are the contribution of risk prediction sections due to the lack of independent gold standard labels (“Assessing the impact of imputation on prediction accuracy within LTPI: eMERGE-CKD analysis”, “Construction of LTPI and LTFH scores for disease risk prediction”, “Performance assessment of disease risk prediction accuracy”). In all these three sections, the prediction labels are used to predict liability, making it hard to interpret the results. It is important to come up with independent gold standard labels, at least for example diseases. eMERGE-CKD analysis uses eGFR to predict liability, while evaluating it on eGFR generated labels; the latter two sections use disease labels (PheCode-FH) to predict posterior liability, while evaluating it on highly correlated PheCodes. The authors aim to compare relative performance to LTFH, but this is not the best comparison (as models use different information), and the labels are used as a predictor. I suggest for all these analyses to exclude the target disease label and highly correlated traits to focus on the improved prediction power from the not clearly related non-target traits (i.e. excluding non-target traits that clearly denote the same disease).

Minor comments/questions:

1. “non-diagonal elements represent covariances due to non-genetic factors.”... “Similarly, the environmental covariance matrix Σ can be estimated in LDSC.” Please clarify how you compute non-diagonal elements of the environmental covariance matrix using LDSC?

2. “Maximum Likelihood Estimator (MLE; $LTP1$) for the target disease’s genetic liability, leveraging additional continuous non-target phenotypes.” Please explain the rationale for using posterior mean in $LTP1b$ but use MLE for $LTP1c$; why not just get a posterior mean estimate using both continuous and binary non-target traits (when both are available)?

3. In the methods section of ATSA: “where R represents the variance of genetic liability on T explained by O ”; based on the derivation, here ROT represents the proportion of genetic variance that are explained (which is different from the “variance of genetic liability”).

4. “We assumed no correlations in environmental liability between traits” Environment liability is independent from genetic liability across individuals for a given trait; however, environment liability between traits is correlated. It is worth testing whether the model is robust with strong environmental correlation between traits.

5. It is unclear from Figure 2 what is the number of binary vs. quantitative traits in the red dots (mix of binary and quantitative).

6. Figure 2: what is the observed upper limit of R^2 between binary observation and LIABgen? If that observed R^2 between binary and simulated liability value is transformed to liability scale, what is the liability-scale R^2 (equation 8 in Lee et al. 2012 Genetic Epidemiology)? The figure doesn’t present the advantage of transforming observed binary traits to a predicted liability as we have been using liability-scale R^2 /heritability; it just shows the prediction gets better as the model includes more genetically correlated predictors. I think it is more important to clarify why we want continuous liabilities instead of just using the binary labels (for R^2 , we already have the needed tools to transform from binary scale to liability-scale, what is the added value?).

7. It should be clarified in simulation whether the heritability is on observed scale or liability scale (I am guessing that it refers to the latter in simulations from the text).

8. In “Factors affecting LTPI’s GWAS performance”: “due to LTPI’s lower genetic correlation ($\rho=0.1$) against LTFH’s inter-family genetic correlation of 0.5. However, as genetic correlation among different traits increased to 0.3, 0.5, and 0.7, LTPI gained the upper hand”. If there are 10 additional traits used as predictor, each has correlation 0.5 to the target trait, it is unfair to compare to family history which is only one additional variable. A fair comparison should report “variance of genetic liability” (defined in ATAS part) of ten binary traits. The decreased performance could be caused by noise induced when mapping from non-target binary traits to liability.

9. “As disease prevalence varies between 0.01, 0.05, and 0.1, LTPI maintains roughly 43% power.” Is this just varying target disease prevalence?

10. “GHK” approximation. It is unclear from the main text how GHK method is used (e.g. no mention of it in the “Overview of LTPI construction”); it would help the reader to understand the rationale of choosing an approximated method instead of sampling from the exact posterior distribution. Also, in the main text, GHK is mentioned to compute only posterior mean and variance but not approximate the posterior distribution, while later it is mentioned the estimate is “skewed”; this should be

clarified.

11. If eGFR is used as a continuous non-target trait for prediction liability, wouldn't G-stage cases (defined by eGFR) be highly correlated with the liability? This looks like a circular evaluation.

12. Connected to the previous point, (Figure 5) comparing the eGFR defined case-control with the score (liability) predicted using eGFR doesn't make much sense. Figure 5CDIJ is more informative than Figure 5ABGH.

13. Figure 5E: the Phecodes (binary)--G-stage trait (binary) AUC is not well defined (it is just a single number of FPR and TPR) and thus not comparable to liability--G-defined traits; it is more comparable to compute the liability-scale correlation of PheCodes--G-stage vs liability--G-stage.

14. "Assessing the impact of imputation on prediction accuracy within LTPI: eMERGE-CKD analysis" connected to major comments, it is interesting to see whether directly using SoftImpute to predict liability has comparable performance to the proposed model.

Reviewer #2:

Remarks to the Author:

The manuscript presents the Liability Threshold Phenotypic Integration (LTPI) method, aimed at improving disease phenotype classification by integrating genetic and phenotypic data. It introduces LTPI as a novel approach, while also pointing out the need for further clarification, validation, and comparison with current methodologies.

Strengths of the Manuscript:

- * LTPI is lauded for its innovative approach to phenotype classification, potentially surpassing GWAS and family history methods by incorporating multi-trait liability thresholds and genetic correlations.

- * The manuscript is praised for its thorough explanation of the LTPI algorithm, including the rationale behind it and the methodological rigor in its development and assessment.

- * The adoption of importance sampling for estimating genetic liability and an automatic trait selection algorithm are highlighted as significant advantages, promising greater power and accuracy in prediction within large-scale biobanks.

Major Concerns:

- * There are concerns about the control of false positives and the real-world applicability of LTPI, especially in scenarios utilizing only single-modal images or when applying the method to independent cohorts in Stage 1 and Stage 2.

- * Additional simulations or validations are deemed necessary to explore the impact of assumptions being violated, especially considering the reliance on EHR data, which may not accurately reflect the population's comorbidity and prevalence.

- * Questions are raised regarding the practicality of LTPI in comparison to existing multi-trait analysis approaches and its ability to address the complexities of pleiotropy and the genetic architecture of complex diseases.

- * A need for further explanation on how LTPI handles non-target traits and its comparison with the latest GWAS findings, as well as a deeper discussion on potential biases stemming from cohort characteristic differences between datasets such as the UK Biobank and eMERGE.

Minor Points:

- * Improvements in figure clarity are suggested, particularly in depicting the integration of binary and continuous traits into the LTPI score and the influence of disease type on LTPI's accuracy.

- * A more comprehensive evaluation of LTPI across various disease categories is recommended, along with a discussion on its translational implications for clinical research and precision medicine.

- * Clarification is sought on the incorporation of family history into the LTPI score and the rationale behind the reported high AUC values.

- * Concerns are noted regarding genomic inflation observed in several GWAS results within Table S19, prompting a call for discussion on its origins.

Version 1:

Decision Letter:

9th August 2024

Dear Iuliana,

Your revised Article "Liability threshold model-based disease risk prediction based on Electronic Health Record phenotypes" has been seen by the original referees. You will see from their comments below that, while Reviewer #1 is satisfied and has no further requests, Reviewer #2 has requested a few additional revisions. We remain interested in the publishing your study in Nature Genetics, but we would like to see your response to these points in the form of a further revision before we make a decision on publication and provide final formatting guidelines.

We therefore invite you to revise your manuscript taking into account Reviewer #2's comments. Please highlight all changes in the manuscript text file. At this stage, we will need you to upload a copy of the manuscript in MS Word .docx or similar editable format.

When revising your manuscript:

*1) Include a "Response to referees" document detailing, point-by-point, how you addressed each referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the referees along with the revised manuscript.

*2) If you have not done so already, please begin to revise your manuscript so that it conforms to our Article format instructions, available

[here](http://www.nature.com/ng/authors/article_types/index.html).

Refer also to any guidelines provided in this letter.

*3) Include a revised version of any required Reporting Summary: <https://www.nature.com/documents/nr-reporting-summary.pdf>

It will be available to referees (and, potentially, statisticians) to aid in their evaluation if the manuscript goes back for peer review.

A revised checklist is essential for re-review of the paper.

Please be aware of our [guidelines](https://www.nature.com/nature-research/editorial-policies/image-integrity) on digital image standards.

Please use the link below to submit your revised manuscript and related files:

Link Redacted

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We hope to receive your revised manuscript within four weeks. If you cannot send it within this time, please let us know.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

Nature Genetics is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information, please visit www.springernature.com/orcid.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Sincerely,
Kyle

Kyle Vogan, PhD
Senior Editor
Nature Genetics
<https://orcid.org/0000-0001-9565-9665>

Referee expertise:

Referee #1: Genetics, computational methods, electronic health records

Referee #2: Genetics, computational methods, electronic health records

Reviewers' Comments:

Reviewer #1:

Remarks to the Author:

I appreciate the comprehensive response from the authors to all major and minor points raised. I think the revised manuscript addressed all my concerns.

Reviewer #2:

Remarks to the Author:

Overall, the authors have addressed the major and minor concerns raised during the review process effectively. The distinctions between LTPI and existing multi-trait analysis approaches were clarified, and the explanations regarding how LTPI handles non-target traits, along with the integration of recent GWAS findings, were well-articulated. Additionally, the discussion on potential biases stemming from differences between datasets, as well as the concern about genomic inflation in GWAS results, were appropriately addressed.

However, I noticed that Figure 1 rendered incorrectly in both documents, and I still find it confusing. Specifically, I'm unclear whether the y-axis of the boxes represents phenotypes. The term "dichotomous" is used, but is this simply the same as binary? I assume the dots on the density plots represent the final scores, but this isn't mentioned explicitly. Moreover, it's unclear why these dots are indicated in the Step 1 panel at all. Although I eventually understood the figure, I believe the authors should revisit and potentially revise it to improve clarity. Additionally, I feel that R2m1 has not been sufficiently addressed.

Given the overall improvements and the remaining issues with Figure 1 and R2m1, I recommend the paper for pre-acceptance with the condition that these concerns are addressed. The authors have made significant strides, but further refinement of these aspects would enhance the manuscript's clarity and impact.

Version 2:

Decision Letter:

Our ref: NG-A64248R1

8th October 2024

Dear Iuliana,

Thank you for submitting your revised manuscript "Liability threshold model-based disease risk prediction based on Electronic Health Record phenotypes" (NG-A64248R1). In light of your responses to Reviewer #2, we will be happy in principle to publish your study in Nature Genetics as an Article pending final revisions to comply with our editorial and formatting guidelines.

We are now performing detailed checks on your paper, and we will send you a checklist detailing our editorial and formatting requirements soon. Please do not upload the final materials or make any revisions until you receive this additional information from us.

Thank you again for your interest in Nature Genetics. Please do not hesitate to contact me if you have any questions.

Sincerely,
Kyle

Kyle Vogan, PhD
Senior Editor
Nature Genetics
<https://orcid.org/0000-0001-9565-9665>

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Response to Reviewers

Reviewer #1:

This paper aims to address an interesting challenge to predict continuous disease value (liability) underlying binary traits. It follows past work on Liability-threshold models such as Pedersen et al. 2022 AJHG, Hujoel et al. 2022 Cell Genomics, and Hujoel et al. 2020 Nature Genetics. The most important application presented in this work is improving GWAS power; the authors also tried to use imputed liability to predict binary labels, which is less convincing as the model intrinsically uses the binary label. The paper aligns with recent efforts on leveraging additional traits to increase GWAS power (Turley et al. 2018 Nature Genetics, Cortes et al. 2017/2020 Nature Genetics, Zhang et al. 2023 Cell Genomics, Dahl et al. 2023 Nature Genetics, and An et al. 2023 Nature Genetics), which could be included for comparison to better illustrate the contribution of this work.

Response: Thank you for the helpful comments and summary. Although we agree that the increase in GWAS power is the most important application presented here, it is important to emphasize that *LTPI*, as a method, provides predictions of the posterior mean genetic liability for each individual, and hence provides a way to identify individuals who have a missed diagnosis or who are likely to develop specific diseases in the future. Although the overall lack of gold standard labels in the field does lead to limited ability to evaluate the accuracy of such predictions, we have performed additional analyses by using state-of-the-art phenotyping algorithms for CKD and Breast Cancer.

We also now include comparisons with SoftImpute. *LTPI*, as *LTFH*, is built on a popular model in genetics, the liability threshold model. When family history is not available, *LTPI* provides an efficient method to leverage related traits in EHR to substantially increase power in downstream genetic analyses relative to standard PheCode based GWAS. Note however that *LTPI* is different from phenotype imputation methods such as SoftImpute because it can leverage non-target traits in the EHR to *transform a binary phenotype into a continuous one*, even in the absence of missing data. Therefore, *LTPI* is different in nature relative to phenotype imputation methods. Another difference is that given our automatic trait selection algorithm to select traits that are genetically correlated to the target trait, *LTPI* attempts to reduce the number of false positive findings that can arise due to inclusion of non-target traits. By comparison, imputation methods such as SoftImpute also leverage correlations among phenotypic features but cannot separate genetic from environmental correlations. More details are given below.

There are several merits of this paper, including clear explanation of statistical methodology, comprehensive simulations, good separation of training and validation data sets (eMERGE and UKB), and statistically principled trait selection method (termed “ATSA” in the paper). However, I do have a few major comments/concerns that I think the authors need to address plus several minor comments/questions.

Main comments:

R1M1. Additional methods need to be compared. *LT-FH* models (Pedersen et al. 2022 AJHG, Hujoel et al. 2022 Cell Genomics, Hujoel et al. 2020 Nature Genetics) are designed specifically for including family history using liability-threshold model; however, the proposed model uses a wide range of phenotypes to predict liability, which is closer to phenotype imputation models (Dahl et al. 2023 Nature Genetics, An et al. 2023 Nature Genetics, Bastarache et al. 2018 Science). It is unclear whether the proposed method, which uses non-target traits for imputation, has performance comparable to simpler methods such as matrix factorization (e.g. SoftImpute used in Dahl et al. 2023). As the proposed model could use both Family-history and phenotype information, it would suffice to show that (1) the proposed model has comparable performance to state-of-art methods when only using traits and (2) when proposed model uses both FH and non-target traits, it improves GWAS power compared to models that use only one set of the information.

Response: Thank you for the helpful suggestions. In this revision, we have added comparisons with SoftImpute (Dahl et al. 2023 Nature Genetics) as suggested (we originally tried to include also AutoComplete in our comparisons, but the software provided was quite challenging to use and the method required substantial parameter tuning on the part of a user, so we decided to focus on SoftImpute). Note that an important advantage of *LTPI* is that, unlike phenotype imputation methods such as SoftImpute and AutoComplete, it can leverage non-target traits in the EHR to transform a phenotype from a binary trait definition to a continuous one, even in the absence of missing data and hence can improve power relative to a pure GWAS approach as shown in the Figure below. Therefore, *LTPI* goes beyond what imputation methods are designed to do. Also, as discussed below and in the manuscript, phenotype imputation methods

such as SoftImpute need more careful investigation as their performance and potential to improve GWAS power are not necessarily uniform across conditions.

For this simulation, we have revised our default scheme from the previous version to include 200 traits. We assumed an AR(1) correlation structure for 10 genetically correlated traits and a similar AR(1) structure for 20 non-genetically or environmentally correlated traits, with $\rho = 0.3, 0.5, 0.7$. We elaborated on the default simulation setting on page 4 (section “Validation of *LTPI*’s performance using simulations”).

We compared *LTPI* with GWAS and *LTFH* in the absence of missing data. As before, we show substantial increases in power for *LTPI* vs. GWAS and *LTFH* as we increase the level of correlation between the target and non-target traits (panel **A**). We note that values of $\rho = 0.7$ for AR(1) structure are quite realistic in practice as there are usually closely related non-target traits for any given target trait. Similarly, combining *LTPI* and FH (*LTPI*+FH) leads to additional increases in power.

Next, we compared *LTPI* against SoftImpute using simulations in scenarios with missing data (missing completely at random), where the missing rates for the target phenotype were 30% and 50% (panels **B** and **C**). *LTPI* can capitalize on phenotypic features in the EHR that are genetically correlated to the target trait to improve power. Indeed, power increases substantially as the correlation between non-target traits and the target trait increases. In contrast, SoftImpute’s power remains stable.

We found the performance of mainstream SoftImpute, as implemented by Dahl et al., to be suboptimal for imputing binary traits in imbalanced scenarios as typical for biobank scale EHR data. The loss function in SoftImpute assumes continuous variables, and as such, its application to binary traits is affected by the binary nature of disease traits and by the heavily imbalanced case:control ratio in these large biobank scale studies – i.e., the imputation is biased towards the majority class (0’s), and hence SoftImpute’s power is similar to that of GWAS. We have confirmed that these results make sense in personal communication with Andy Dahl. It is also important to note that more work needs to be done to assess the performance of SoftImpute in the settings of binary data in EHR. There do exist alternatives for matrix completion for binary matrices that need to be evaluated in future work – see, for example, <https://academic.oup.com/imaiai/article-abstract/3/3/189/714205>. At a minimum, the loss function needs to be changed from the current squared error loss to a cross-entropy loss, for example. The imbalance issue also needs to be addressed. And finally, it is important to also consider under what settings a low-rank approximation is appropriate for phenotype matrices. We have added the following paragraph to the manuscript on page 6:

“It is important to note that the performance of SoftImpute is very similar to that of GWAS. In effect, the mainstream SoftImpute implemented in Dahl et al. does not leverage the other traits to its advantage in these simulations. There are several reasons for this suboptimal performance for SoftImpute. First, SoftImpute is a latent factor model that assumes that the data are continuous and so are the latent factors, hence this method is optimized for continuous traits, whereas our target trait and many of the non-target traits are binary. In particular, the squared error loss function could be modified to a cross-entropy loss function to account for the binary nature of the data. Second, the target trait is imbalanced with respect to the case:control ratio which leads to imputed values for cases very close to 0, and hence similar to values imputed for the control group. Third, SoftImpute assumes that the phenotype matrix can be well approximated by a low-rank matrix, which may not always be true in practice. For example, for matrices of high rank, a low-rank approximation, as in SoftImpute, may not be sufficient to capture the underlying correlations. “

The power results for these settings are reported below.

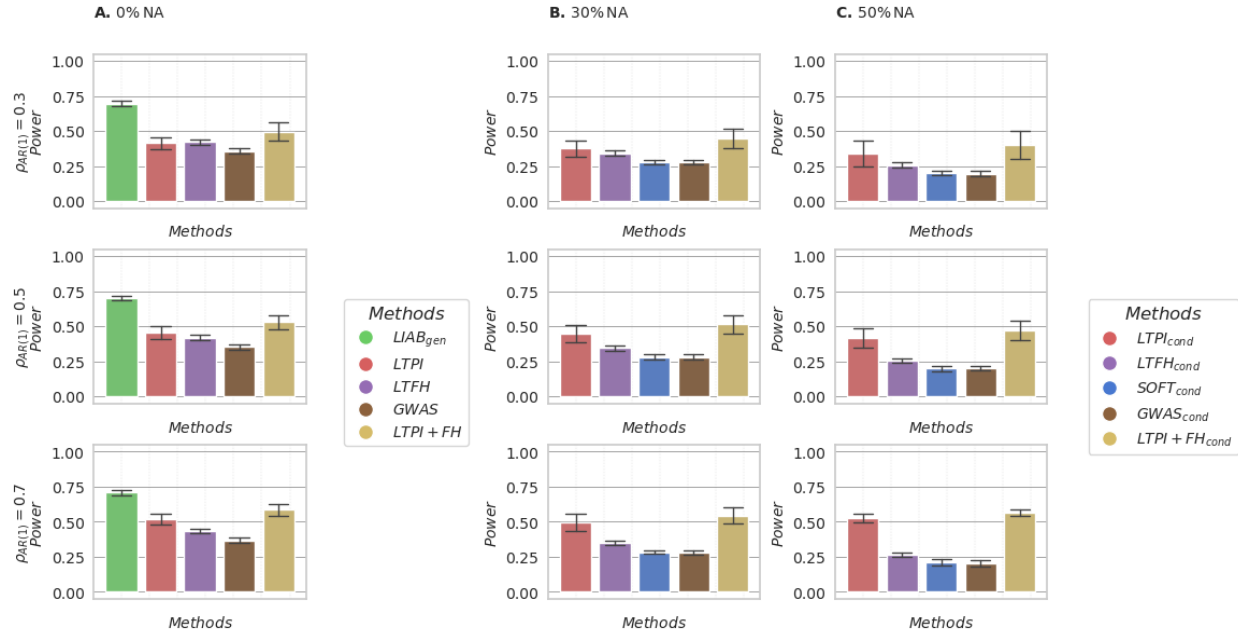


Fig. 1. Power simulation. We conducted simulations under default settings, varying the rate of missing target phenotypes (A. 0%, B. 30%, and C. 50%) and $\rho_{AR(1)} = \{0.3, 0.5, 0.7\}$. Each bar represents the average power from 25 iterations, defined as the proportion of causal SNPs with p-value $< 5 \times 10^{-8}$, and the black error bars represent the 95% confidence interval.

From left to right

- No missing data. The target trait is coded as 1/0 if the corresponding target disease phecode is present/absent. In this case, the imputation method SoftImpute is not applicable as there is nothing to impute.
- Assuming a missing rate of 30% for the target trait. We used the subscript “cond” to denote performance with missing phenotypes. For $GWAS_{conds}$ we ignored all missing phenotypes.
- Assuming missing rate 50% for the target trait.

In summary, *LTPI* increases power over phenotype imputation methods such as *SoftImpute* even when not using family history information. Incorporating family history can further increase the power.

R1M2. The specificity of GWAS on predicted liability needs to be evaluated. Specifically, in GWAS analyses on imputed traits (in this case, non-target traits span all PheCodes), GWAS signals could be spurious signals of other PheCodes that are used to impute/predict the liability. These effects should be evaluated (Dahl et al. 2023 Nature Genetics) and could be potentially corrected (Miao et al. 2024 <https://www.medrxiv.org/content/10.1101/2024.01.03.24300779v1>). Despite mentioning “controlling the false positive rate (FPR) driven by non-target traits” in the text, I did not find detailed methods for this. Figure 9BD shows that a significant proportion of variants are not replicated in the GWASs, suggesting potential false positives.

Response: Thank you for this comment. Indeed, the potential non-specificity of the detected associations is a valid concern for all the different methods that use non-target traits to impute or predict a target trait of interest. We want to emphasize that *LTPI* does leverage genetically correlated traits to improve power, as mentioned at the previous point, and, in particular, it will borrow information at pleiotropic loci from other traits to improve the power at those loci for the target trait. However, we do not think of those loci as false positives; instead, we think of them as disease loci that are shared across traits. Given our ATSA algorithm to select traits that are genetically correlated to the target trait, there is no reason for *LTPI* to be more susceptible to false positives relative to imputation methods such as

SoftImpute that also leverage correlations among phenotypic features but cannot separate genetic from environmental correlations.

Note that our previous GWAS validation results were not based on the latest GWAS studies for those diseases. We have revised the results in Figure 9 (Figure 8 in the current manuscript) to include some of the largest GWAS studies for the four diseases. We show that with these larger sample sizes, the sign consistency between *LTPI* discovered loci and validation GWAS is substantially improved, and very high indeed (see Figure below), suggesting that the newly detected *LTPI* loci are likely risk loci for the target disease. However, these loci may be of lower effect and missed in the original GWAS. The power of *LTPI* is higher because it borrows information from other diseases and traits at pleiotropic loci.

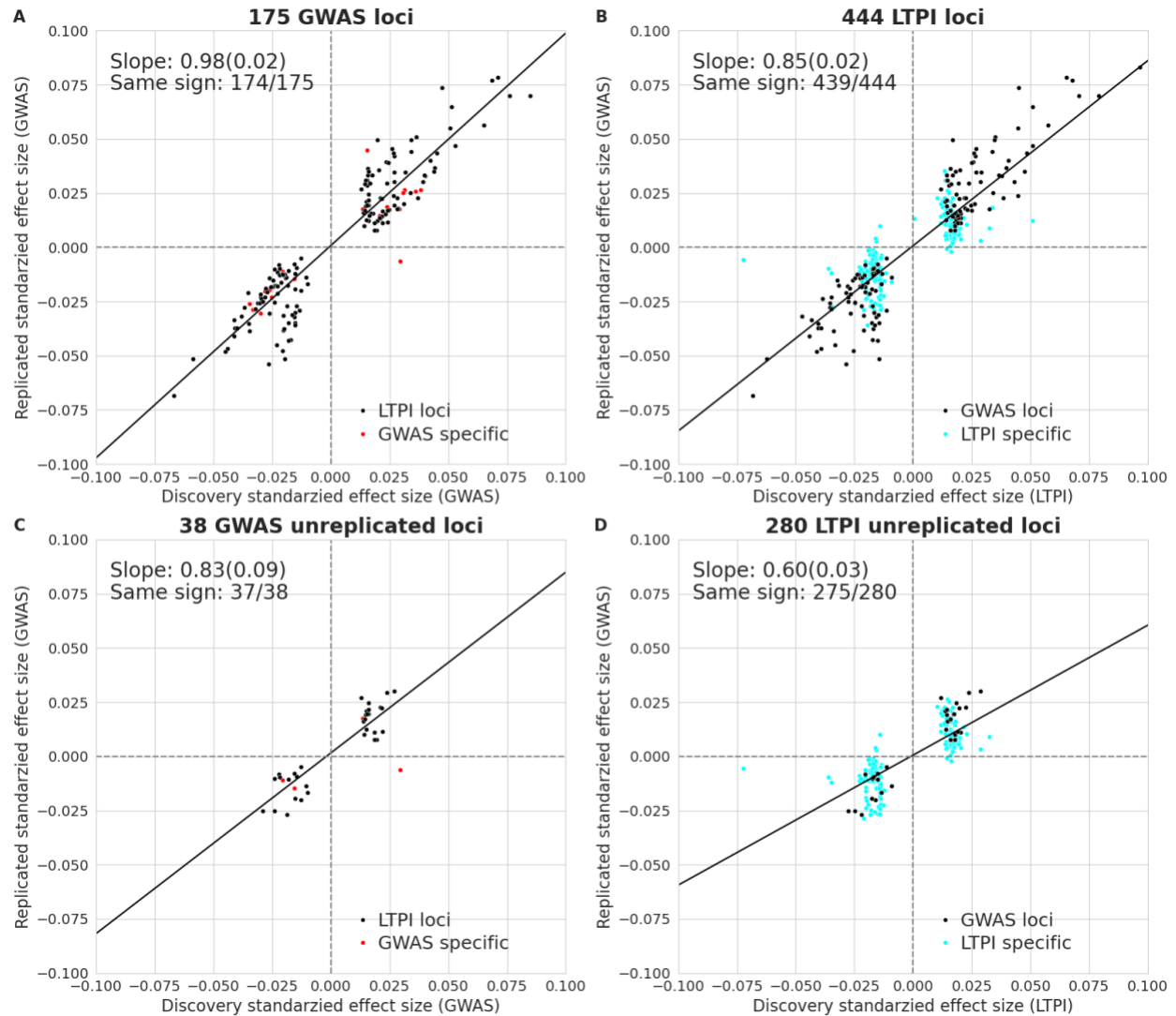


Fig. 2. Replication analysis using GWAS validation data sets. We generated scatter plots of standardized effect sizes ($Z/\sqrt{N_{\text{eff}}}$) in the validation GWAS data vs. the UK Biobank discovery data, aggregated across 4 diseases (CAD, T2D, breast cancer, and prostate cancer) for **A.** 175 loci identified by GWAS, **B.** 444 loci identified by *LTPI*, **C.** 38 loci identified by GWAS but not replicated by the validation GWAS data, and **D.** 280 loci identified by *LTPI* but not replicated by the validation GWAS data. Each point indicates the lead SNP at an independent locus, with the color indicating whether the locus contains genome-wide significant variants identified by both *LTPI* and GWAS.

As suggested, we have also used the PRSpleiotropy measure in Dahl et al. to assess the specificity of the GWAS findings for the different methods first in simulations and then in the real data applications. However, a higher PRS pleiotropy can simply reflect higher pleiotropy among the target and non-target traits. The simulations below represent this scenario as the causal variants are shared across the target and non-target traits, and therefore, there are no false positives but genuine sharing of genetic effects among the different traits.

Our simulations evaluated the PRS pleiotropy for the competing methods in the three scenarios above (no missingness, 30% or 50% missingness for the target phenotype). We found that, indeed, *the PRS pleiotropy increases as we increase the level of correlation among genetically correlated traits with the target trait*, therefore a higher PRS pleiotropy in itself does not imply higher rate of false positives. It is a reflection of the fact that *LTPI* leverages those genetically correlated non-target traits to build the resulting score. Although we expect *LTPI* to have higher PRS pleiotropy relative to methods that do not borrow information from non-target traits (such as *LTFH* and *GWAS*), the higher PRS pleiotropy for *LTPI* vs. *SoftImpute* can appear surprising at first, as *SoftImpute* is also borrowing information from other traits by leveraging the underlying correlations.

As described at the previous point, *SoftImpute*'s performance is suboptimal in these simulations and in fact its performance is very similar to that of *GWAS*, hence it is not surprising that its PRS ratio is also very similar to that of *GWAS*.

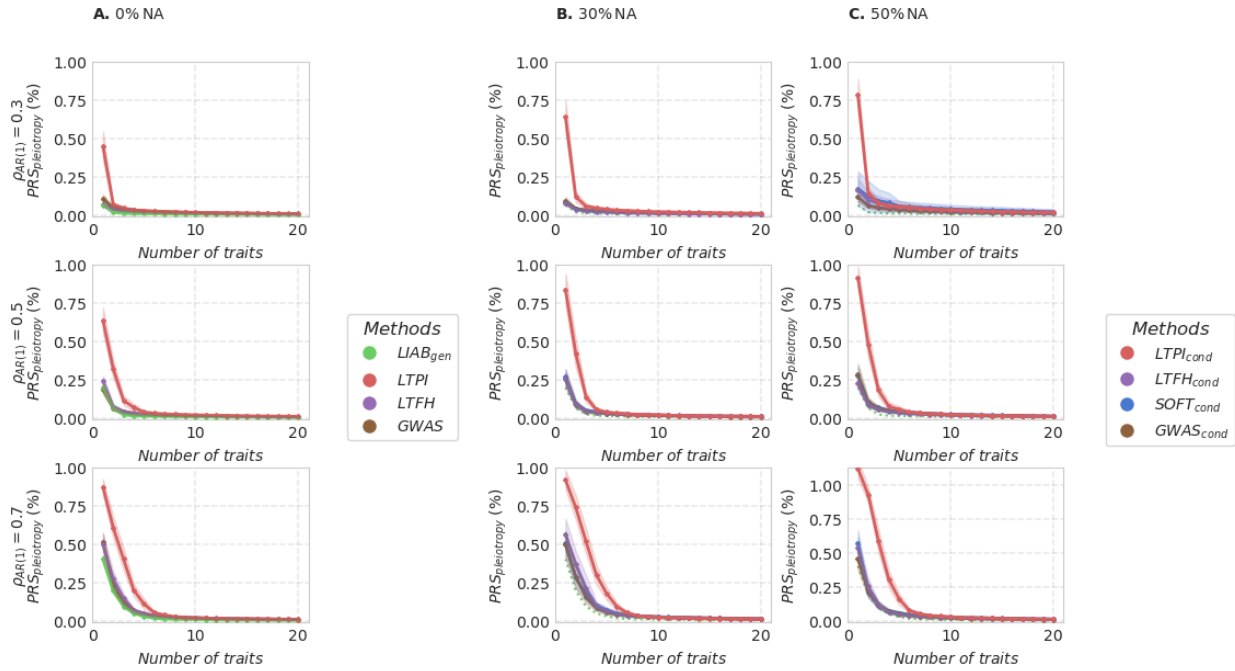


Fig. S1. PRSpleiotropy in default simulations. We evaluated average PRSpleiotropy (%) under default settings, varying the rate of missing target phenotypes (**A.** 0%, **B.** 30%, and **C.** 50%) and $\rho_{AR(1)}$ values (0.3, 0.5, and 0.7). Each point and shaded region of each line represent the mean and 95% confidence interval, with different colors representing different methods. The dotted green line represents the ideal PRSpleiotropy observed under the genetic liability scale. Each point and width of the shaded region are based on estimates from 25 simulations.

We have also computed the PRSpleiotropy ratio for the real GWAS analyses. All methods showed a similar pattern of PRSpleiotropy values, with *LTPI* having the largest values as expected given *LTPI*'s leverage of pleiotropic signals to enhance power. *LTFH* showed lower PRSpleiotropy values even relative to *GWAS*, indicating enhanced target specificity for *LTFH*.

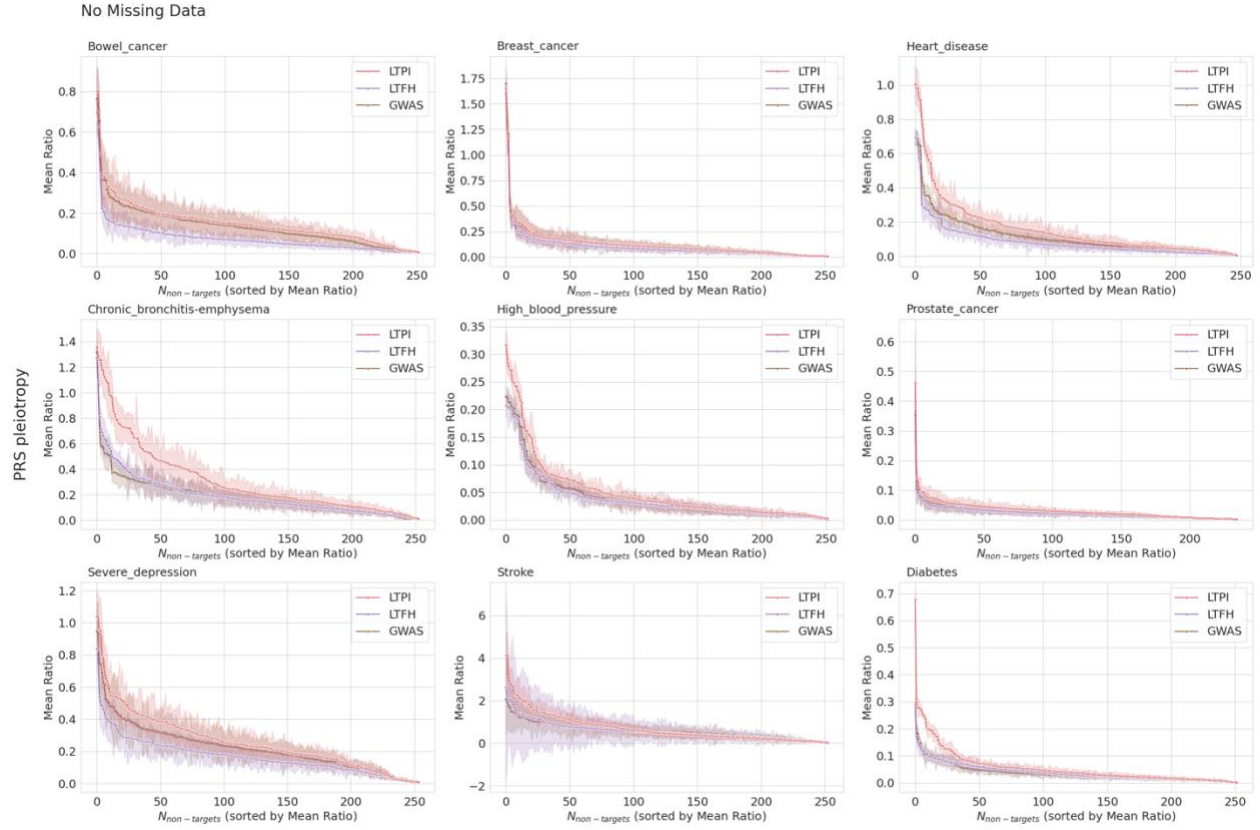


Fig. S2. PRSpleiotropy across nine traits in UKBB-FH analysis with no missing for the target phenotype. Each panel label denotes a target trait. We assessed PRSpleiotropy for three different methods: *LTPI* (red), *LTFH* (orange), and *GWAS* (blue). Each point represents the mean PRSpleiotropy calculated from ten-fold cross-validation. Values were sorted in descending order.

We performed additional analyses by introducing random missingness (30%) to the target phenotype. In this analysis, we also compared with SoftImpute. The PRSpleiotropy patterns remain similar to the scenario with non-missing data, with SoftImpute showing similar performance to *GWAS*. Interestingly, for some diseases such as bowel cancer, *GWAS* has a larger PRSpleiotropy than other methods.

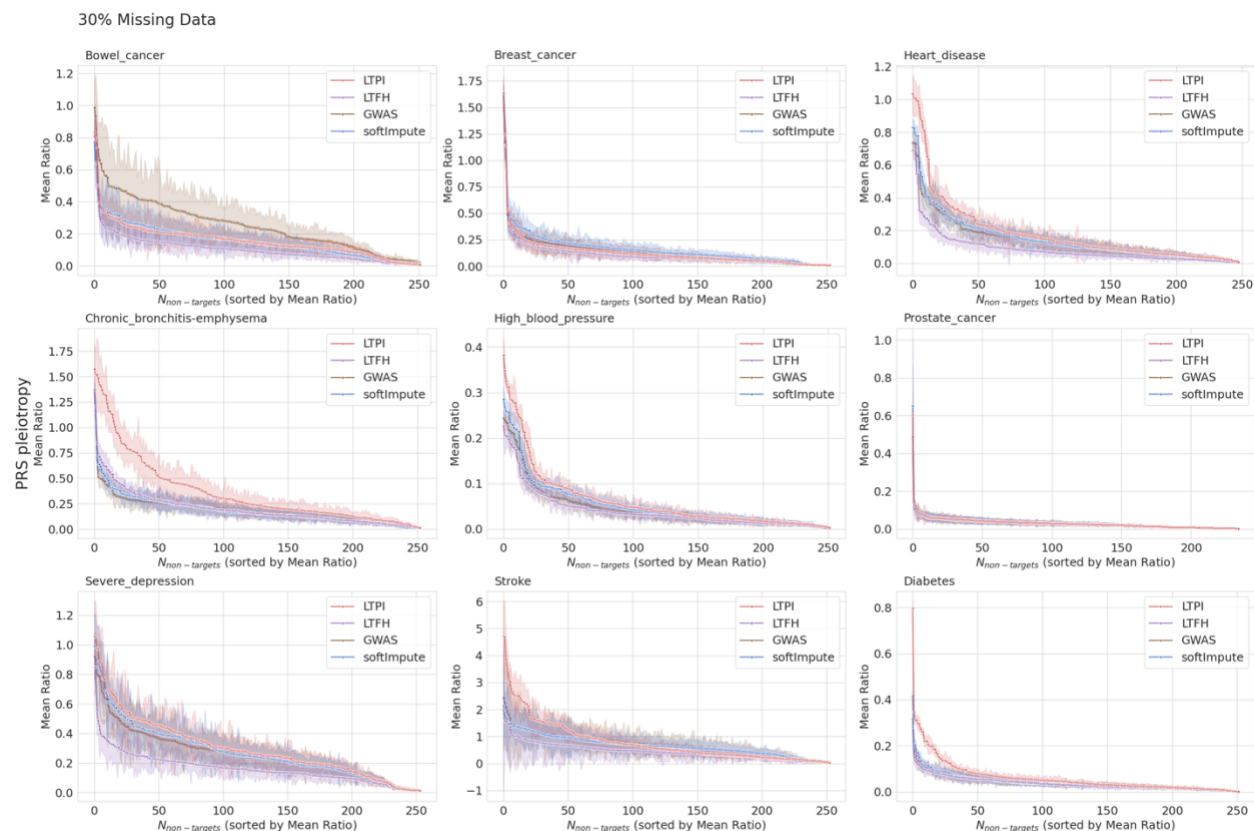


Fig. S3. PRSpleiotropy across nine traits in UKBB-FH analysis with 30% missing for the target phenotype. Each panel label denotes a target trait. We assessed PRSpleiotropy for four different methods: *LTPI* (red), *LTFH* (orange), *SoftImpute* (dark blue), and *GWAS* (light blue). Each point represents the mean PRSpleiotropy calculated from ten-fold cross-validation. Values were sorted in descending order.

R1M3. It is unclear what are the contribution of risk prediction sections due to the **lack of independent gold standard labels** (“Assessing the impact of imputation on prediction accuracy within *LTPI*: eMERGE-CKD analysis”, “Construction of *LTPI* and *LTFH* scores for disease risk prediction”, “Performance assessment of disease risk prediction accuracy”). In all these three sections, **the prediction labels are used to predict liability**, making it hard to interpret the results. It is important to come up with independent gold standard labels, at least for example diseases. eMERGE-CKD analysis uses eGFR to predict liability, while evaluating it on eGFR generated labels; the latter two sections use disease labels (PheCode-FH) to predict posterior liability, while evaluating it on highly correlated PheCodes. The authors aim to compare relative performance to *LTFH*, but this is not the best comparison (as models use different information), and the labels are used as a predictor. I suggest for all these analyses to exclude the target disease label and highly correlated traits to focus on the improved prediction power from the not clearly related non-target traits (i.e. excluding non-target traits that clearly denote the same disease).

Response: Thank you for this comment. We acknowledge that there was some circularity in our analyses. Our initial idea was to show that we can obtain an accurately predicted phenotype based on selecting features (using ATSA or manual selection). But to avoid this circularity, we have now removed eGFR from *LTPI* construction, and have performed some additional analyses by excluding traits that were likely to be a result of the CKD diagnosis, as advised by our clinical collaborator, Dr. Kiryluk. See Figure below (and more on page 9 of the manuscript).

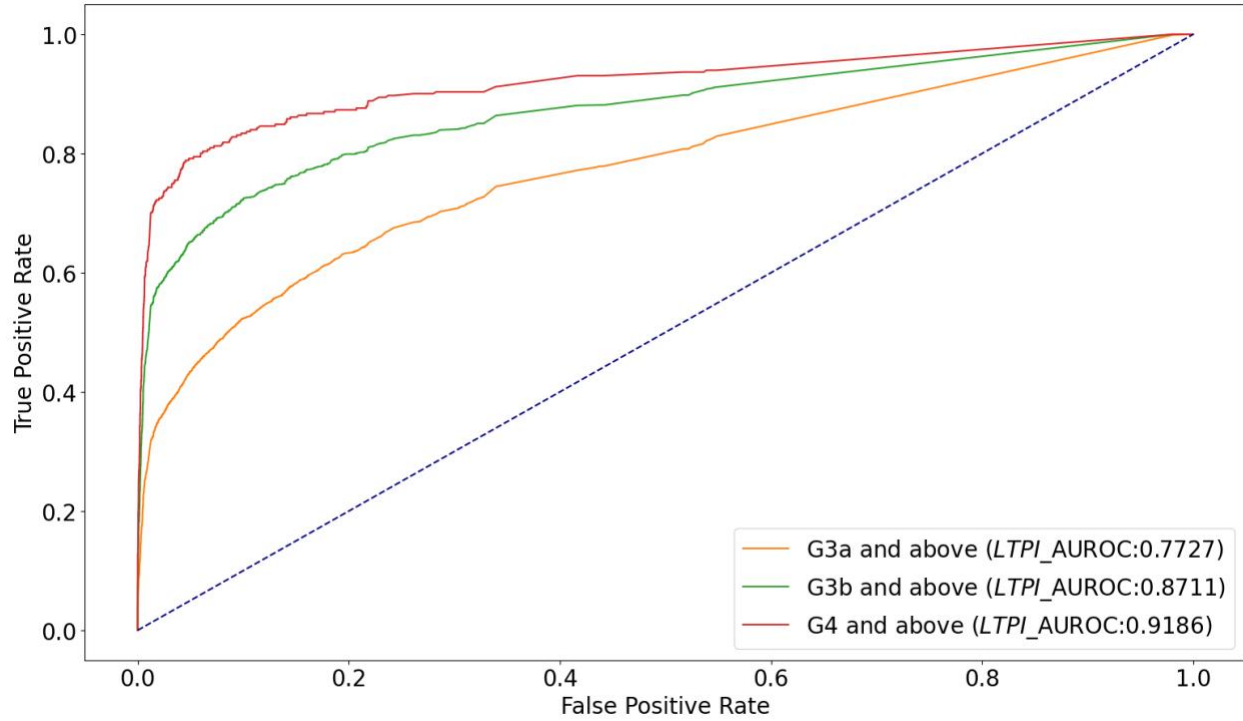


Fig. S4. Validation of risk prediction performance in eMERGE-CKD analysis using a stringent trait selection criterion. LTPi analysis results of eMERGE-CKD with the target phenotype (CKD) and 32 non-target binary traits (*Manual*₃₂). The ROC curves show LTPi scores predicting CKD stages ('G3' to 'G4') in G-STAGE, with stages color-coded. The diagonal navy line represents no discrimination. AUC values in the legend indicate prediction performance.

Below, we present LTPi results on eMERGE-CKD without using eGFR measurements. Overall, LTPi is quite robust, maintaining similar (although slightly worse) performance after removing eGFR. This figure is now included in Fig. 5 in the manuscript.

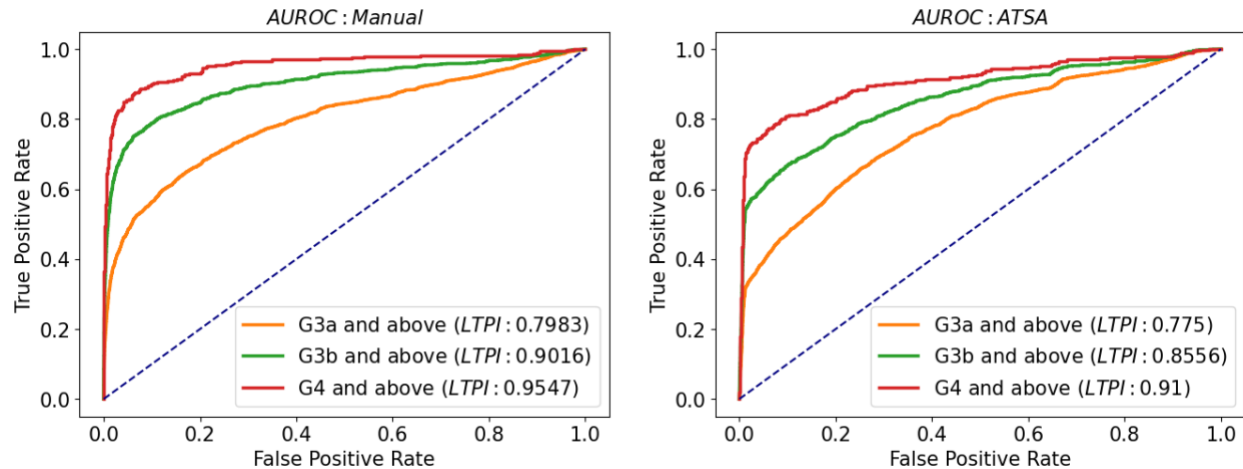
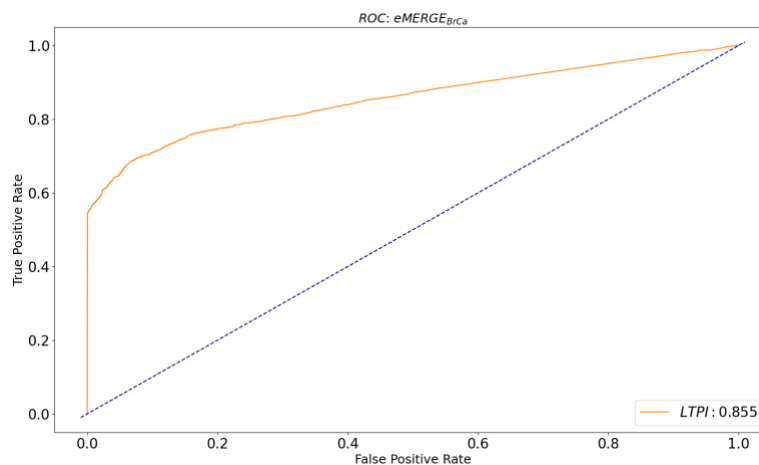


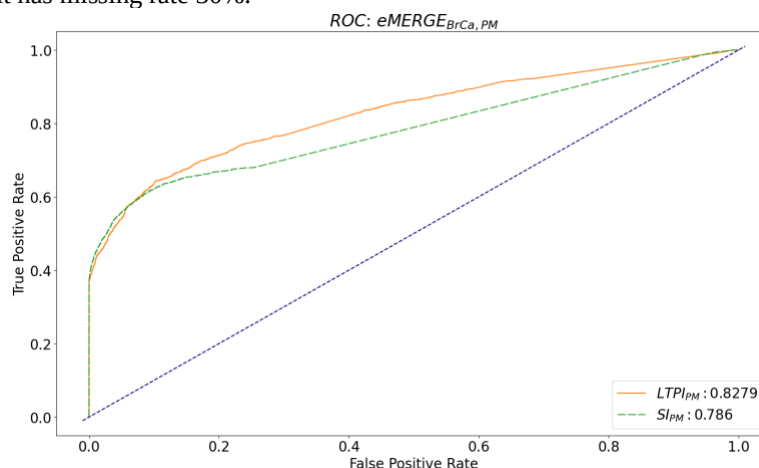
Figure. ROC curves for LTPi scores in predicting CKD G-stages ('G3' to 'G4') in G-STAGE. Left: LTPi analysis results using manual selection of non-target traits (*Manual*). Right: LTPi analysis results using the Automatic Trait Selection Algorithm (*ATSA*). Stages are color-coded. The diagonal navy line represents no-discrimination. AUC values in the legend indicate prediction performance.

To address the comment on using gold standard labels: we tried to find such phenotype labels, but we found it very difficult to get sufficient gold standard phenotype data for the datasets we used in applications. For UK Biobank, such information is not available. We did, however, conduct an additional analysis on breast cancer using the eMERGE dataset (eMERGE-BrCa), where we utilized the eMerge-III BrCa phenotypes, a phenotype definition developed by BrCa experts and validated by manual chart reviews according to the phenotype validation protocol of eMERGE-network (the same protocol was used to validate the CKD phenotype). Although these are not gold-standard labels, they represent labels from state-of-the-art phenotyping algorithm for breast cancer that has been validated against manual chart reviews. We selected 50 non-target traits using ATSA (we excluded all BrCa-related PheCodes from ATSA analysis, including the target PheCode "174"). We derived *LTPI* scores on 36,478 European individuals. More details are in the manuscript (page 7). We have also compared the prediction accuracy with that of SoftImpute under the three scenarios above, and we included these results as supplementary figures (**Fig. S7** and **S8**). The results are below.

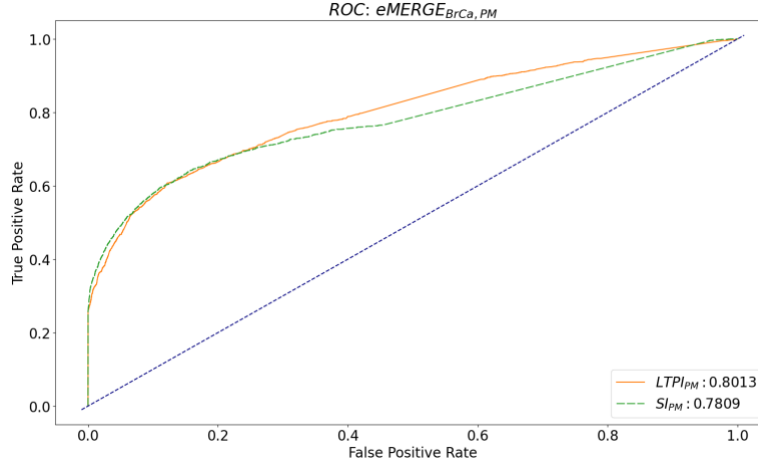
- (1) The target trait is coded as 1/0 if the corresponding target disease phecode is present/absent. In this case, the imputation method SoftImpute is not applicable as there is nothing to impute.



- (2) The target trait has missing rate 30%.



- (3) The target trait has missing rate 50%.



Minor comments/questions:

R1m1. “non-diagonal elements represent covariances due to non-genetic factors.”... “Similarly, the environmental covariance matrix Σ can be estimated in LDSC.” Please clarify how you compute non-diagonal elements of the environmental covariance matrix using LDSC?

Response: Thank you for this comment. Below please find the clarification which we have added to the manuscript on page 17.

“Estimating C_e and Σ ”

In the LDSC output, the “Intercept” value in the “Genetic Covariance” section can serve as a measure for environmental correlation^{32,60}. An environmental correlation matrix (C_e) is constructed by estimating this intercept for all possible pairs of traits.

Next, the environmental correlation matrix (C_e) is converted into an environmental covariance matrix (Σ) using the heritability estimates (h^2). Specifically, the transformation is performed as follows:

$$\Sigma = \text{diag}(\sqrt{1 - h^2}) \cdot C_e \cdot \text{diag}(\sqrt{1 - h^2})$$

where $\text{diag}(\sqrt{1 - h^2})$ is a diagonal matrix containing $\sqrt{1 - h^2}$ for each trait along its diagonal.”

R1m2. “Maximum Likelihood Estimator (MLE; $LTP1'$) for the target disease’s genetic liability, leveraging additional continuous non-target phenotypes.” Please explain the rationale for using posterior mean in $LTP1b$ but use MLE for $LTP1c$; why not just get a posterior mean estimate using both continuous and binary non-target traits (when both are available)?

Response: We use the Bayesian approach to estimate the posterior mean of genetic liability when integrating binary traits. After estimating the genetic liability’s posterior mean for target traits, we combine it with observed continuous non-target traits for combined analysis. We assume these continuous traits to follow a normal distribution (possibly after transformation). When both genetic liabilities and continuous traits follow normal distributions, the Maximum Likelihood Estimation (MLE) method is a natural approach given MLE’s statistical efficiency with normally distributed data. The posterior mean converges to the MLE when the sample size is large, and the prior distribution is uniform. Therefore, both methods should result in similar estimates with large sample sizes and uniform priors.

R1m3. In the methods section of ATSA: “where R represents the variance of genetic liability on T explained by $\mathbf{O}$ ”; based on the derivation, here ROT represents the proportion of genetic variance that are explained (which is different from the “variance of genetic liability”).

Response: Thank you for this comment. We have corrected the language in the text on page 18.

“ R_{OT}^2 represents the proportion of the variance in genetic liability for T that is explained by $\mathbf{O}$; $1 - R_{OT}^2$ describes the residual heritability of the target trait.”

R1m4. “We assumed no correlations in environmental liability between traits” Environment liability is independent from genetic liability across individuals for a given trait; however, environment liability between traits is correlated. It is worth testing whether the model is robust with strong environmental correlation between traits.

Response: Thank you for the suggestion. We have modified our simulations to now include environmental correlations among traits (see section “Validation of *LTPI*’s performance using simulations” on page 4) and we present results under this assumption.

R1m5. It is unclear from Figure 2 what is the number of binary vs. quantitative traits in the red dots (mix of binary and quantitative).

Response: We have revised the analyses depicted in Figure 2 and updated the figure caption to provide more details about the number of traits used in the simulation setting. See next response.

R1m6. Figure 2: what is the observed upper limit of R^2 between binary observation and *LIABgen*? If that observed R^2 between binary and simulated liability value is transformed to liability scale, what is the liability-scale R^2 (equation 8 in Lee et al. 2012 Genetic Epidemiology)? The figure doesn’t present the advantage of transforming observed binary traits to a predicted liability as we have been using liability-scale R^2 /heritability; it just shows the prediction gets better as the model includes more genetically correlated predictors. I think it is more important to clarify why we want continuous liabilities instead of just using the binary labels (for R^2 , we already have the needed tools to transform from binary scale to liability-scale, what is the added value?).

Response: Figure 2 compares the prediction performance when including more binary traits and the value of adding quantitative traits. Note that we have revised the simulations as explained in the paper. We now assume a more realistic AR(1) structure, which makes the effect of adding more traits less obvious than before when we were assuming constant correlation values for additional traits. We show that including quantitative traits in addition to binary traits can significantly improve performance; nonetheless, binary traits are more readily available, and we show that including an increasing number of binary traits can help performance. Continuous liabilities, as inferred by *LTPI* based on **additional** non-target traits in EHR, can provide more information for disease risk prediction and for GWAS analyses beyond a simple transformation from the binary scale to the liability scale. We have also added the R^2 between the target binary trait and *LIABgen* to the plot.

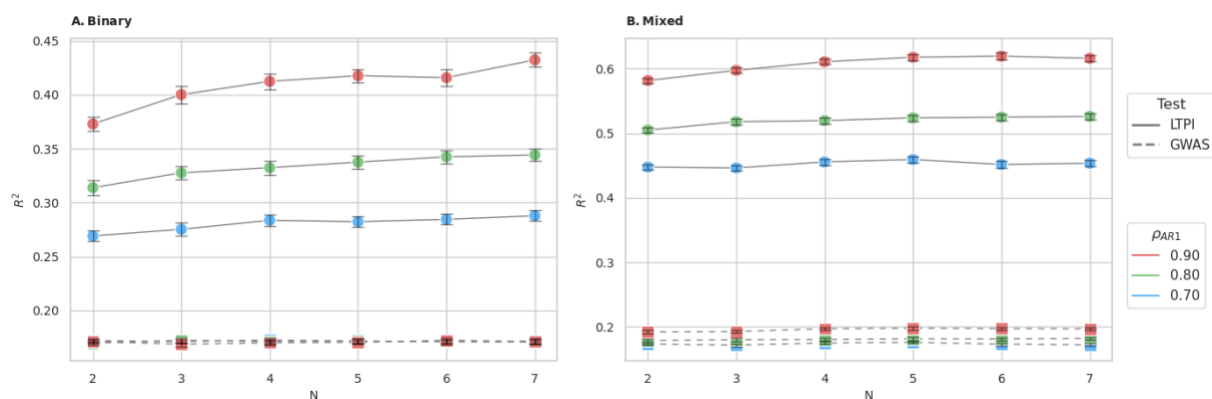


Fig. 3. R^2 in relation to *LIABgen* based on trait count. *LTPI* analyses were conducted with: **A.** A model aggregating 3N binary non-target trait. **B.** A model combining 3N binary and five continuous non-target traits. The x-axis shows the trait count ($N = \{2, 3, 4, 5, 6, 7\}$). Points connected by lines represent the mean R^2 from 100 replications, with black error bars indicating 95% confidence intervals. Trait correlations were generated using AR(1) structure with $\rho_{AR(1)}$: 0.7 (blue), 0.8 (green), and 0.9 (red). We estimated R^2_{LTPI} (Circle) and R^2_{GWAS} (Square) for *LIABgen*.

R1m7. It should be clarified in simulation whether the heritability is on observed scale or liability scale (I am guessing that it refers to the latter in simulations from the text).

Response: Indeed, it is on the liability scale. We have clarified this aspect in the manuscript (section “Validation of LTPI’s performance using simulations”).

R1m8. In “Factors affecting LTPI’s GWAS performance”: “due to LTPI’s lower genetic correlation ($\rho=0.1$) against LTFH’s inter-family genetic correlation of 0.5. However, as genetic correlation among different traits increased to 0.3, 0.5, and 0.7, LTPI gained the upper hand”. If there are 10 additional traits used as predictor, each has correlation 0.5 to the target trait, it is unfair to compare to family history which is only one additional variable. A fair comparison should report “variance of genetic liability” (defined in ATSA part) of ten binary traits. The decreased performance could be caused by noise induced when mapping from non-target binary traits to liability.

Response: Thank you for your comment. We agree that the initial comparison was unfair; however, it was intended to illustrate the potential of leveraging additional traits (for instance, when family history is unavailable, or even when it is available) to enhance performance.

For a fairer comparison, we conducted additional *LTPI* analyses using 3 binary traits (*LTPI_{FH}*), which utilized the same phenotype matrix and genetic correlation structure as the *LTFH* analysis using parent-offspring trio. Our results indicate that in this scenario, *LTPI* and *LTFH* exhibit similar performance, as expected (see figure below). However, a key advantage of *LTPI* is its ability to leverage information from numerous binary and quantitative traits.

We included these results in section “B” of **Figure 4** in the manuscript, in which we investigated the factors affecting *LTPI*’s GWAS performance.

More detailed simulation results and settings are provided on page 6 of the manuscript.

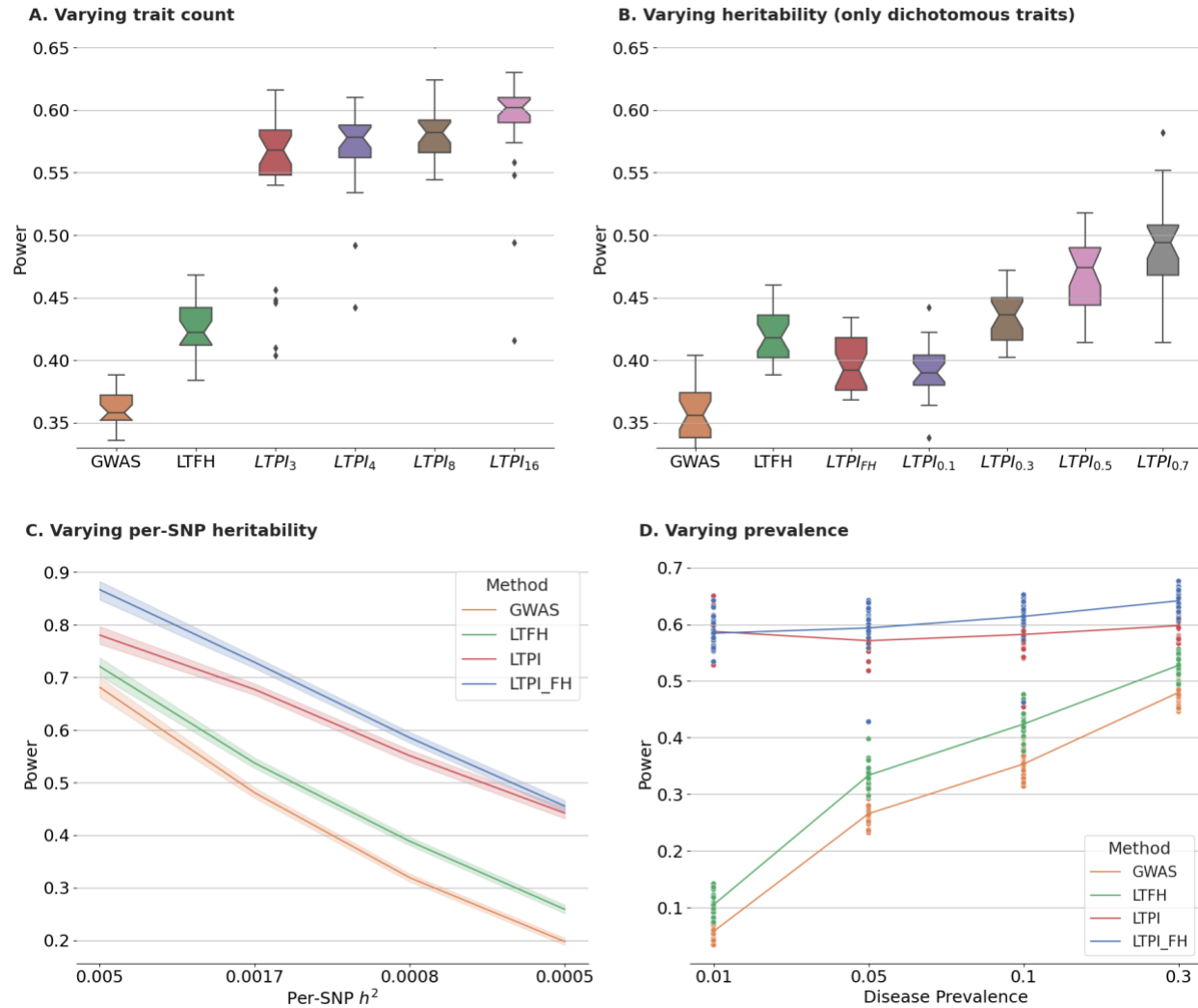


Fig. 4. GWAS power comparisons across various simulations. **A.** Power comparison for varying trait counts. **B.** Power comparison for varying heritability. We include dichotomous traits to the simulation. **C.** Power comparison for varying per-SNP heritability. **D.** Power comparison for varying disease prevalence. Boxplots represent the average power from 25 replicates, with whiskers extending to 1.5 times the IQR. Lines illustrate the average power over these replicates, and error bars showcase the 95% confidence interval for power estimates.

R1m9. “As disease prevalence varies between 0.01, 0.05, and 0.1, LTPI maintains roughly 43% power.” Is this just varying target disease prevalence?

Response: We have revised the text and added clarifications in section “Factors affecting LTPI’s GWAS performance” on page 6. The disease prevalence is varied for both target and non-target traits.

R1m10. “GHK” approximation. It is unclear from the main text how GHK method is used (e.g. no mention of it in the “Overview of LTPI construction”); it would help the reader to understand the rationale of choosing an approximated method instead of sampling from the exact posterior distribution. Also, in the main text, GHK is mentioned to compute only posterior mean and variance but not approximate the posterior distribution, while later it is mentioned the estimate is “skewed”; this should be clarified.

Response: We apologize for the imprecise language. It is assumed that the posterior distribution is a normal distribution and therefore the mean and variance are enough to specify the distribution. We have revised the text and added more details in Step 1 of the LTPI framework on page 3:

“To approximate the posterior mean genetic liability, we use an importance sampling method called the GHK (Geweke, Hajivassiliou and Keane) algorithm³⁵⁻³⁷, which can provide very accurate approximations for conditional moment problems with multivariate Gaussian distributions.” Also, in the Methods section, we mention the rationale of using GHK instead of sampling from the exact posterior distribution.

R1m11. If eGFR is used as a continuous non-target trait for prediction liability, wouldn't G-stage cases (defined by eGFR) be highly correlated with the liability? This looks like a circular evaluation.

Response: Yes, we agree with the reviewer's point. Indeed, *LTPI* with eGFR and GSTAGE have a correlation of 0.582, higher when compared to 0.533 for the correlation between *LTPI* derived without eGFR and GSTAGE. Please see our response to next question for *LTPI* results without using eGFR.

R1m12. Connected to the previous point, (Figure 5) comparing the eGFR defined case-control with the score (liability) predicted using eGFR doesn't make much sense. Figure 5CDIJ is more informative than Figure 5ABGH.

Response: We appreciate this suggestion. We removed *LTPI* scores using eGFR from Figure 5. The new results are shown below.

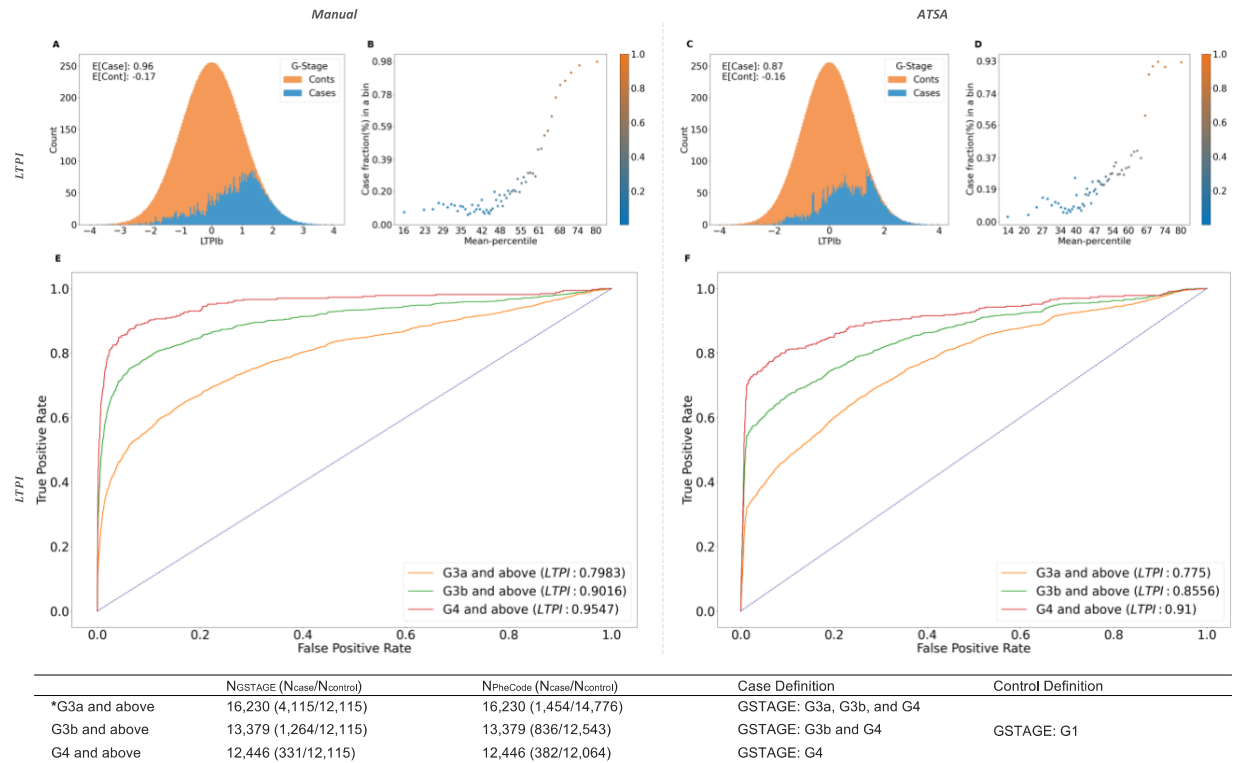


Fig. 5. *LTPI* analysis results for eMERGE-CKD. Left. *LTPI* analysis results using manual selection of non-target traits (*Manual*). Right. *LTPI* analysis results using the Automatic Trait Selection Algorithm (*ATSA*). **A, C:** Histograms of *LTPI* score distributions, segmented by PheCode classification. The y-axis shows the count of ranked *LTPI* scores. **B, D:** Scatter plots displaying case fractions (%) across mean-percentile bins. Each point's coordinates represent its mean percentile (x-axis) and case fraction (y-axis). **E, F:** ROC curves for *LTPI* scores in predicting CKD stages ('G3' to 'G4') in G-STAGE. Stages are color-coded. The diagonal navy line represents no-discrimination. AUC values in the legend indicate prediction performance. The accompanying table lists the sample sizes and definitions of case-control status used in the analysis. Each line in the AUROC corresponds to a row in this table. An asterisk (*) indicates the G-stage definition used to generate the histograms and scatter plots. *LTPI* scores were standardized using rank-based inverse normal transformation (RINT).

R1m13. Figure 5E: the Phecodes (binary)--G-stage trait (binary) AUC is not well defined (it is just a single number of FPR and TPR) and thus not comparable to liability--G-defined traits; it is more comparable to compute the liability-scale correlation of PheCodes--G-stage vs liability--G-stage.

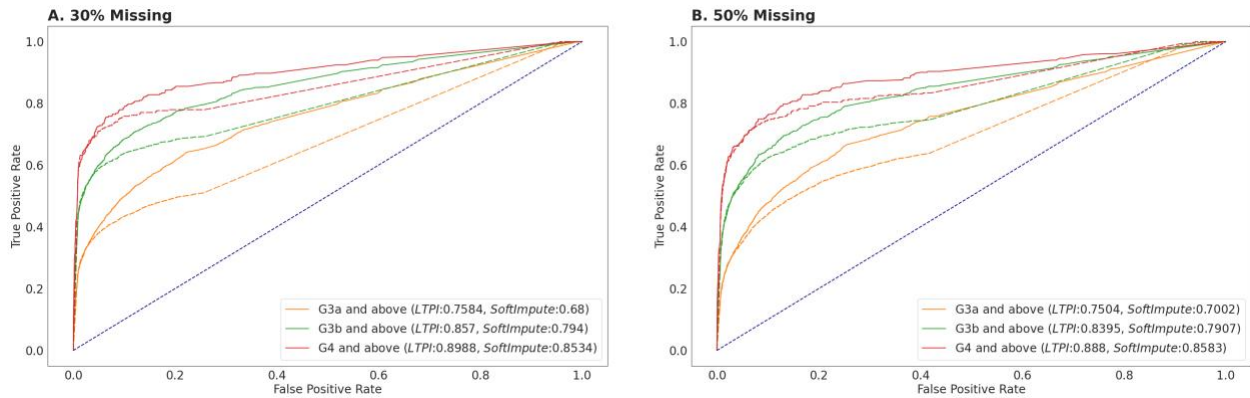
Response: Thank you for the comment. We have removed the phecode from the AUROC analyses in Fig. 5. and now focus on comparisons between *LTPI* and other methods that provide continuous scores (such as *LTFH* and *SoftImpute*). Additionally, we evaluated the Nagelkerke R^2 between G-STAGE and PheCode and G-STAGE and *LTPI* scores. These results are included in Table S6, as follows:

	$N_{G\text{STAGE}}(N_{\text{case}}/N_{\text{control}})$	$Prev_{\text{sample}}$	R^2 (SE)		
			<i>PheCode</i>	<i>LTPI</i> _{Manual}	<i>LTPI</i> _{Manual32}
G3a and above	16,230(4,115/12,115)	0.254	0.252(0.008)	0.325(0.01)	0.303(0.008)
G3b and above	13,379(1,264/12,115)	0.094	0.419(0.015)	0.504(0.015)	0.458(0.014)
G4 and above	12,446(331/12,115)	0.027	0.487(0.024)	0.589(0.024)	0.515(0.024)

Table S1. Prediction accuracy in R^2 for *LTPI* and PheCode against G-stage in eMERGE-CKD analysis. The table presents the risk prediction accuracy in Nagelkerke R^2 . The number of cases and controls, sample prevalence, and the corresponding R^2 values are also reported. We evaluated standard error (SE) of the estimated R^2 using bootstrap with 100 iterations.

R1m14. “Assessing the impact of imputation on prediction accuracy within *LTPI*: eMERGE-CKD analysis” connected to major comments, it is interesting to see whether directly using *SoftImpute* to predict liability has comparable performance to the proposed model.

Response: We removed this analysis since we no longer include eGFR in the *LTPI* construction to avoid the above mentioned circularity. We have implemented *SoftImpute* in the GWAS studies, as we explained in Major comment 2 (R1M2). We have performed some additional analyses assuming missing data for the target phenotype and find that *LTPI* outperforms *SoftImpute* in terms of AUROC values.



AUROC comparison between *LTPI* and *SoftImpute* in eMERGE-CKD analysis. *LTPI* and *SoftImpute* results using *Manual32* with different rates of missing target phenotypes: **A.** 30% missing and **B.** 50% missing. ROC curves show *LTPI* scores predicting CKD stages ('G3' to 'G4') in G-STAGE, with stages color-coded. AUC values in the legend indicate prediction performance.

Reviewer #2:

The manuscript presents the Liability Threshold Phenotypic Integration (LTPI) method, aimed at improving disease phenotype classification by integrating genetic and phenotypic data. It introduces LTPI as a novel approach, while also pointing out the need for further clarification, validation, and comparison with current methodologies.

Strengths of the Manuscript:

** LTPI is lauded for its innovative approach to phenotype classification, potentially surpassing GWAS and family history methods by incorporating multi-trait liability thresholds and genetic correlations.*

** The manuscript is praised for its thorough explanation of the LTPI algorithm, including the rationale behind it and the methodological rigor in its development and assessment.*

** The adoption of importance sampling for estimating genetic liability and an automatic trait selection algorithm are highlighted as significant advantages, promising greater power and accuracy in prediction within large-scale biobanks.*

Response: Thank you for the helpful comments and summary. We address below point-by-point the individual comments.

Major Concerns:

R2M1. *There are concerns about the control of false positives and the real-world applicability of LTPI, especially in scenarios utilizing only single-modal images or when applying the method to independent cohorts in Stage 1 and Stage 2.*

Response: Indeed, controlling false positives is essential for all methods that use non-target traits to predict target traits. We should emphasize that the principle behind the proposed method is to leverage genetically correlated traits to improve GWAS power, and, in particular, *LTPI* will borrow information at pleiotropic loci from other traits to improve the power at those loci for the target trait (See power figure below). Even though those loci may not be specific to the target trait (and may be shared with other traits), there is great advantage in borrowing information from other traits to detect such loci. We also note that given our ATSA algorithm to select traits that are genetically correlated to the target trait, there is no reason for *LTPI* to be more susceptible to false positives relative to recently proposed imputation methods such as SoftImpute (Dahl et al. Nat Genet 2023) that also leverage correlations among phenotypic features but cannot separate genetic from environmental correlations.

We have done some additional analyses, both simulations and GWAS analyses to investigate power and potential for false positives.

For the real GWAS data, we have updated the results in Figure 9 (Figure 8 in the current manuscript) to include some of the largest GWAS studies for the four diseases, and we show that with these larger sample sizes, the sign consistency between *LTPI* discovered loci and validation GWAS is substantially improved, and very high indeed (see Figure below), suggesting that the newly detected loci are likely risk loci for the target disease, although these loci may be of lower effect and missed in the original GWAS. The power of *LTPI* is higher because it borrows information from other diseases and traits at pleiotropic loci.

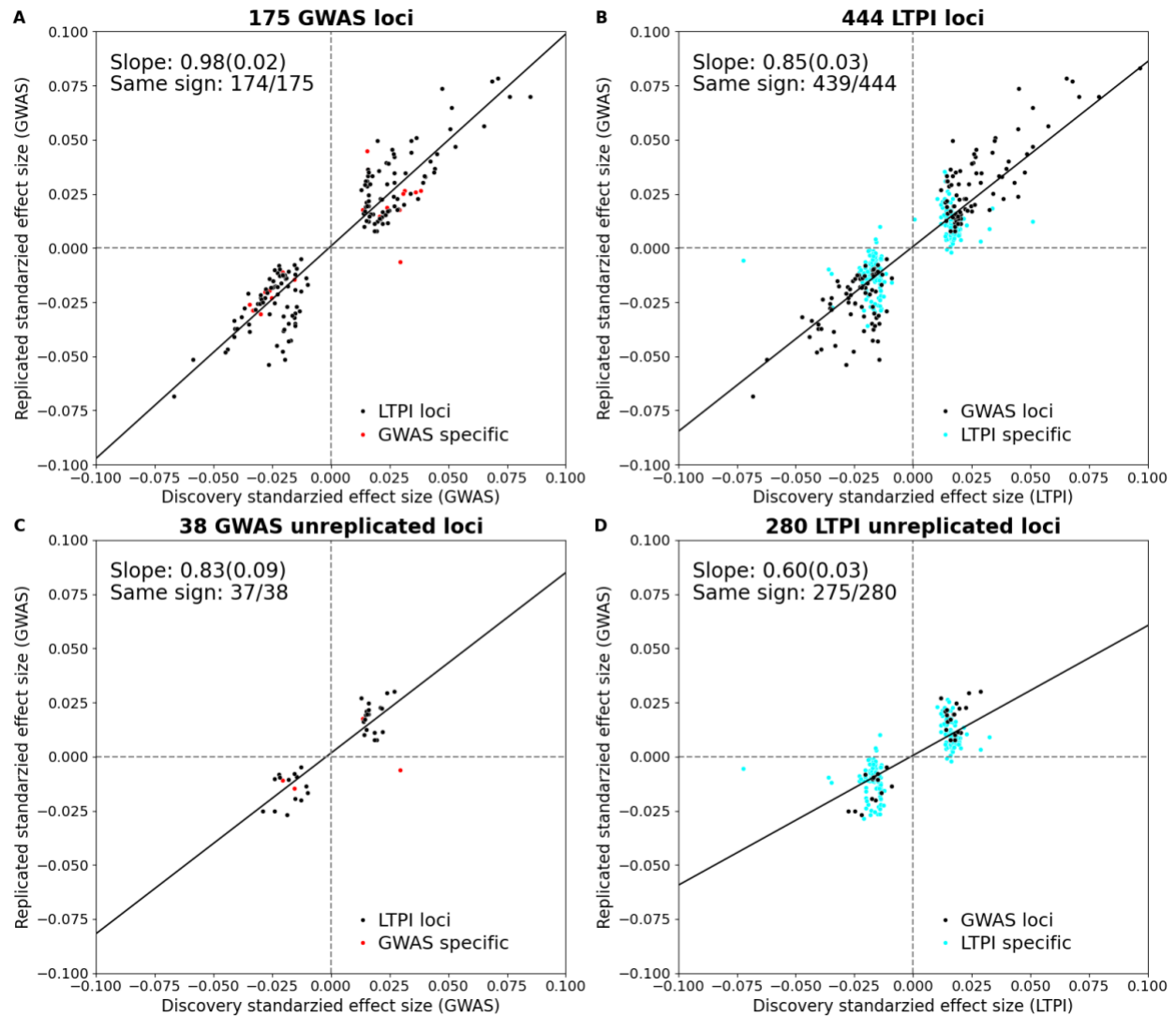


Fig. 6. Replication analysis using GWAS validation data sets. We generated scatter plots of standardized effect sizes ($Z/\sqrt{N_{\text{eff}}}$) in the validation GWAS data vs. the UK Biobank discovery data, aggregated across 4 diseases (CAD, T2D, breast cancer, and prostate cancer) for **A**, 175 loci identified by GWAS, **B**, 444 loci identified by LTPI, **C**, 38 loci identified by GWAS but not replicated by the validation GWAS data, and **D**, 280 loci identified by LTPI but not replicated by the validation GWAS data. Each point indicates the lead SNP at an independent locus, with the color indicating whether the locus contains genome-wide significant variants identified by both LTPI and GWAS.

In simulations we show that our method controls the type 1 error (Fig. S2). We have also used the PRSpLeiotropy measure in Dahl et al. to assess the specificity of the GWAS findings for the different methods first in simulations and then in the real data applications. However, a higher PRSpLeiotropy can simply reflect higher pleiotropy among the target and non-target traits. The simulations below represent this scenario as the causal variants are shared across the target and non-target traits, and therefore, there are no false positives but genuine sharing of genetic effects among the different traits.

Our simulations evaluated the PRSpLeiotropy for the competing methods in three scenarios (no missingness, 30% or 50% missingness for the target phenotype). We found that, indeed, *the PRSpLeiotropy increases as we increase the level of correlation among genetically correlated traits with the target trait*, therefore a higher PRSpLeiotropy in itself does not imply higher rate of false positives. It is a reflection of the fact that LTPI leverages those genetically correlated non-target traits to build the resulting score. Although we expect LTPI to have higher PRSpLeiotropy relative to methods that do not borrow information from non-target traits (such as LTFH and GWAS), the higher PRSpLeiotropy

for *LTPI* vs. *SoftImpute* can appear surprising at first, as *SoftImpute* is also borrowing information from other traits by leveraging the underlying correlations.

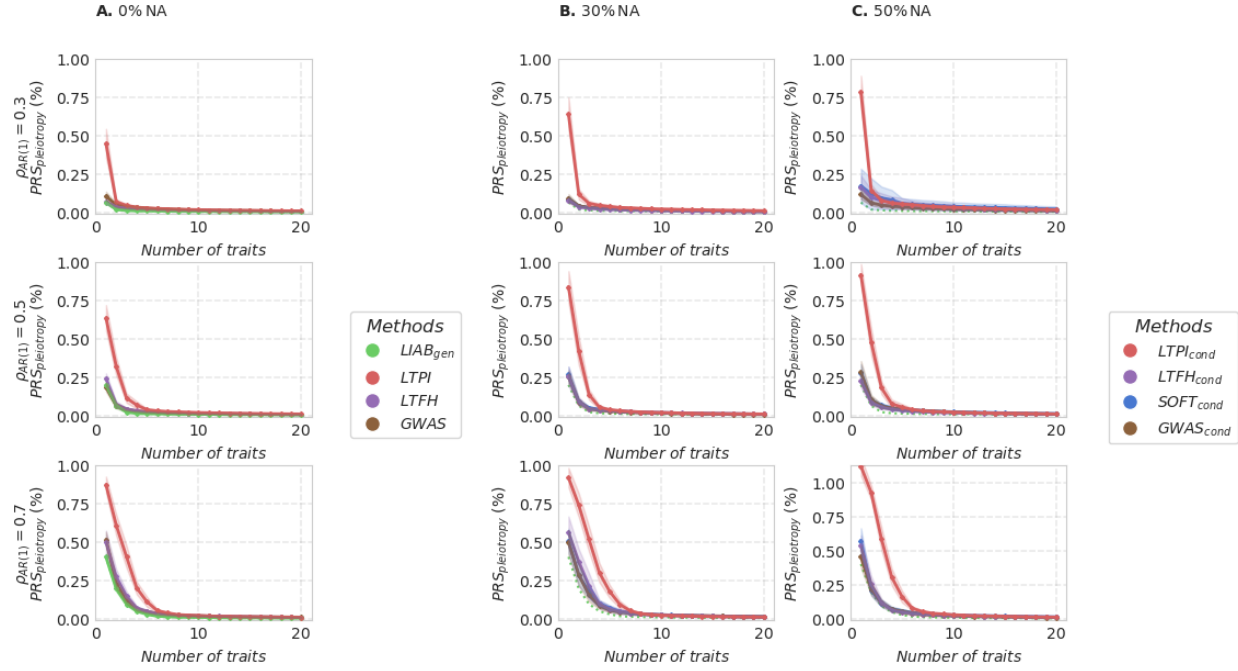


Fig. S5. PRSpleiotropy in default simulations. We evaluated average PRSpleiotropy (%) under default settings, varying the rate of missing target phenotypes (A. 0%, B. 30%, and C. 50%) and $\rho_{AR(1)}$ values (0.3, 0.5, and 0.7). Each point and shaded region of each line represent the mean and 95% confidence interval, with different colors representing different methods. The dotted green line represents the ideal PRSpleiotropy observed under the genetic liability scale. Each point and width of the shaded region are based on estimates from 25 simulations.

However, as shown in the power plot below, *SoftImpute*'s power is similar to *GWAS* in these simulations (due to its poor imputation performance when dealing with binary data, especially with imbalanced case:control ratio as in biobanks) which explains the lower PRSpleiotropy relative to our method.

We are not sure we understand the issues with “single-modal images”, maybe this is an auto-correct issue? Stages 1 and 2 of the *LTPI* are supposed to be applied to the same set of individuals, although it is possible that some individuals may have missing data for some quantitative traits in Stage 2. Nonetheless, *LTPI* will treat that as missing data and proceed to estimate the posterior mean genetic liability.

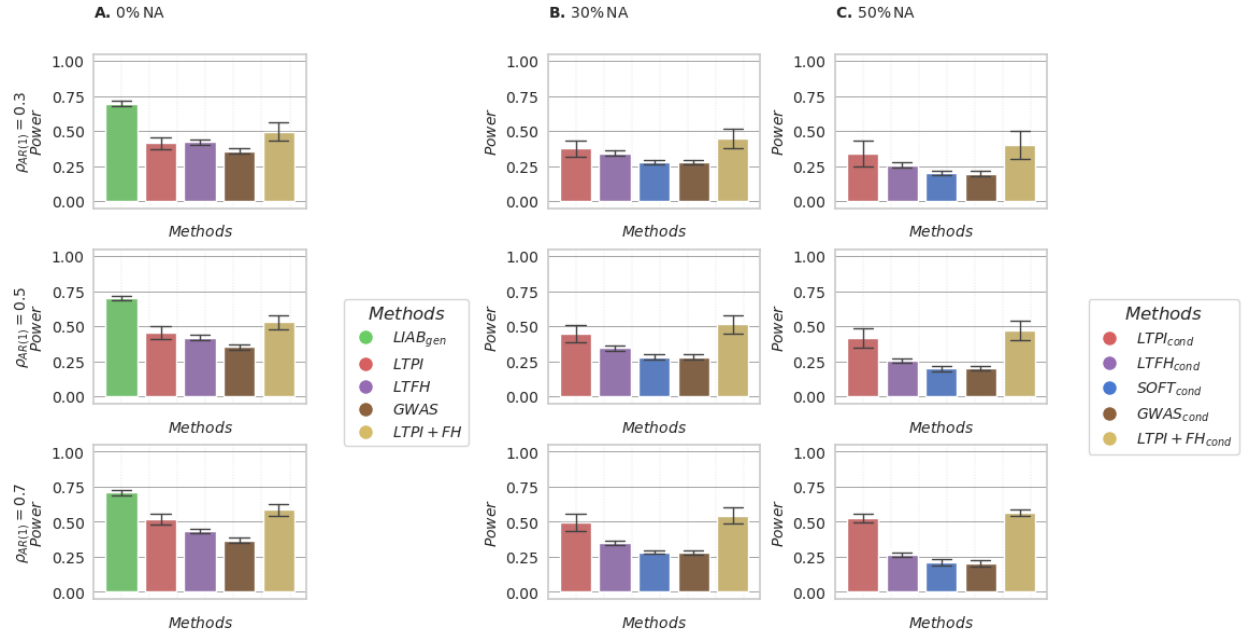


Fig. 7. Power simulation. We conducted simulations under default settings, varying the rate of missing target phenotypes (A. 0%, B. 30%, and C. 50%) and the $\rho_{AR(1)} = \{0.3, 0.5, 0.7\}$. Each bar represents the average power from 25 iterations, defined as the proportion of causal SNPs with $p\text{-value} < 5 \times 10^{-8}$, and the black error bars represent the 95% confidence interval.

R2M2. Additional simulations or validations are deemed necessary to explore the impact of assumptions being violated, especially considering the reliance on EHR data, which may not accurately reflect the population's comorbidity and prevalence.

Response: Thank you for the comment. We assessed $LTPI$'s power under default settings ($LTPI$) and the power under misspecification ($LTPI_X$, where X denotes the value of misspecification). We applied the same changes to both target and non-target traits. Specifically, we changed the heritability to $\{0.3, 0.5, 0.7\}$, prevalence to $\{0.05, 0.1, 0.3\}$, and correlation $\rho_{AR(1)}$ to $\{0.3, 0.5, 0.7, 0.9\}$ in turn and evaluated power under misspecification. Default values are 0.1 for the prevalence, 0.5 for the heritability and $\rho_{AR(1)} = 1$ for genetic correlation among traits. Results are below. Overall, mis-specifying the prevalence and heritability has limited impact (compare $LTPI$ vs. $LTPI_X$), but mis-specifying the genetic correlations can have a significant impact, highlighting the importance of accurate correlation estimation.

	$LTPI$	$LTPI_{0.3}$	$LTPI_{0.5}$	$LTPI_{0.7}$
Power(%)	62.6 (0.44)	62.05 (0.46)	62.62 (0.44)	62.74 (0.42)
$E[\chi^2_{null}]$	1.00 (0.01)	1.00 (0.01)	1.00 (0.01)	1.00 (0.01)
$E[\chi^2_{causal}]$	126.3 (1.88)	122.1 (2.07)	126.3 (1.89)	127.8 (1.8)

Table S2. Impact of heritability misspecification on $LTPI$ power. We assessed $LTPI$'s power under default settings ($LTPI$; shaded) and the power under misspecification of disease heritability ($LTPI_X$, where X denotes the value of misspecification). We applied this change to both target and non-target. Values are % mean power from 25 simulations, with standard errors in parentheses.

	$LTPI$	$LTPI_{0.05}$	$LTPI_{0.1}$	$LTPI_{0.3}$
Power(%)	61.9 (0.46)	61.76 (0.46)	61.92 (0.47)	62.03 (0.48)
$E[\chi^2_{null}]$	1.00 (0.01)	1.00 (0.01)	1.00 (0.01)	1.00 (0.01)
$E[\chi^2_{causal}]$	122.3 (1.5)	121.5 (1.43)	122.3 (1.5)	123.22 (1.63)

Table S3. Impact of prevalence misspecification on $LTPI$ power. We assessed $LTPI$'s power under default settings ($LTPI$; shaded) and the power under misspecification of disease prevalence ($LTPI_X$, where X denotes the value of misspecification). We applied this change to both target and non-target. Values are % mean power from 25 simulations, with standard errors in parentheses.

	<i>LTPI</i>	<i>LTPI</i> _{0.3}	<i>LTPI</i> _{0.5}	<i>LTPI</i> _{0.7}	<i>LTPI</i> _{0.9}
<i>Power</i> (%)	63.35 (0.26)	44.42 (0.57)	49.95 (0.45)	55.5 (0.33)	60.79 (0.3)
$E[\chi^2_{null}]$	1.00 (0.01)	1.00 (0.01)	1.00 (0.01)	1.00 (0.01)	1.00 (0.01)
$E[\chi^2_{causal}]$	126.2 (1.8)	49.98 (1.25)	64.09 (1.49)	83.57 (1.5)	109.1 (1.62)

Table S4. Impact of genetic correlation misspecification on *LTPI* power. We assessed *LTPI*'s power under default settings (*LTPI*; shaded) and the power under misspecification of $\rho_{AR(1)}$ (*LTPI*_{*X*}, where *X* denotes the value of misspecification). Values are % mean power from 25 simulations, with standard errors in parentheses.

R2M3. Questions are raised regarding the practicality of *LTPI* in comparison to existing multi-trait analysis approaches and its ability to address the complexities of pleiotropy and the genetic architecture of complex diseases.

Response: Thank you for this comment. *LTPI* is a phenotype prediction method with distinct advantages relative to existing multi-trait methods. *LTPI* builds on the liability threshold model, a workhorse model in genetics, and leverages many traits from EHR to perform (1) phenotype prediction and (2) improve GWAS power. *LTPI* is also helpful in settings with EHR data where there could be missing data. Therefore, it also has connections to phenotype imputation methods, such as SoftImpute, with which we have already performed comparisons as described in the previous comment **R2M1**.

Relative to existing multi-trait analysis methods, like MTAG (PLEIO, HOPS), which aggregate GWAS summary statistics from multiple sources and provide updated effect sizes for a target trait by leveraging other related traits, *LTPI* calculates the posterior mean of genetic liability at the individual level for the target trait. *LTPI* also implements an automatic trait selection algorithm (ATSA) that carefully assesses the genetic correlation between a target trait and non-target traits, minimizing the risk of false positives that can come from incorporating information from non-target traits that are not genetically related to the target trait. Finally, *LTPI* can also be used to identify individuals who have a missed diagnosis or will develop specific diseases in the future, which is impossible for multi-trait methods based on summary statistics.

With respect to pleiotropy, *LTPI* can capitalize on phenotypic features in the EHR that are genetically correlated to the target trait to improve power and therefore pleiotropy is essential for the improved performance of *LTPI*.

R2M4. A need for further explanation on how *LTPI* handles non-target traits and its comparison with the latest GWAS findings, as well as a deeper discussion on potential biases stemming from cohort characteristic differences between datasets such as the UK Biobank and eMERGE.

Response: *LTPI* calculates the posterior mean genetic liability conditional on the target trait and non-target traits. Although the number of possible non-target traits is large, we use an automatic trait selection algorithm (ATSA) to select a small number of non-target traits that are then incorporated into the *LTPI* framework to derive the posterior mean of genetic liability for the target trait. This is advantageous because it provides a less biased way to identify relevant non-target traits than manual selection based on clinical knowledge and reduces the risk of false positives induced by incorporating non-target traits that are not genetically correlated with the target trait.

For the real GWAS data, we have updated the results in Figure 9 (Figure 8 in the current manuscript) to include some of the largest GWAS studies for the four diseases, and we show that with these larger sample sizes, the sign consistency between *LTPI* discovered loci and validation GWAS is substantially improved, and very high indeed (see Figure below), suggesting that the newly detected loci are likely risk loci for the target disease, although these loci may be of lower effect and missed in the original GWAS. The power of *LTPI* is higher because it borrows information from other diseases and traits at pleiotropic loci.

Although different biobanks can have different cohort characteristics, *LTPI* utilizes information on prevalence and trait heritability that are specific to the cohort. But as with any other study, the generalizability of the results to other individuals or populations will depend on selection bias and how representative the study is of the broader population.

As for the differences between UK Biobank and eMERGE, we have added the following paragraph to the Discussion section.

“Indeed, the two EHR systems used in our applications differ in some fundamental ways. The UKBB is a population-based cohort of individuals aged 40 to 70 in the UK and aims to represent the general population broadly, whereas eMERGE targets a patient demographic aged 20 to 80 in the US, often involving hospital visitors during the collection period. This fundamental divergence in design implies that while eMERGE's patient-based approach may allow for deeper phenotype analyses and improved disease signal detection, it introduces a higher risk of ascertainment bias compared to UKBB's broader population sample.”

Minor Points:

R2m1. Improvements in figure clarity are suggested, particularly in depicting the integration of binary and continuous traits into the LTPI score and the influence of disease type on LTPI's accuracy.

Response: Thank you for the suggestions. We have tried our best to improve the clarity of the figures and legends. In particular, we have revised Figure 1 depicting the overall flow of the LTPI framework as below.

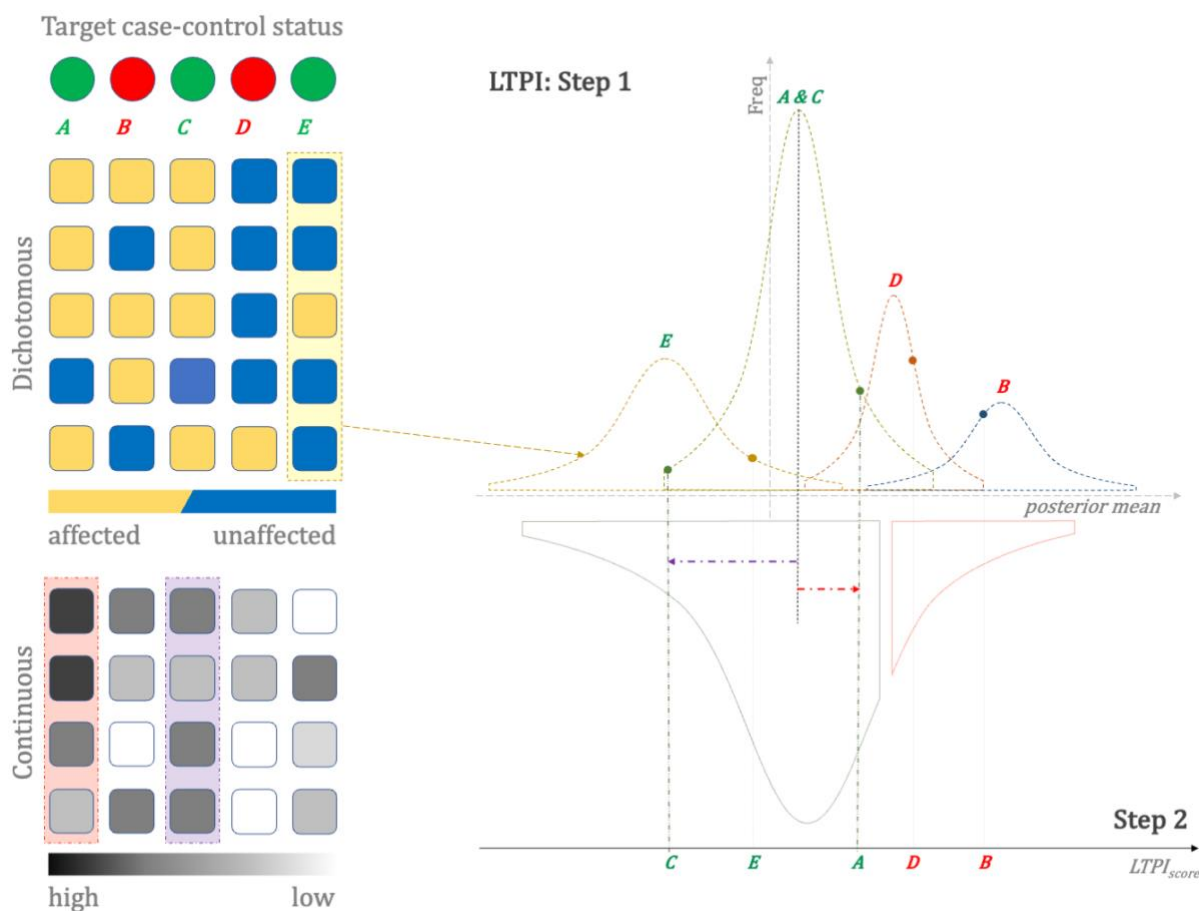


Fig. 8. Summary of the LTPI score construction. The LTPI method uses a two-step approach to aggregate non-target phenotype information and estimate latent disease risk of binary target phenotypes. **Step 1:** Focuses on dichotomous phenotypes (e.g., affected vs. unaffected). It estimates the posterior mean genetic liability and standard error (hence the conditional distribution) per individual. **Step 2:** Integrates continuous phenotypes using a maximum likelihood framework. This step refines the initial estimates from **Step 1** by incorporating continuous phenotype data, such as observed biomarkers or lab measurements (indicated by the red and purple shades on the left). The arrows highlight the effect of adding continuous phenotypes on the final LTPI score of two individuals with the same LTPI score in **Step 1**.

R2m1. A more comprehensive evaluation of LTPI across various disease categories is recommended, along with a discussion on its translational implications for clinical research and precision medicine.

Response: Thank you for the comment. We note that our paper includes analyses of 2 diseases in eMERGE and 9 diseases in UK Biobank, which is much more than other papers (for example, SoftImpute (Dahl et al. 2023) used only one disease (MDD) as illustration). These diseases have different genetic architectures and belong to different categories (cardiovascular, cancers, neurological, and psychiatric diseases). There are direct applications to clinical research and precision medicine. By predicting the posterior mean genetic liability for each individual (based on non-target traits and family history), we can identify individuals who have a missed diagnosis or who will develop a disease in the future. With longitudinal data it will be possible to provide risk predictions at different time points and therefore *LTPI* provides a way to monitor individuals for increased risk to particular diseases. We have added a paragraph on this topic to the discussion.

R2m2. Clarification is sought on the incorporation of family history into the *LTPI* score and the rationale behind the reported high AUC values.

Response: Incorporating family history (whenever available) into *LTPI* is crucial due to its strong genetic correlation with the target disease. The integration of parent-offspring trios into *LTPI* is straightforward. To aggregate family history information, we treat the case-control status of each family member as a non-target trait. For genetic and environmental correlation, the intra-family structure is fixed, meaning the genetic correlation between a parent and offspring is 0.5, while the correlation between parents is 0. This approach uses the offspring's between-trait correlation to estimate the correlation between traits for each family member.

It is true that the AUC values are high. In our analyses, we note that disease cases do have large *LTPI* scores (Figure 5B and 5D in the manuscript and below), so there is good separation between cases and controls based on the *LTPI* scores, as reflected in the high AUROC values. Part of the reason for the high AUC values is that there is some inherent circularity in our analyses, in the sense that the phenotype label used for evaluation is based on phenotype algorithms that also use various EHR features. Our initial idea was to show that we can obtain a “good” predicted phenotype based on selecting features (using ATSA or manual selection). Ideally, one would use independent gold-standard labels for these analyses, but unfortunately, such labels are not available for the traits we looked at, and in general, even when available, they are only available for a very restricted set of individuals.

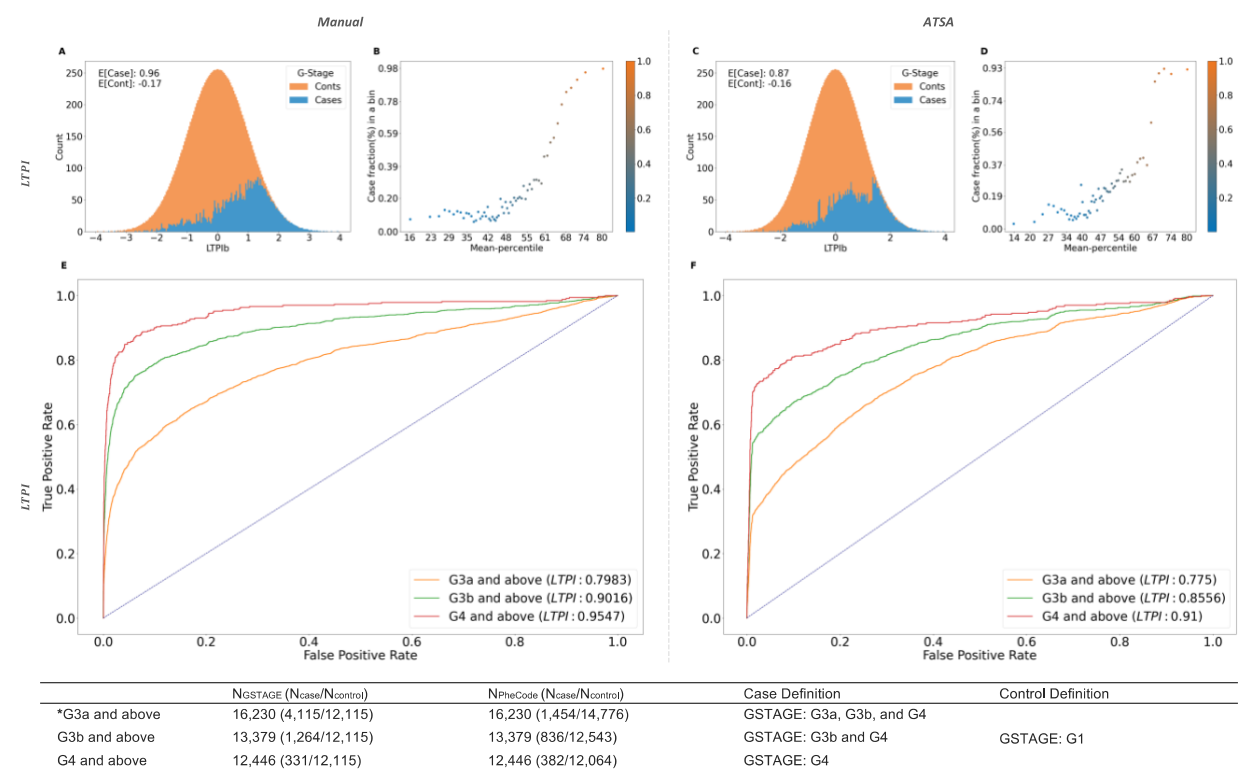


Fig. 9. Summary of *LTPI* analysis on eMERGE-CKD. Left. *LTPI* analysis results using manual selection of non-target traits (*Manual*). Right. *LTPI* analysis results using the Automatic Trait Selection Algorithm (*ATSA*). **A, C: Histograms of *LTPI* score distributions, segmented by PheCode classification. The y-axis shows the count of ranked *LTPI* scores. **B, D:** Scatter plots displaying case fractions (%) across mean-percentile bins. Each point's coordinates represent its mean percentile (x-axis) and case fraction (y-axis). **E, F:** ROC curves for *LTPI* scores in predicting CKD stages ('G3' to 'G4') in G-STAGE. Stages are color-coded. The diagonal navy line represents no-discrimination. AUC values in the legend indicate prediction performance. The accompanying table lists the sample sizes and definitions of case-control status used in the analysis. Each line in the AUROC corresponds to a row in this table. *LTPI* scores were standardized using rank-based inverse normal transformation (RINT).**

To mitigate this circularity, we have removed eGFR from *LTPI* construction in eMERGE-CKD. Below, we present results on eMERGE-CKD without using eGFR measurements when constructing *LTPI* for comparison. Overall, *LTPI* is quite robust, maintaining similar (although slightly worse) performance after removing eGFR.

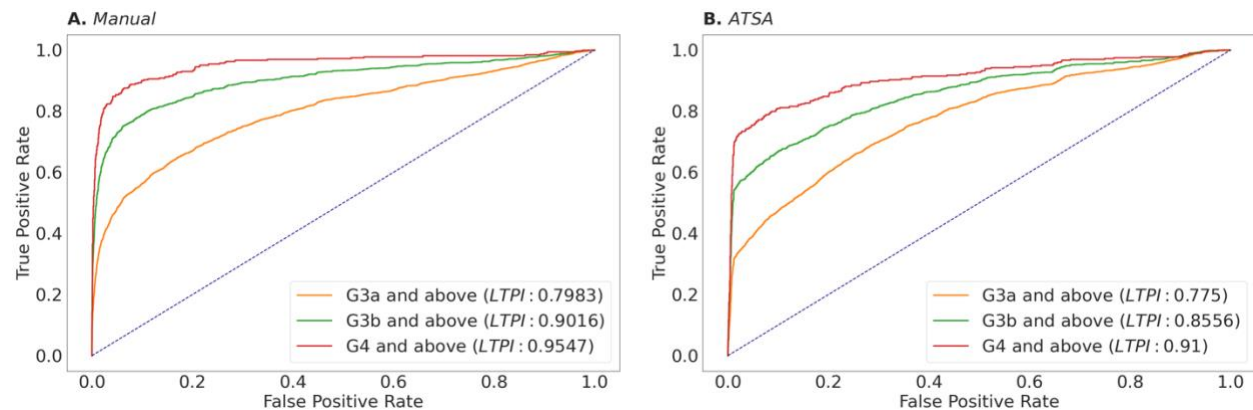


Figure. ROC curves for *LTPI* scores in predicting CKD stages ('G3' to 'G4') in G-STAGE. **A. *LTPI* analysis results using manual selection of non-target traits (*Manual*). **B.** *LTPI* analysis results using the Automatic Trait Selection Algorithm (*ATSA*). Stages are color-coded. The diagonal navy line represents no-discrimination. AUC values in the legend indicate prediction performance.**

We have also performed some additional analyses by excluding traits that were likely to be a result of the CKD diagnosis, as advised by our clinical collaborator, Dr. Kiryluk (see figure below). Even though AUROC values decrease, they are still high.

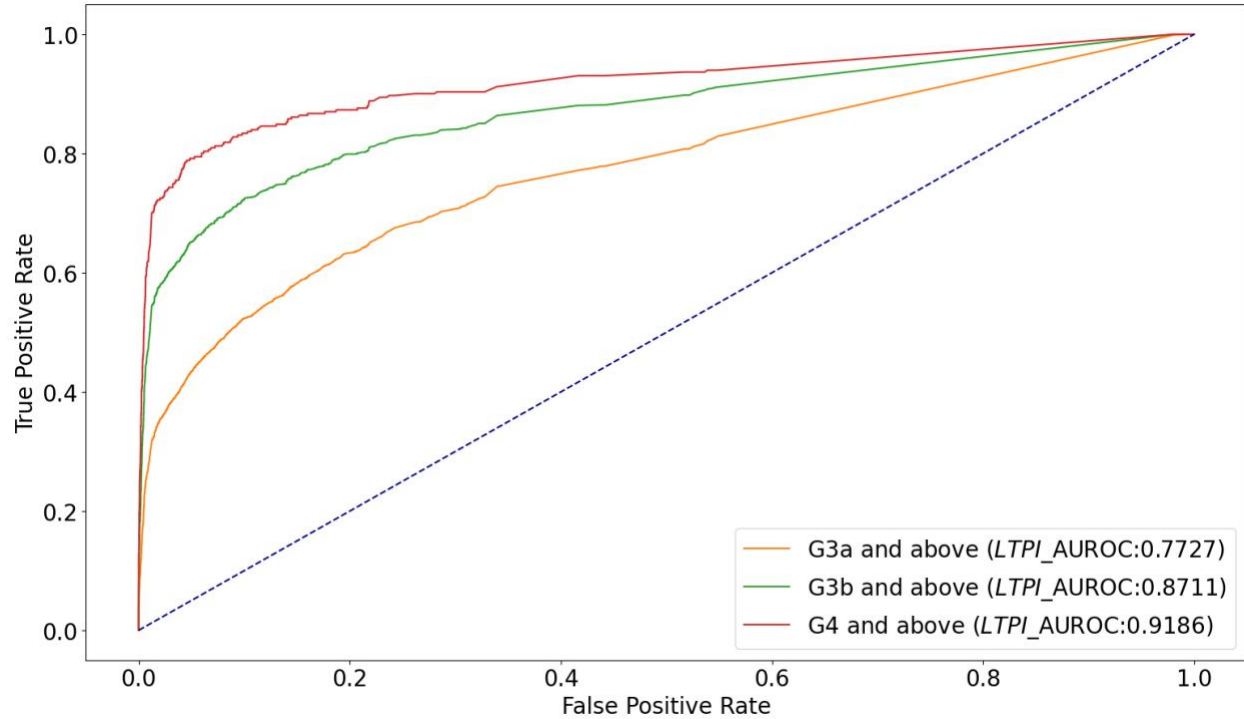


Fig. S6. Validation of risk prediction performance in eMERGE-CKD analysis using a stringent trait selection criterion. *LTPI* analysis results of eMERGE-CKD with the target phenotype and 32 non-target binary traits (*Manual*₃₂). The ROC curves show *LTPI* scores predicting CKD stages ('G3' to 'G4') in G-STAGE, with stages color-coded. The diagonal navy line represents no discrimination. AUC values in the legend indicate prediction performance.

R2m3. Concerns are noted regarding genomic inflation observed in several GWAS results within Table S19, prompting a call for discussion on its origins.

Response: The genomic inflation factors in Table S19 (Table S20 in the current manuscript) are largely similar between *LTPI* and *LTFH*. We have added several sentences in the Discussion on the need to validate the results from such analyses. The validations using existing large GWAS studies for the target trait, as performed by us, can further validate the results (see below).

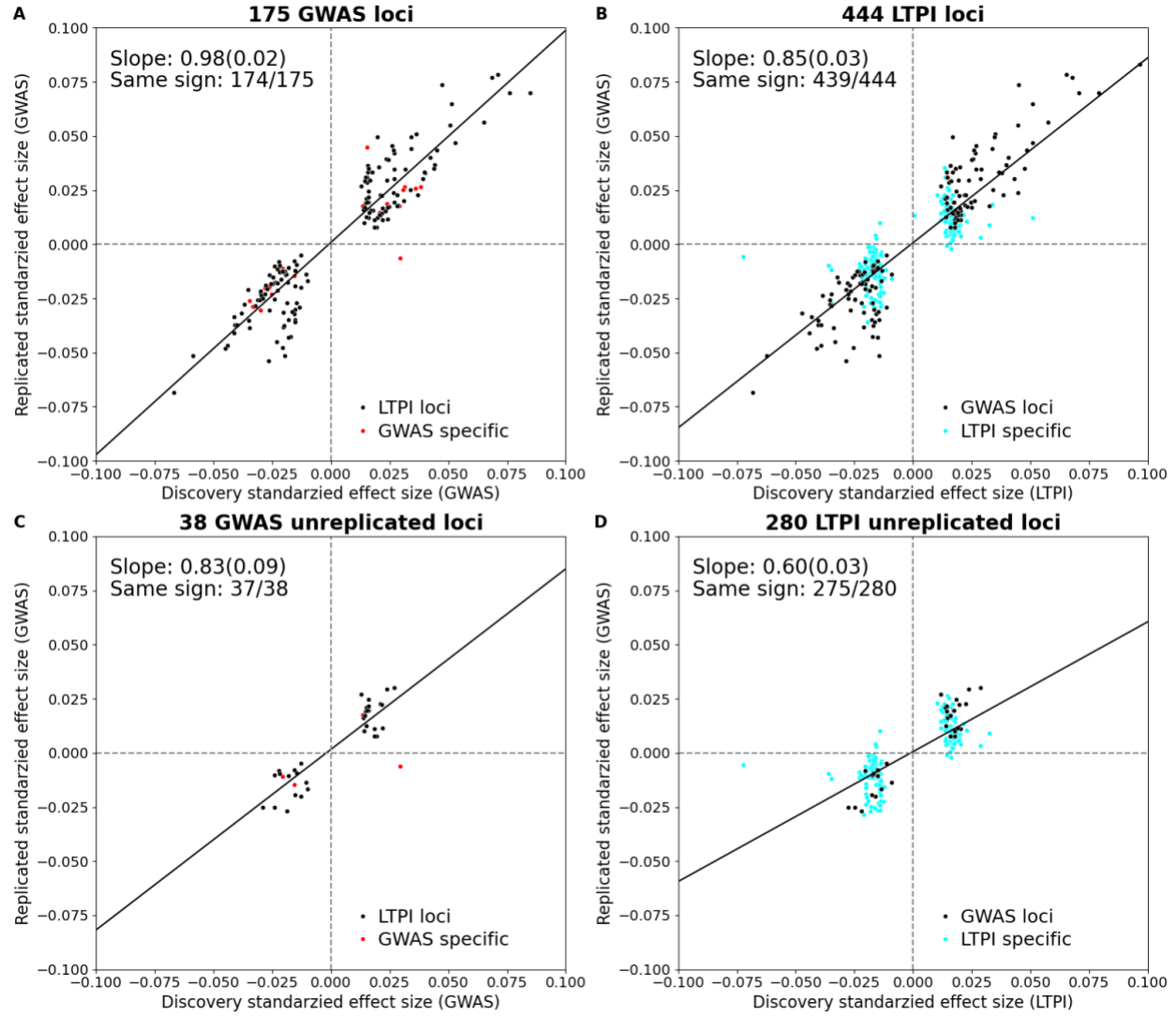


Fig. 10. Replication analysis using validation data sets. We generated scatter plots of standardized effect sizes ($Z/\sqrt{N_{\text{eff}}}$) in the validation GWAS data vs. the UK Biobank discovery data, aggregated across 4 diseases (CAD, T2D, breast cancer, and prostate cancer) for **A.** 175 loci identified by GWAS, **B.** 444 loci identified by LTPI, **C.** 38 loci identified by GWAS but not replicated by the validation GWAS data, and **D.** 280 loci identified by LTPI but not replicated by the validation GWAS data. Each point indicates the lead SNP at an independent locus, with the color indicating whether the locus contains genome-wide significant variants identified by both LTPI and GWAS.

Response to Reviewers

Reviewer #1:

I appreciate the comprehensive response from the authors to all major and minor points raised. I think the revised manuscript addressed all my concerns.

Response: Thank you very much for your thoughtful comments which have helped us improve the manuscript during the revision.

Reviewer #2:

Overall, the authors have addressed the major and minor concerns raised during the review process effectively. The distinctions between LTPI and existing multi-trait analysis approaches were clarified, and the explanations regarding how LTPI handles non-target traits, along with the integration of recent GWAS findings, were well-articulated. Additionally, the discussion on potential biases stemming from differences between datasets, as well as the concern about genomic inflation in GWAS results, were appropriately addressed.

Response: We are grateful for your detailed review and comments, which significantly enhanced the quality and clarity of our manuscript.

However, I noticed that Figure 1 rendered incorrectly in both documents, and I still find it confusing. Specifically, I'm unclear whether the y-axis of the boxes represents phenotypes. The term "dichotomous" is used, but is this simply the same as binary? I assume the dots on the density plots represent the final scores, but this isn't mentioned explicitly. Moreover, it's unclear why these dots are indicated in the Step 1 panel at all. Although I eventually understood the figure, I believe the authors should revisit and potentially revise it to improve clarity. Additionally, I feel that R2m1 has not been sufficiently addressed.

Response: Thank you for your constructive feedback regarding Figure 1. We apologize that in the previous submission there seemed to be an issue with Figure 1 after converting the files to pdf. We have carefully revised Figure 1 to enhance its clarity and address the concerns. The legend now explicitly describes both binary and continuous traits and the significance of the dots in the density plots. We have also expanded our description in the legend to ensure a clearer understanding of how the proposed methods leverage both binary and continuous phenotypes to compute the final LTPI scores.

We updated **Figure 1** in the revised manuscript as follows:

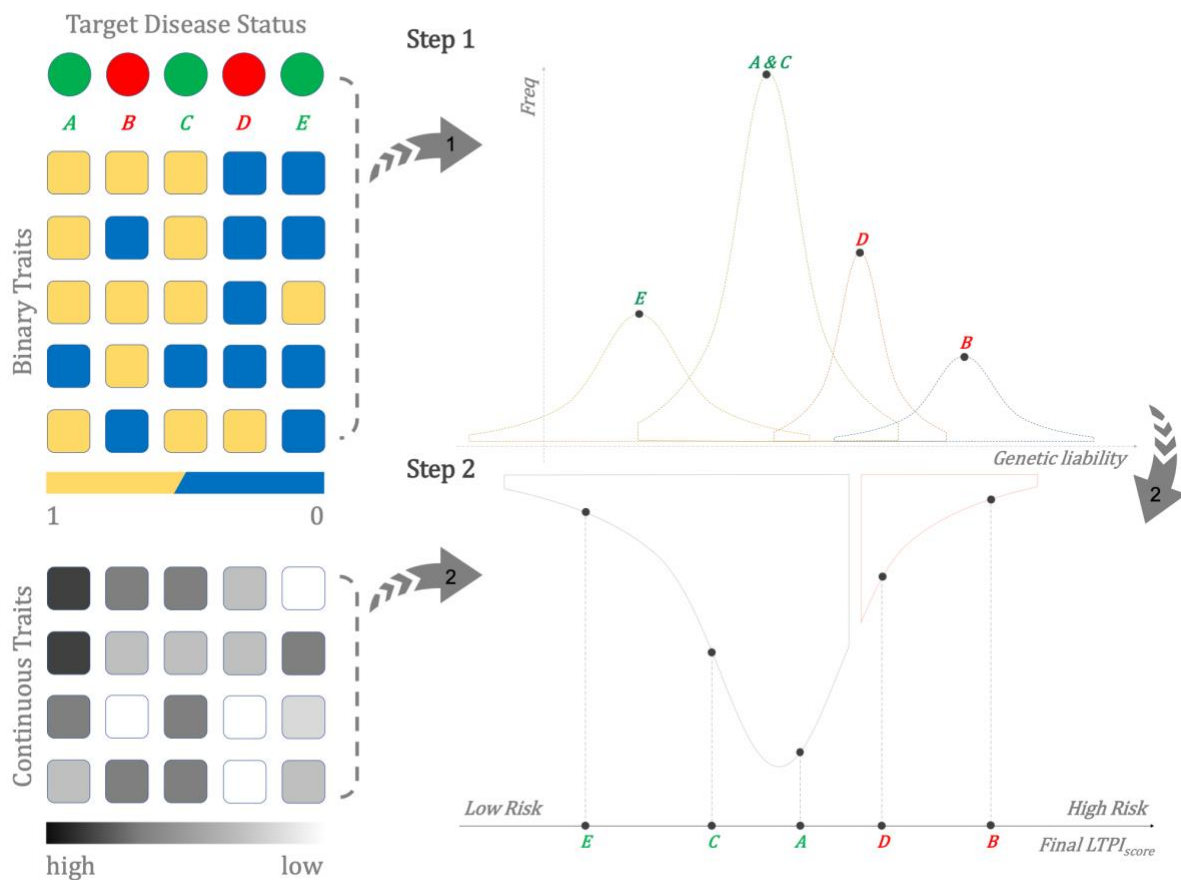


Fig. 1. Summary of the LTPI score construction. The LTPI method employs a two-step process to integrate non-target phenotype information and estimate latent disease risk for binary target phenotypes. The left panel displays a phenotype matrix with individuals (A to E) in columns and phenotypes in rows. The top row indicates target disease status (red for cases, green for controls), followed by non-target binary traits (yellow/blue) and continuous traits (shades of gray). **Step 1:** This step focuses on binary traits (1/0), estimating the posterior mean genetic liability and standard error for each individual. The top right panel visualizes these estimates on a plot where the x-axis represents genetic liability. Points indicate posterior mean genetic liabilities for individuals (A-E), noting that identical values can be assigned to multiple individuals sharing the same phenotypic features (as observed for A & C). **Step 2:** In this step continuous traits are integrated into the estimates from **Step 1** to produce the final LTPI score.