nature portfolio

Peer Review File

UKB-MDRMF: A Multi-Disease Risk and Multimorbidity Framework Based on UK Biobank Data

Corresponding Author: Dr Hongtu Zhu

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
In their manuscript, “UKBFound: A Foundation Model for Multi-Disease Prediction and Individual Risk Assessment Based on UK Biobank Data,” the authors present a novel modeling approach using large-scale multimodal biomedical data to facilitate disease prediction and patient risk assessment. The authors integrate demographic, lifestyle, lab, environmental, genetic, and imaging data through a robust processing pipeline that accounts for elements such as missingness and inputs the processed data into a variety of machine learning models to predict occurrence of diseases and long-term survival probability for individual patients. Lastly, the authors present a brief comorbidity analysis using output from their network-based machine learning models to identify potential interactions across phenotypes.

Overall, the authors have presented an important next step in the field of precision medicine, highlighting how multiomic data integration can benefit patients in the prediction of disease and survival. Here are some points of feedback that I think would serve to improve the quality of this paper

- The authors define their model as a “foundation model,” and while I appreciate that a foundation model technically refers to any AI model trained on vast amounts of data, I do not think that it necessarily applies in this case. Particularly given the popularity of biomedical papers involving generative AI, I think this terminology would be confusion for readers at a first glance. The authors should rename their method, or include an explanation of why their method counts as a foundation model. At the very least, the term ‘foundation model’ should be defined within the manuscript.

- The authors note that in their classification of data, type-II factors refer to instances “where data is missing because the variable is irrelevant or non-applicable to certain subgroups, such as sex-specific factors.” I would argue that this is a significant oversight and that sex-specific factors are highly relevant to both disease onset and progression (see the plethora of literature on genotype-by-sex effects, such as https://www.nature.com/articles/s41588-021-00912-0). If feasible, the authors should integrate sex as a variable under consideration, or a discussion of this significance of this factor should be included in the limitations of the paper.

- I would like to see more biological context discussed in the results of this manuscript. For instance, the authors note that bipolar disorder / depression status as well as age at recruitment were highly important across data categories. However, there is no discussion of a potential biomedical explanation for why these results come to light. Why are the disease categories that were identified more significant for disease prediction vs. risk assessment?

- I found the authors’ approach to defining multimorbidity networks based upon the final layer of the developed neural network models to be a significant and novel approach for identifying potential comorbidities between phenotypes, and I very much appreciate that this portion of the analysis was included. However, it is well known in the field of network medicine that different methods for network generation from biological data can yield vastly different networks. If multimorbidity analysis is to be included as a portion of this manuscript, then it is important that the authors include a robust comparison of the networks to one another in terms of elements such as overall network statistics or identified sub-clusters. If feasible, the authors can also compare their derived multimorbidities to comorbidities derived from the UKB EHR

- Lastly, I was excited to see the Shiny application that was developed for the manuscript - I think it is extremely important to have ways for users to directly play with the output of a paper even if they don’t have access to their own data to which they

can apply the methods. However, I think more context could be added to the application to improve its usability. Can a background section be included to explain the purpose of the app? Can tooltips be included to clarify the intention of the different dropdown options and the different tabs? With respect to the multimorbidity section of the application, I appreciate the ability to highlight different disease categories, but more detail is needed. Can network statistics for the data (i.e. attributes of individual disease nodes) be included (see https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giac002/6528770)? Or different options for data point clustering? Node sizing by a variable of choice? Incorporating such aspects will improve the usability and impact of your application.

(Remarks on code availability)
The code provided in the GitHub repository is is well-documented and readily accessible. The Shiny app developed by the authors is also very useful to facilitate exploration of the data.

Reviewer #2

(Remarks to the Author)
The paper proposes a "foundation model" for disease risk prediction and individual risk assessment from UK Biobank data.

There are a number of important aspects of the work that I found unclear which prevented my ability to provide a detailed review of the finer aspects of the work.

First, it is unclear to me why this work is described as a "foundation model". If we go by the original Stanford HAI definition of an AI model that is trained on broad data at scale and is adaptable to a wide range of downstream tasks then it is unclear quite how this work fits into this definition. The downstream task seems quite well defined and this model appears to me like a prediction model -- albeit a complicated one. There is no use of self-supervised or transfer learning and no sense of what the re-usable components (i.e. representations) are might underpin what is typically understood as a "foundation model".

Going further, it is unclear how the foundation model is generalisable in application outside of UK Biobank. In the majority of possible prediction applications, it is unlikely that the user will have access to the extensive wealth of information present in UK Biobank. The authors themselves suggest this because of the lack of validation data. Can this foundation model support applications where there is more limited access to information?

I was also unsure how individual-level predictions could be derived for any particular individual - both from the textual description and the Shiny app.

It was also unclear in the methods description (e.g. Figure 11) how the sampling times of different modalities of data was handled and how this entered into the learning strategy for the model. The work should define the prediction setup more thoroughly as it is currently unclear what the relevant input-outcome specification is for the prediction tasks. There was no mention of TRIPOD guidelines and the relevance (or not -- with justification) to this work.

In general, significantly more detail is required for the model construction and design as the schematic diagrams alone are not adequate.

(Remarks on code availability)
The code has been made available and there is information about the organisation of the code on the repository. However, the wider documentation is sorely lacking. There is insufficient information to understand how to use the code to replicate the experiments and to use it for alternative tasks. Python scripts and notebooks are not throughly commented. Scripts and some instructions are provided for the input data preparation but I would not be confided that there is sufficient detail to reproduce the work.

Reviewer #3

(Remarks to the Author)
Thank you for giving me the opportunity to review this manuscript. 'UKBFOUND' is a cohort study which uses data from UK Biobank data to build a prediction model using demographic, lifestyle factors, biomarkers and GWAS. The model was build to predict the risk of 1560 diseases and several machine learning techniques were used.

I will highlight some strengths.
1) Comprehensive set of predictors are considered. Use of UK Biobank allows this kind of analysis wide range of biomarkers, environmental factors and genetic factors are available.
2) Several methodological approaches were evaluated: logistic regression, Random Forest, XGBoost, Light GBM, FCNN, POPDx and CATI. I am not an expert in ML so critique of choosing these methods is out of my scope.
3) The presentation of technical details of methods and results from complicated analyses has been done very well with comprehensive coverage of relevant details.
4) Figures are good quality, missing data analysis is impressive.
5) For the sake of transparency and replicability, all codes have been provided on github. Appropriate comparison has been made with existing literature in this field.

I do have some significant concerns and reservations:

1) While technically sound, I am not sure what is the practice or policy related utility of this analysis.
2) Most important variables identified had hardly any surprises with majority related to demographic, socio-economic status and metabolic health/mental health. When a relatively less commonly used variable was included, it was hard to ascertain, how this predictor can change/influence practice (e.g. carotid ultrasound).
3) Baseline health status (i.e. conditions participants had the start of the follow-up) is likely to be one of the strongest predictor of future health trajectory and multimorbidity trajectory. The analysis should have considered this crucial factor.
4) The list of outcomes considered (>1000 phenotypes) were too broad. It would have been better to select a small list of outcomes which are most important from public health perspective.
5) I could not seen any adjustment for multiple testing, very important considering the number of outcomes.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have done an excellent job at incorporating my and the other reviewers' feedback into their manuscript. I particularly appreciate the incorporation of further biological interpretation in the Discussion, extended network analysis in the Supplement, and the revisions to the Shiny application.

(Remarks on code availability)
The authors have provided a comprehensive, well-documented overview of the code used in this study. I do note that the current repository is currently the only repository saved under the account of "anonymous-researcher01," which at first glance seems like a temporary account to me - I hope that this resource will continue to remain online upon publication of the manuscript.

Reviewer #2

(Remarks to the Author)
The authors have provided an updated version of their work, now entitled “UKB-MDRMF: A Multi-Disease Risk & Multimorbidity Framework Based on UK Biobank Data”.

Foundation model

Unfortunately, with clarification that this is *not* a foundation model, the level of novelty of the work is substantially diminished as the results are now, as previously identified, more of a large scale series of prediction models that are coupled in their training. Methodologically the authors have used fully-connected neural networks and DeepSurv which are standard, pre-existing approaches. The specification of the loss functions also suggest that the loss sub-component for each Phecode is independent. It is not clear that there is any coupling between models for different precedes?

Validation

The authors used All of Use for validation purposes. All of Use is an emerging population research cohort but I should clarify that my original query was more related to routinely collected point-of-care clinical data. There will clearly be similarities in the data collected between population research cohorts.

Nonetheless, the use of All of Us is a promising addition by the authors. However, I was confused by the statement “we retrained the FCNN and DeepSurv models from the UKB-MDRMF using the All of Us data”. If the intention is to validate the models then shouldn’t the trained model from UKB be transported and used as-is on All of Us without retraining? If the model is retrained, this would be a validation of the process rather than the models.

Shiny app

The authors have provided a limited demonstration of the individual risk prediction models. While this addition is welcomed, I did not feel that it was done particularly well. The survival plots are not described, what does the “survival probability” on the vertical axis mean? Is this the risk of the disease? If so, why do the survival curves seem to asymptote after 80 years of age at a non-trivial probability?

Methodological details

The “enhanced” Figure 11 is still a high-level schematic. Unless I have missed it, the authors have not really provided an item-by-item report against the TRIPOD checklist which could be incorporated into Supplementary Information. For example, Item 4b iii) “The length of follow-up and prediction horizon/time frame are reported”, is a checklist item where the authors would benefit from giving a more explicit response. It is unclear what the prediction horizons are for the framework. It is unclear how deaths post-enrolment are handled. The removal of disease occurrences prior to the covariate collection leads

to significant biases and limitations in the prediction model particularly if pre-existing diseases highly correlate with others. The authors have not considered competing risk implications? The methodological description could be improved significantly further.

Software repository

The files on the Github have been updated and documentation improved.

Overall comments

While there are some improvements to the presentations. I found some substantial details missing, in particular, those related to methods. This continues to prevent me from a more comprehensive understanding of the paper, the appropriateness of the analyses and the results. While the change from a foundation model to predictive framework results in a less novel concept, the multi-disease approach would be potentially interesting. However, it is not clear as stated above, that there is any dependence introduced between diseases.

(Remarks on code availability)
The code repository has been substantially revised and is significantly more usable.

Reviewer #3

(Remarks to the Author)
Thank you for your response and revised manuscript. The revised submission has improved significantly and a greater deal of clarity has been added to the analysis and subsequent presentation.
However, the two crucial points I raised in the previous review have not been addressed.
Firstly, the baseline health status has not been considered in the analysis which is very likely to have a hugely significant impact on the health outcomes studied.
Secondly, a very wide variety of clinical outcomes are considered. Authors argue why this approach is relevant for their analysis. However, a sensitivity analysis with much smaller number but more clinically relevant outcomes will be more useful.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
The authors have addressed my concerns comprehensively. The revisions include significantly more detail to improve the explanation and reproducibility of large parts of the work (especially methods and data curation) and now addresses key limitations of the work. The use of TRIPOD has been applied and more text has been provided for the Shiny app. This is now a substantially improved piece of work compared to the original.

(Remarks on code availability)
The code is well documented and well structured for reproduction.

Reviewer #3

(Remarks to the Author)
The authors have improved their manuscript significantly and addressed all of my concerns. I do not have any further comments and would like to congratulate the authors.

(Remarks on code availability)

Response to Reviewer 1

Thank you very much for your constructive comments and your encouragement! We have revised our paper carefully to address the concerns. The responses to your comments are given in the following.

1. *In their manuscript, "UKBFound: A Foundation Model for Multi-Disease Prediction and Individual Risk Assessment Based on UK Biobank Data," the authors present a novel modeling approach using large-scale multimodal biomedical data to facilitate disease prediction and patient risk assessment. The authors integrate demographic, lifestyle, lab, environmental, genetic, and imaging data through a robust processing pipeline that accounts for elements such as missingness and inputs the processed data into a variety of machine learning models to predict occurrence of diseases and long-term survival probability for individual patients. Lastly, the authors present a brief comorbidity analysis using output from their network-based machine learning models to identify potential interactions across phenotypes.*

   Overall, the authors have presented an important next step in the field of precision medicine, highlighting how multiomic data integration can benefit patients in the prediction of disease and survival. Here are some points of feedback that I think would serve to improve the quality of this paper

   - The authors define their model as a "foundation model," and while I appreciate that a foundation model technically refers to any AI model trained on vast amounts of data, I do not think that it necessarily applies in this case. Particularly given the popularity of biomedical papers involving generative AI, I think this terminology would be confusion for readers at a first glance. The authors should rename their method, or include an explanation of why their method counts as a foundation model. At the very least, the term 'foundation model' should be defined within the manuscript.

   Response: Thank you for your insightful comment. We agree that using the term "foundation model" might cause confusion, as it is commonly associated with large language models and generative AI. After careful consideration, we have renamed our method to eliminate ambiguity and more accurately reflect its core functionalities. Our model is now termed the **UKB** **M**ulti-**D**isease **R**isk & **M**ultimorbidity **F**ramework (UKB-MDRMF). This name more precisely encapsulates our approach, focusing on multi-disease risk assessment and exploring multimorbidity

patterns within the UK Biobank dataset. The framework offers a comprehensive pipeline for processing diverse data types for prediction and risk assessment, distinguishing it from the foundation models typically associated with broader, self-supervised, or generative tasks. The manuscript has been revised to include these changes and additional clarifications to better define the model's scope and objectives

2. - *The authors note that in their classification of data, type-II factors refer to instances "where data is missing because the variable is irrelevant or non-applicable to certain subgroups, such as sex-specific factors." I would argue that this is a significant oversight and that sex-specific factors are highly relevant to both disease onset and progression (see the plethora of literature on genotype-by-sex effects, such as https://www.nature.com/articles/s41588-021-00912-0). If feasible, the authors should integrate sex as a variable under consideration, or a discussion of this significance of this factor should be included in the limitations of the paper.*

   Response: We appreciate this important point and apologize for any ambiguity in the manuscript. To clarify, we have not omitted crucial demographic variables such as sex or age; indeed, these are among the most significant predictors in our model. Sex-specific factors, despite their high rates of missing data, are also integrated.

   To prevent confusion with existing terminology, we have renamed Type-II missingness to 'Pattern-II missingness'. Please see Section 4.4, "Missing data processing", on Fig. 10. This refers to missing values that occur when a variable is not applicable to a certain subgroup. For example, all male participants have missing data for female-specific variables like 'number of live births' and 'age at first live birth'. We address these missing values in a manner similar to the well-established missing indicator method [1]. However, we diverge from the typical approach by not adding separate missing indicators to our prediction model, since these indicators (e.g., the sex variable, which indicates the absence of sex-specific factors) are already included in our set of variables. This approach avoids redundancy and ensures that comprehensive information, including sex-specific factors, is preserved.

   In the revised manuscript, we have provided a clearer explanation of such missing data scenarios and enhanced the clarity of our presentation. We have also included Supplementary Table D5 and the file "S2_missing_pattern.csv" in the Supplementary Materials detailing our classification of missing patterns.

3. - *I would like to see more biological context discussed in the results of this manuscript. For instance, the authors note that bipolar disorder / depression status as well as age at recruitment were highly important across data categories. However, there is no discussion of a potential biomedical explanation for why these results come to light. Why are the disease cate-*

gories that were identified more significant for disease prediction vs. risk assessment?

Response: Thank you! We have now included a more comprehensive biological explanation for the significance of key variables, including depression-related factors (e.g., bipolar disorder and depression), age, and other important features identified in our model. We have also elaborated on the observed differences between disease prediction and risk assessment, providing a clearer rationale for the results.

Our findings highlight the role of mental health-related variables, particularly depression and bipolar disorder, in multi-disease prediction and risk assessment. These conditions are associated with alterations in the neuroendocrine system, which affect neurotransmitter levels such as serotonin, dopamine, and norepinephrine [13, 9]. These neurotransmitter imbalances have been linked to a higher risk for multiple diseases, including cardiovascular disease, diabetes, and chronic inflammatory conditions [14, 5, 3]. Additionally, depression has been shown to impair immune system function, leading to increased susceptibility to infections, autoimmune diseases, and chronic inflammatory conditions [4]. These mechanisms provide a plausible explanation for why depression-related features were consistently identified as key variables in both prediction and risk assessment models.

Age was another prominent variable influencing multi-disease outcomes. The biological impact of aging is well-documented, as metabolic, immune, and cellular repair processes gradually decline with age, leading to an increased risk of multiple chronic diseases [12]. Aging is also associated with changes in epigenomic and transcriptomic profiles, resulting in non-linear shifts in multi-omics data [16]. Our framework captures the importance of age across disease categories, as reflected in both the disease prediction and risk assessment models. These age-related biological changes contribute to disease heterogeneity, making age a crucial factor in understanding multimorbidity.

Furthermore, our findings highlight the significance of body fat percentage, body mass index (BMI), waist circumference, lymphocyte percentage, and microalbumin in urine as key predictors in multi-disease outcomes. Excess body fat, particularly abdominal obesity reflected by waist circumference, contributes to insulin resistance, systemic inflammation, and dyslipidemia, increasing the risk of metabolic diseases such as type 2 diabetes, cardiovascular disease, and non-alcoholic fatty liver disease [10, 6]. Similarly, lymphocyte percentage reflects immune system function, where alterations often indicate immune dysregulation or chronic inflammation, key contributors to autoimmune diseases, infections, and cancer [8]. In addition, microalbumin in urine serves as an early marker of kidney damage and systemic vascular injury, commonly observed in individuals with diabetes and hypertension, reflecting glomerular dysfunction and heightened cardiovascular risk [15]. By integrating these factors, our framework offers a more comprehensive view of the biological mechanisms underlying

multimorbidity, enhancing its ability to identify key pathways for disease prediction and risk assessment.

Besides, we have further clarified the observed differences between disease prediction and risk assessment. Although both tasks aim to model disease occurrence, they have distinct objectives and methodologies:

- Disease Prediction: This task focuses on predicting the likelihood of disease onset as a classification problem. It utilizes multimodal data to forecast disease occurrence in the future.
- Risk Assessment: This task models the time-to-event for disease onset, taking into account censored data and longitudinal disease trajectories. The survival-based approach provides a richer assessment of long-term risk and disease progression over time.

Our analysis revealed that certain disease categories were more prominent in disease prediction compared to risk assessment. For instance, reproductive diseases showed high prediction accuracy in classification-based disease prediction models, while sensory diseases had lower predictive performance for both disease prediction and risk assessment. The differences may be attributed to the nature of the data and the availability of early predictive signals for certain diseases. For example, reproductive diseases may have clear early indicators (e.g., hormone-related biomarkers) that enhance predictive accuracy, while sensory diseases (e.g., vision or hearing loss) may not present obvious early warning signs.

We have expanded the Discussion section on page 9 to provide a more in-depth analysis of the biological context. We now discuss the roles of key variables, including age, depression-related factors, and metabolic health, and their connections to the observed multimorbidity patterns. Additionally, we have clarified the conceptual differences between disease prediction and risk assessment, highlighting how our framework offers a comprehensive approach to modeling multimorbidity. This revision provides a stronger biological basis for our findings and a more complete discussion of the results.

4. *- I found the authors' approach to defining multimorbidity networks based upon the final layer of the developed neural network models to be a significant and novel approach for identifying potential comorbidities between phenotypes, and I very much appreciate that this portion of the analysis was included. However, it is well known in the field of network medicine that different methods for network generation from biological data can yield vastly different networks. If multimorbidity analysis is to be included as a portion of this manuscript, then it is important that the authors include a robust comparison of the networks to one another in terms of elements such as overall network statistics or identified sub-clusters. If feasible, the authors can also compare their derived multimorbidities to comorbidities derived from the UKB EHR.*

Response: Thank you for highlighting the importance of comparing networks derived from different methods and validating their clinical relevance. In the revised manuscript, we have expanded the multimorbidity network analysis, which is now detailed in both the main text and the Supplementary Information. Please see Section 2.6, "Multimorbidity and trend analysis of disease risks", on pages 6-7 and Section D.3.1, "Multimorbidity network", in the Supplementary Information.

Our multimorbidity network was constructed using embeddings derived from the penultimate layer of neural network models, specifically FCNN for disease prediction and DeepSurv for risk assessment. Each Phecode was represented as a 100-dimensional vector, and we calculated pairwise correlations between these embeddings to form a correlation matrix. Only significant correlations were retained, producing undirected graphs where nodes represent Phecodes and edges signify strong correlations. The Louvain method was applied to detect communities within the graph [7], revealing distinct clusters of diseases with significant comorbid relationships.

Figures 1 and 2 (Fig. D8 and D9 in the Supplementary Information) illustrate the six major disease communities derived from the FCNN and DeepSurv models, respectively. These clusters encompass various disease categories, including reproductive health, mental health, metabolic disorders, and cardiovascular diseases. The identification of these primary clusters is consistent with findings from other comorbidity studies using large-scale biomedical data [2, 7], affirming the robustness of our methodology.

Moreover, we validated the multimorbidity clusters identified in our networks against clinically observed comorbidities, enhancing the clinical relevance of our findings. For example, we identified clusters like Hypertensive chronic kidney disease (Phecode: 401.22) within urinary-related diseases, often associated with cardiovascular conditions, and Anemia of chronic disease (Phecode: 285.2), traditionally classified under blood disorders. These observations align with other research [11] and demonstrate our approach's capability to uncover subtle disease relationships that may be obscured in conventional classification systems.

The revised manuscript includes these updates in the Results section and Supplementary Information, supplemented with new figures and analyses for a comprehensive comparison and validation of the multimorbidity networks.

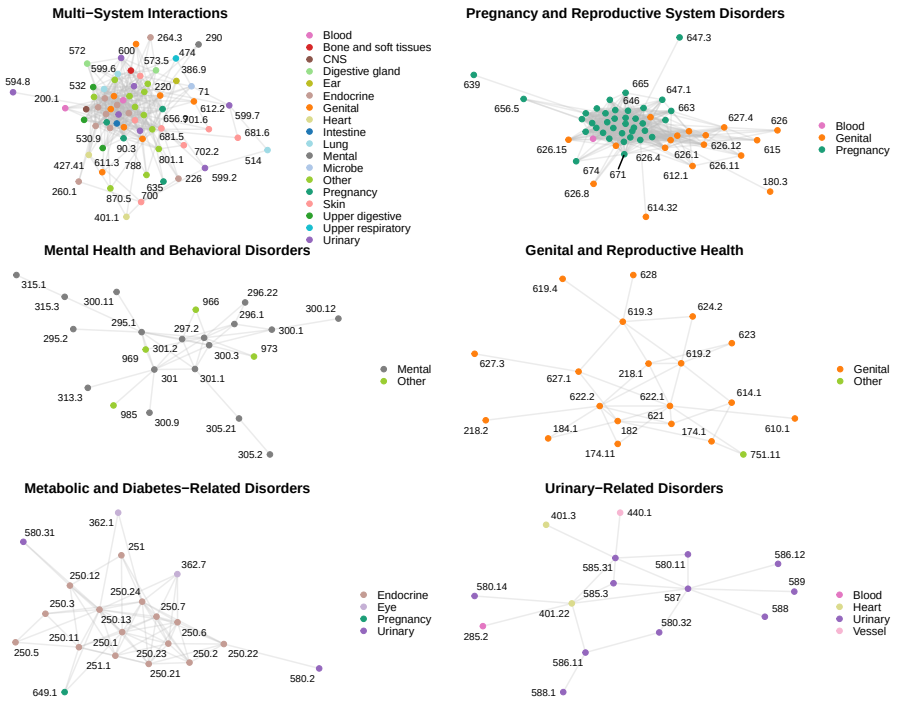

Figure 1: **Network visualization of six disease communities identified by Phecode clusters based on FCNN.** Each node represents a specific Phecode, with edges indicating potential comorbid relationships between diseases. The node colors correspond to different disease categories, such as blood, urinary, reproductive health, mental health, and metabolic disorders, derived from their Phecode classification. This visualization highlights the comorbidity patterns captured through the FCNN model's penultimate layer embeddings.

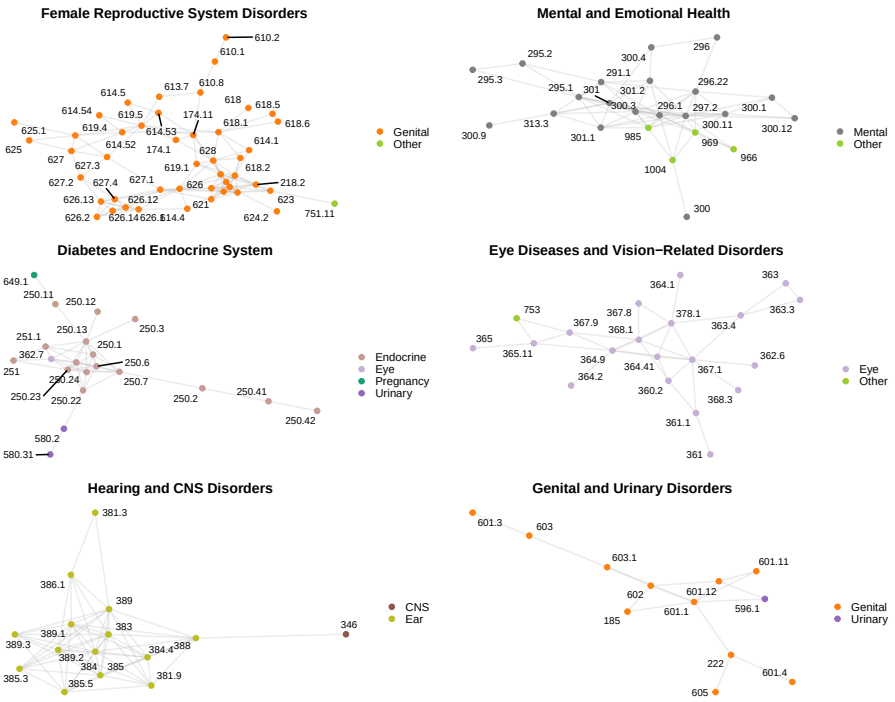

Figure 2: **Network visualization of six disease communities identified by Phecode clusters based on DeepSurv.** Each node represents a specific Phecode, with edges indicating potential comorbid relationships between diseases. The node colors correspond to distinct disease categories such as reproductive health, mental health, metabolic disorders, vision-related disorders, hearing, and central nervous system (CNS) disorders, and genital and urinary disorders, derived from their Phecod.

5. - *Lastly, I was excited to see the Shiny application that was developed for the manuscript - I think it is extremely important to have ways for users to directly play with the output of a paper even if they don't have access to their own data to which they can apply the methods. However, I think more context could be added to the application to improve its usability. Can a background section be included to explain the purpose of the app? Can tooltips be included to clarify the intention of the different dropdown options and the different tabs? With respect to the multimorbidity section of the application, I appreciate the ability to highlight different disease categories, but more detail is needed. Can network statistics for the data (i.e. attributes of individual disease nodes) be included (see https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giac002/6528770)? Or different options for data point clustering? Node sizing by a variable of choice? Incorporating such aspects*

will improve the usability and impact of your application.

Response: Thank you for the positive feedback on the Shiny application and your thoughtful suggestions for enhancing its usability. In response, we have added a comprehensive background section to the application, providing users with context on the data, methods, and overall purpose of the tool (`https://luminite.shinyapps.io/ukb-mdrmf/`). Additionally, we have included detailed explanations for the functionality of interactive buttons and dropdown menus to create a more user-friendly experience.

To further improve the multimorbidity section, we have integrated an interactive comorbidity network based on Louvain clustering. Users can now adjust model types, set correlation thresholds for edge connections, and explore specific diseases along with their corresponding Phecodes interactively. These updates significantly enrich the representation and exploration of multimorbidity within the application. Additionally, we have introduced individual-level baseline predictions to enhance the tool's overall functionality, further improving its usability and impact.

6. *The code provided in the GitHub repository is well-documented and readily accessible. The Shiny app developed by the authors is also very useful to facilitate exploration of the data.*

   Response: We appreciate the positive remarks regarding the code availability. We will continue to ensure that the code is well-maintained and documented, and we are open to any further suggestions for improvement.

References

[1] Paul D Allison. Missing data. *The SAGE handbook of quantitative methods in psychology*, pages 72–89, 2009.

[2] Thomas Beaney, Jonathan Clarke, David Salman, Thomas Woodcock, Azeem Majeed, Paul Aylin, and Mauricio Barahona. Identifying multiresolution clusters of diseases in ten million patients with multimorbidity in primary care in england. *Communications Medicine*, 4(1):102, 2024.

[3] Eléonore Beurel, Marisa Toups, and Charles B Nemeroff. The bidirectional relationship of depression and inflammation: double trouble. *Neuron*, 107(2):234–256, 2020.

[4] Beatriz Cañas-González, Alonso Fernández-Nistal, Juan M Ramírez, and Vicente Martínez-Fernández. Influence of stress and depression on the immune system in patients evaluated in an anti-aging unit. *Frontiers in psychology*, 11:1844, 2020.

[5] Silvia Capellino, Maren Claus, and Carsten Watzl. Regulation of natural killer cell activity by glucocorticoids, serotonin, dopamine, and epinephrine. *Cellular & molecular immunology*, 17(7):705–711, 2020.

[6] Jean-Pierre Després and Isabelle Lemieux. Abdominal obesity and metabolic syndrome. *Nature*, 444(7121):881–887, 2006.

[7] Guiying Dong, Jianfeng Feng, Fengzhu Sun, Jingqi Chen, and Xing-Ming Zhao. A global overview of genetically interpretable multimorbidities among common diseases in the uk biobank. *Genome medicine*, 13:1–20, 2021.

[8] David Furman, Judith Campisi, Eric Verdin, Pedro Carrera-Bastos, Sasha Targ, Claudio Franceschi, Luigi Ferrucci, Derek W Gilroy, Alessio Fasano, Gary W Miller, et al. Chronic inflammation in the etiology of disease across the life span. *Nature medicine*, 25(12):1822–1832, 2019.

[9] Gregor Hasler. Pathophysiology of depression: do we have any solid evidence of interest to clinicians? *World psychiatry*, 9(3):155, 2010.

[10] Gökhan S Hotamisligil. Inflammation and metabolic disorders. *Nature*, 444(7121):860–867, 2006.

[11] Ushani Jayamanna and JAA Sampath Jayaweera. Childhood anemia and risk for acute respiratory infection, gastroenteritis, and urinary tract infection: a systematic review. *Journal of Pediatric Infectious Diseases*, 18(02):061–070, 2023.

[12] Matt Kaeberlein, Peter S Rabinovitch, and George M Martin. Healthy aging: the ultimate preventative medicine. *Science*, 350(6265):1191–1193, 2015.

[13] Jung Goo Lee, Young Sup Woo, Sung Woo Park, Dae-Hyun Seog, Mi Kyoung Seo, and Won-Myong Bahk. Neuromolecular etiology of bipolar disorder: possible therapeutic targets of mood stabilizers. *Clinical Psychopharmacology and Neuroscience*, 20(2):228, 2022.

[14] SM Matt and PJ Gaskill. Where is dopamine and how do immune cells see it?: dopamine-mediated immune cell function in health and disease. *Journal of Neuroimmune Pharmacology*, 15:114–164, 2020.

[15] Bruce A Perkins, Linda H Ficociello, Bijan Roshan, James H Warram, and Andrzej S Krolewski. In patients with type 1 diabetes and new-onset microalbuminuria the development of advanced chronic kidney disease may not require progression to proteinuria. *Kidney international*, 77(1):57–64, 2010.

[16] Xiaotao Shen, Chuchu Wang, Xin Zhou, Wenyu Zhou, Daniel Hornburg, Si Wu, and Michael P Snyder. Nonlinear dynamics of multi-omics profiles during human aging. *Nature aging*, pages 1–16, 2024.

Response to Reviewer 2

Thank you very much for your constructive comments and your encouragement! We have revised our paper carefully to address the concerns. The responses to your comments are given in the following.

1. *The paper proposes a "foundation model" for disease risk prediction and individual risk assessment from UK Biobank data.*

   There are a number of important aspects of the work that I found unclear which prevented my ability to provide a detailed review of the finer aspects of the work.

   First, it is unclear to me why this work is described as a "foundation model". If we go by the original Stanford HAI definition of an AI model that is trained on broad data at scale and is adaptable to a wide range of downstream tasks then it is unclear quite how this work fits into this definition. The downstream task seems quite well defined and this model appears to me like a prediction model – albeit a complicated one. There is no use of self-supervised or transfer learning and no sense of what the re-usable components (i.e. representations) are might underpin what is typically understood as a "foundation model".

   Response: Thank you for your observation. We fully recognize the potential confusion surrounding the term "foundation model" and agree that it might not accurately reflect the scope of our work, especially given the lack of self-supervised or transfer learning components. In response to your feedback, we have removed the term "foundation model" and now describe our approach as the UKB Multi-Disease Risk & Multimorbidity Framework (UKB-MDRMF). This name more accurately represents its focus as a specialized tool for multi-disease prediction and risk assessment. The UKB-MDRMF leverages UK Biobank data specifically for disease risk prediction and multimorbidity analysis. Its structure is tailored to explore complex relationships between diseases and risk factors, rather than providing broad adaptability across unrelated tasks. We believe this revision more clearly defines the specialized nature of our work.

2. *Going further, it is unclear how the foundation model is generalisable in application outside of UK Biobank. In the majority of possible prediction applications, it is unlikely that the user will have access to the extensive wealth of information present in UK Biobank. The authors themselves sug-*

gest this because of the lack of validation data. Can this foundation model support applications where there is more limited access to information?

Response: Thank you for your insightful comment. We recognize your concerns regarding the generalizability of our model beyond the UK Biobank, particularly in contexts where comprehensive data may not be available. In response, we have implemented two strategies to address these issues and enhance the applicability of our framework:

- **Synthetic Data Generation for Broader Application**: To aid in the understanding and usage of our model, we have created a synthetic dataset, detailed in our GitHub repository (`https://github.com/yourrepository/scripts/randomdatademo.ipynb`). This dataset simulates more typical real-world scenarios, allowing users to run our code and validate the approach even when extensive datasets like the UK Biobank are not accessible. This adaptation ensures that our model remains practical and versatile for various predictive applications.
- **Validation with the All of Us Dataset**: We have also validated our model using the All of Us dataset, a comprehensive and diverse biomedical resource from the United States, similar to the UK Biobank. This dataset includes a wide array of participants, representing a broad demographic spectrum. The results, which we presented on page 7, Section 2.7, "Multi-center Validation Using All of Us Data," with further details in Appendix D.3.2, "Multi-center Validation Using All of Us Data," show that our model performs comparably well, affirming its robust predictive capabilities across different datasets.

These enhancements ensure that our framework is not only flexible but also broadly applicable beyond the UK Biobank, accommodating use cases with limited data availability. We have detailed these improvements in the revised manuscript to highlight our framework's adaptability and practicality.

3. *I was also unsure how individual-level predictions could be derived for any particular individual - both from the textual description and the Shiny app.*

Response: We apologize for any confusion regarding individual-level predictions. Our framework generates these predictions by processing an individual's complete feature set through the model to estimate disease risk or survival probabilities, which are then presented as probabilities for various diseases or outcomes. Importantly, our population-level results are actually aggregated from these individual predictions. To clarify this process, we have enhanced the content in our Shiny app to better reflect the results at the individual level.

In the updated Shiny app (`https://luminite.shinyapps.io/ukb-mdrmf/`), users can now input specific features such as age, sex, and lifestyle factors. The app utilizes the trained model to generate precise predictions for the entered individual. Recognizing the need for clearer guidance, we have also improved the app interface and textual descriptions. Detailed instructions and tooltips have been added to assist users in how to accurately enter data, interpret the predictions made, and explore different health outcomes.

4. *It was also unclear in the methods description (e.g. Figure 11) how the sampling times of different modalities of data was handled and how this entered into the learning strategy for the model. The work should define the prediction setup more thoroughly as it is currently unclear what the relevant input-outcome specification is for the prediction tasks. There was no mention of TRIPOD guidelines and the relevance (or not – with justification) to this work.*

   Response: Thank you for your feedback. We apologize that our previous version did not clearly explain the handling of sampling times across modalities. In the revised manuscript, we have elaborated on how the temporal information of each modality was integrated into the model. For different types of predictor data (X), we used the baseline time for most variables of each sample (Field 53). For the timing of responses from different sources (Y), each disease record has a specific corresponding time. When data are merged, we retain the earliest occurrence time for individuals with multiple occurrences of a disease. Moreover, during the alignment of predictor and response times, we only consider records where responses occur after the predictors; records where responses precede X are marked as "NA," ensuring that the model construction is logically sound by not using cases where the response precedes the predictor data. Additionally, we have enhanced Figure 11 to clarify the methodological approach, and a detailed explanation has been included in the methods section of the manuscript.

   We also acknowledge the omission of the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) guidelines in the initial manuscript. We have now incorporated the TRIPOD guidelines, enhancing our approach to model development, validation, and reporting to meet standards of transparency and reproducibility. Specific enhancements and details have been added to the Online Methods section, starting on page 16, particularly in Subsection 4.5, "Model Construction," where we provide a comprehensive description of our refined practices.

5. *In general, significantly more detail is required for the model construction and design as the schematic diagrams alone are not adequate.*

   Response: Thank you for your feedback. In response, we have enhanced

the manuscript by providing a more detailed description of the model construction and training process. We've included specific loss functions and computational formulas as mentioned in the article. These additions have been made in the revised Online Methods section. Additionally, we have updated the schematic diagrams to align with and support this more comprehensive explanation.

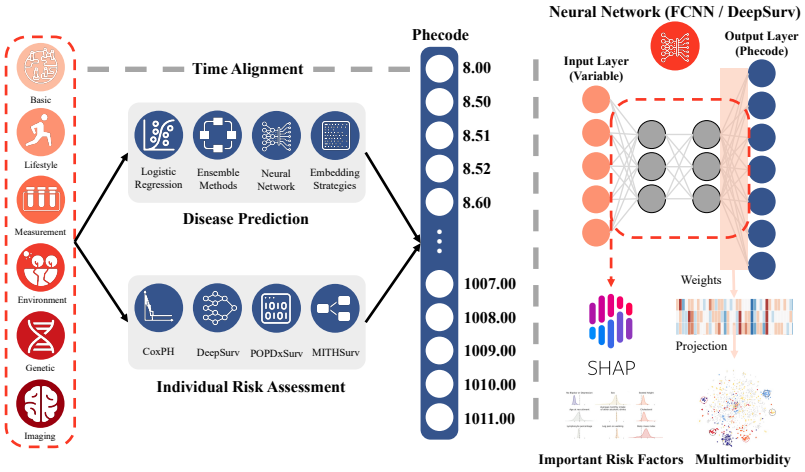

Figure 1: **Construction of UKB-MDRMF**. The entire model construction process is divided into two parts: one for disease prediction and the other for risk assessment. The complete model's input variables are divided into six data types. The prediction model can forecast future occurrences, while the survival model is used to estimate future risk for each phecode. The series of models on the right side of the neural network can be used to show the importance of risk factors using SHAP values and the multimorbidity of different diseases.

6. *The code has been made available and there is information about the organisation of the code on the repository. However, the wider documentation is sorely lacking. There is insufficient information to understand how to use the code to replicate the experiments and to use it for alternative tasks. Python scripts and notebooks are not throughly commented. Scripts and some instructions are provided for the input data preparation but I would not be confided that there is sufficient detail to reproduce the work.*

   Response: Thank you for your feedback! We have thoroughly updated our GitHub repository to include detailed instructions for replicating the experiments, ensuring a comprehensive guide on data preprocessing,

model training, and evaluation. We've also enhanced the Python scripts and notebooks with more detailed comments to aid in understanding and implementation. Furthermore, to accommodate accessibility restrictions of UKB data, we have included examples of generated data that mimic patterns from the original dataset. This addition will facilitate better usage and testing of our model.

Response to Reviewer 3

Thank you very much for your constructive comments and your encouragement! We have revised our paper carefully to address the concerns. The responses to your comments are given in the following.

1. *Thank you for giving me the opportunity to review this manuscript. 'UKB-FOUND' is a cohort study which uses data from UK Biobank data to build a prediction model using demographic, lifestyle factors, biomarkers and GWAS. The model was build to predict the risk of 1560 diseases and several machine learning techniques were used.*

   I will highlight some strengths.

   (a) *Comprehensive set of predictors are considered. Use of UK Biobank allows this kind of analysis wide range of biomarkers, environmental factors and genetic factors are available.*

   (b) *Several methodological approaches were evaluated: logistic regression, Random Forest, XGBoost, Light GBM, FCNN, POPDx and CATI. I am not an expert in ML so critique of choosing these methods is out of my scope.*

   (c) *The presentation of technical details of methods and results from complicated analyses has been done very well with comprehensive coverage of relevant details.*

   (d) *Figures are good quality, missing data analysis is impressive.*

   (e) *For the sake of transparency and replicability, all codes have been provided on github. Appropriate comparison has been made with existing literature in this field.*

   I do have some significant concerns and reservations:

   While technically sound, I am not sure what is the practice or policy related utility of this analysis.

   Response: Thank you for this important comment. We agree that while the technical aspects of the model are essential, its practical and policy-related utility is equally critical. To address this, we have developed a framework to integrate our model into real-world clinical workflows. Specifically, we are designing an AI-powered clinical decision support agent that enables healthcare professionals to interact with the model using

natural language inputs. This tool allows clinicians to explore patient risk profiles, disease trajectories, and the impact of specific risk factors. To demonstrate the practical utility and extension of our model, we provide a preliminary version of this clinical decision support agent for testing. Below are the detailed steps for interacting with the platform for interface examples (Access the platform at: `http://165.227.78.169:8501/`):

(a) **Input OpenAI API Key:**

- In the top-left corner of the interface, locate the input box labeled *"OpenAI API Key"*.
- Paste your OpenAI API Key (you can generate your own API Key from `https://platform.openai.com/api-keys/`).
- Press **Enter** to confirm. Once successful, the system will load a sample dataset, which is anonymized UK Biobank (UKB) data, used here to demonstrate the extended utility of our model.
- If you do not have an OpenAI API Key, we have provided a temporary key for testing within the review system (`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`). Please note that the OpenAI API Key input does not contain any spaces.

(b) **Select a Question or Input Your Own:**

- At the bottom of the interface, you will see a chat input area with three *candidate questions* provided for initial testing.
- Users can also input their own custom questions. However, due to system stability in this preliminary version, we recommend starting with the candidate questions.
- Once a question is submitted, the system will begin processing, and a **"Running"** icon will appear in the top-right corner to indicate that the system is working.

(c) **Model Selection and Processing:**

- If the first question is related to *data description*, the system will directly display the output.
- If the question involves specific model analysis (e.g., disease prediction), the system may prompt you to choose a model with a message such as *"Please Select Disease Diagnosis Model"*.
- Select the desired model for deeper analysis. The system will then resume processing, and the **"Running"** icon will reappear until the task is completed.

(d) **Receive the Output:**

- Once processing is complete, the system will display the final analysis results.

- During the process, you may encounter occasional **red warning boxes**. These warnings can be safely ignored; continue following the outlined steps.

(e) **Important Notes:**

- This platform is a **preliminary version** designed to showcase the extended applications and potential value of our model. The demo is for **review purposes only**.

These steps, combined with the attached images, provide a clear demonstration of the platform's interface and functionality, showcasing how our model can be interactively used to explore patient risk profiles, disease trajectories, and predictive insights.

For broader public use, we are also extending this tool through the Shiny platform (`https://luminite.shinyapps.io/ukb-mdrmf/`), specifically via the Individual Risk Prediction tab, allowing individuals to input their personal health information and receive personalized risk assessments and insights, empowering them to make more informed decisions about their health. We believe this approach bridges the gap between technical performance and real-world application, making the model valuable for both clinical practice and public health policy. Additionally, the Shiny application we have developed already includes individual-level prediction functionalities, and we will continue to enhance its usability to better support decision-making processes.

2. *Most important variables identified had hardly any surprises with majority related to demographic, socio-economic status and metabolic health/mental health. When a relatively less commonly used variable was included, it was hard to ascertain, how this predictor can change/influence practice (e.g. carotid ultrasound).*

   Response: Thank you for your feedback! Indeed, the most significant variables identified in our study, such as demographic, socio-economic, metabolic, and mental health factors, are well-established predictors of health outcomes and align with existing literature. However, our inclusion of less commonly used variables, such as carotid ultrasound, opens up unique opportunities for advancing predictive modeling and enhancing clinical insights.

   Carotid ultrasound, though not yet a staple in routine clinical practice, offers valuable insights for stratifying cardiovascular risk, especially in subgroups with elevated risk profiles. This underutilized predictor can detect subtle risk patterns that might elude more conventional measures. Although not universally adopted in healthcare settings, the demonstrated predictive value of carotid ultrasound in large-scale datasets like the UK Biobank indicates its potential to enhance personalized and precise care in the future.

Additionally, emphasizing the importance of such variables encourages further research and clinical trials to validate their utility and promote their integration into broader patient care protocols. This strategy not only boosts the predictive power of our models but also broadens the range of clinical variables considered in both research and practice.

Crucially, our selection of key variables—both conventional and novel—enables our methodology to be readily adapted to datasets from other centers. This adaptability facilitates the rapid development of predictive or survival models by focusing on the most impactful variables, thereby improving the scalability and practicality of our approach across various healthcare environments.

3. *Baseline health status (i.e. conditions participants had the start of the follow-up) is likely to be one of the strongest predictor of future health trajectory and multimorbidity trajectory. The analysis should have considered this crucial factor.*

   Response: Thank you for your important observation. We recognize that baseline health status, such as pre-existing conditions at the start of follow-up, plays a crucial role in determining future health trajectories and multimorbidity patterns. Currently, our framework indirectly accounts for baseline health status through multiple data sources, including inpatient data, death registers, self-reported health information, and primary care records, as depicted in our model (Fig. 9, page 20 of the main text). These sources furnish a comprehensive view of participants' health histories, which are integrated into the response data (Y) after preprocessing, including Phecode transformation and multi-source integration.

   While our analysis presently focuses on utilizing these multimodal data to explore comprehensive risk prediction and multimorbidity patterns, we have not explicitly modeled baseline health status as a distinct variable. Recognizing the value of this approach, we plan to explicitly incorporate baseline health conditions as independent covariates in both prediction and survival models in future iterations of our framework. This enhancement will not only clarify how pre-existing conditions influence disease progression and multimorbidity trajectories but also improve the interpretability and clinical applicability of our framework.

   Thank you for highlighting this opportunity to further strengthen our model. We will incorporate this enhancement in future studies to refine and expand the applicability of our framework.

4. *The list of outcomes considered (>1000 phenotypes) were too broad. It would have been better to select a small list of outcomes which are most important from public health perspective.*

   Response: Thank you for your feedback. We appreciate your concerns about the broad scope of outcomes considered in our study. While our

model incorporates a wide range of phenotypes to demonstrate its versatility across various disease domains, we acknowledge that focusing on a smaller, more targeted set of outcomes could yield deeper insights into diseases of significant public health importance.

However, the primary goal of our research is to facilitate joint prediction and risk assessment across multiple diseases. Utilizing a broader spectrum of phenotypes not only enhances predictive accuracy—as demonstrated in Fig. 2 (subplots c and f) on page 12 of the manuscript—but also enables a more comprehensive analysis of multimorbidity. This approach allows for the exploration of inter-disease relationships and shared mechanisms that might be overlooked when concentrating on a narrower set of outcomes.

While we believe that this broader approach strikes a balance between robust prediction performance and detailed multimorbidity analysis, we also recognize the benefits of more targeted analyses focused on specific public health priorities. In response to your feedback, future studies could indeed refine our framework to concentrate on smaller, more specific sets of diseases, enhancing its practical applicability to specific public health challenges.

5. *I could not seen any adjustment for multiple testing, very important considering the number of outcomes.*

   Response: Thank you for raising this important point. We acknowledge that multiple testing correction is crucial in studies involving genomic analyses or large-scale association studies, where the goal is to control for false positives across numerous independent hypothesis tests. However, the primary objective of our research is to facilitate joint prediction and risk assessment across multiple diseases. Our approach emphasizes discovery and prediction, aiming to develop a unified framework for comprehensive disease prediction, rather than testing specific hypotheses for major disease categories or individual disease outcomes. While we have included a comparison between the predictive performance of the overall multi-disease prediction model and models focusing on single-disease categories, the goal of our study is distinct from scenarios where multiple hypothesis testing is required. As such, multiple testing correction is not directly applicable in this context. Instead, we focus on the overall predictive performance of the model, which provides a more holistic and robust evaluation of its utility.

Response to Reviewer 1

1. *The authors have done an excellent job at incorporating my and the other reviewers' feedback into their manuscript. I particularly appreciate the incorporation of further biological interpretation in the Discussion, extended network analysis in the Supplement, and the revisions to the Shiny application.*

   The authors have provided a comprehensive, well-documented overview of the code used in this study. I do note that the current repository is currently the only repository saved under the account of "anonymous-researcher01," which at first glance seems like a temporary account to me - I hope that this resource will continue to remain online upon publication of the manuscript.

   Response: Thank you for your kind words about the manuscript and the detailed documentation of the code repository. We understand your concern regarding the temporary nature of the "anonymous-researcher01" account. This account was specifically created to maintain the integrity of the double-blind review process. Upon the manuscript's acceptance, we will transfer the repository to a permanent account associated with our research group or institution. This change will ensure long-term accessibility and stability, allowing the broader community to utilize and reproduce our methods effectively.

Response to Reviewer 2

Thank you very much for your constructive comments and your encouragement! We have revised our paper carefully to address the concerns. The responses to your comments are given in the following.

1. *The authors have provided an updated version of their work, now entitled "UKB-MDRMF: A Multi-Disease Risk & Multimorbidity Framework Based on UK Biobank Data".*

   Foundation model

   *Unfortunately, with clarification that this is *not* a foundation model, the level of novelty of the work is substantially diminished as the results are now, as previously identified, more of a large scale series of prediction models that are coupled in their training. Methodologically the authors have used fully-connected neural networks and DeepSurv which are standard, pre-existing approaches. The specification of the loss functions also suggest that the loss sub-component for each Phecode is independent. It is not clear that there is any coupling between models for different precedes?*

   Response: We thank the reviewer for their thoughtful feedback and the opportunity to clarify the contributions of our work. Below, we address the points raised regarding the novelty of our framework and the coupling between models for different Phecodes.

   (a) **Novelty of the Work**: While our methodology incorporates pre-existing approaches such as fully connected neural networks (FCNNs) and DeepSurv, the novelty of UKB-MDRMF lies in its integration and application for large-scale multi-disease risk prediction and multimorbidity analysis. Unlike traditional models that focus on single diseases or broad categories, UKB-MDRMF simultaneously predicts 1,560 diseases within a unified framework. By leveraging the multimodal UK Biobank dataset, the model captures shared and disease-specific risk factors, offering a broader and more comprehensive perspective on disease mechanisms, multimorbidity, and inter-disease correlations. This integration provides significant insights and moves beyond traditional prediction models by addressing the complex interplay between diseases.

   (b) **Coupling Between Models**: Although the loss functions for individual Phecodes are defined independently, the shared neural network

architecture ensures intrinsic coupling between models for different diseases. Specifically:

- The shared network layers create a unified feature space across all diseases, capturing inter-disease dependencies and shared risk factors. During backpropagation, gradients from each Phecode contribute to optimizing the shared representation, thus facilitating coupling across diseases. For example, the loss function for FCNN aggregates gradients across all Phecodes ($j = 1, \ldots, p$):

$$L^{(\text{FCNN})} = \sum_{j=1}^{p} \sum_{i=1}^{N} \left[y_{ij} \log \hat{g}_\theta^{(j)}(\mathbf{x}_i) + (1 - y_{ij}) \log(1 - \hat{g}_\theta^{(j)}(\mathbf{x}_i)) \right].$$

- Multimorbidity patterns are further revealed through shared and disease-specific variable importance metrics, derived from the trained model, enabling a deeper understanding of inter-disease relationships.

However, due to structural differences between prediction (FCNN) and risk assessment tasks (DeepSurv), model weights are not currently shared across tasks. This is an area we plan to improve in future iterations to enhance cross-task learning and performance.

(c) **Scalability and Impact**: Although UKB-MDRMF is not classified as a foundation model, it introduces a large-scale multi-disease prediction and risk assessment framework that addresses critical clinical and research challenges. Covering 1,560 Phecodes, it facilitates comprehensive exploration of multimorbidity mechanisms, providing an integrated view of health risks and disease relationships. Additionally, the streamlined workflow, from data preprocessing to model implementation, makes the framework scalable and reproducible for similar large datasets.

(d) **Interactive Platform**: To improve accessibility and usability, we developed an interactive platform (`https://luminite.shinyapps.io/ukb-mdrmf/`) for clinicians and researchers. This platform enables detailed exploration of model predictions, variable importance, and comorbidities across diseases and risk factors, bridging the gap between advanced modeling and practical application. This resource ensures that the model's insights are actionable and transparent for a diverse range of users.

(e) **Contributions Beyond Prediction**: UKB-MDRMF offers more than high predictive accuracy; it provides novel insights into disease interrelationships and multimorbidity mechanisms. By analyzing disease-disease and disease-risk factor connections, the framework facilitates hypothesis generation and uncovers shared disease pathways. This dual focus on prediction and discovery marks a significant improvement over traditional models, which often lack interpretability and clinical utility in understanding multimorbidity.

We appreciate the opportunity to clarify these aspects to better emphasize the unique contributions and impact of our work.

2. *Validation*

   The authors used All of Us for validation purposes. All of Us is an emerging population research cohort but I should clarify that my original query was more related to routinely collected point-of-care clinical data. There will clearly be similarities in the data collected between population research cohorts.

   Nonetheless, the use of All of Us is a promising addition by the authors. However, I was confused by the statement "we retrained the FCNN and DeepSurv models from the UKB-MDRMF using the All of Us data". If the intention is to validate the models then shouldn't the trained model from UKB be transported and used as-is on All of Us without retraining? If the model is retrained, this would be a validation of the process rather than the models.

   Response: Thank you for raising this important point regarding the validation approach using All of Us data. Due to substantial differences in data structures and variable definitions between the UK Biobank and All of Us datasets, direct model transportation is not feasible. For instance:

   - While both datasets share variables with similar conceptual meanings (e.g., age, BMI, smoking status), the specific definitions, data formats, and categorical encodings often vary significantly. For example, options for categorical variables such as smoking status might differ in their granularity or representation across the two datasets.
   - Additionally, there are differences in the inclusion and representation of certain demographic, clinical, and environmental variables between the two cohorts, reflecting their distinct data collection methodologies.

   As a result, aligning all variables between the two datasets perfectly is challenging, and the direct use of a model trained on UK Biobank data would likely lead to poor performance due to these mismatches. Therefore, instead of transporting the models, we focused on transferring the process underlying UKB-MDRMF. Specifically:

   - The entire data preprocessing workflow and framework design, including several multimodal feature preprocesses, missing variable imputation, and multi-disease modeling using Phecodes as outcomes, were directly applied to the All of Us dataset.
   - For model application and selection, as well as the categorization of Phecodes, we utilized the same framework and codebase from UKB-MDRMF to maintain consistency across the overall modeling process. This allowed us to compare results directly and evaluate how potential structural differences between biobank datasets influence

the final model construction and prediction performance for various disease types.

The retrained models demonstrated results consistent with those obtained from the UK Biobank data, highlighting the robustness of the UKB-MDRMF process and its applicability across datasets. While this approach does not directly validate the original models, it provides strong evidence for the transferability of the framework itself, including its preprocessing pipeline, variable handling methodology, and modeling design.

We acknowledge that validation using routinely collected point-of-care clinical data would provide additional insights into the transportability of the models without retraining. This remains an important area for future work, and we will explore opportunities to extend UKB-MDRMF to such datasets.

3. *Shiny app*

   The authors have provided a limited demonstration of the individual risk prediction models. While this addition is welcomed, I did not feel that it was done particularly well. The survival plots are not described, what does the "survival probability" on the vertical axis mean? Is this the risk of the disease? If so, why do the survival curves seem to asymptote after 80 years of age at a non-trivial probability?

   Response: Thank you for highlighting the areas where our explanation could be more detailed. We provide a more comprehensive clarification below.

   For each disease, we calculate the baseline cumulative risk, $r_{base}(t)$, which increases over time based on the number of cases and the onset time for each individual. Next, we compute the relative risk, r_{rel}, for each individual compared to the baseline, allowing us to determine their individual cumulative risk as:

$$r_{individual}(t) = r_{base}(t) \times r_{rel}.$$

   For instance, if an individual's relative risk for cataracts is 1 (i.e., equal to the baseline risk), their personal cumulative risk curve is shown in Figure 1. This curve increases with age, indicating that as an individual grows older, their risk of developing the disease also rises. A steeper increase in the curve suggests that the risk of developing the disease escalates significantly with age.

   The curve displayed on Shiny represents a survival function, which indicates the probability of remaining "disease-free" (or not developing the disease) over time. This survival probability decreases as age increases and corresponds inversely to the individual risk curve. The survival probability is computed using the following equation:

$$P_{surv}(t) = e^{-r_{individual}(t)},$$

as illustrated in Figure 2. The survival curve gradually declines with age, reflecting the increasing risk of disease onset.

Regarding the observation that the survival curves asymptote at a non-trivial probability after the age of 80, this phenomenon occurs because not all individuals develop the disease within their lifetime. Some individuals may never be diagnosed with the condition, leading to a non-zero survival probability at older ages. This effect is more pronounced for diseases with a lower lifetime incidence rate.

We have provided further detailed explanations of these visualizations and their interpretations in the Shiny application to enhance clarity.

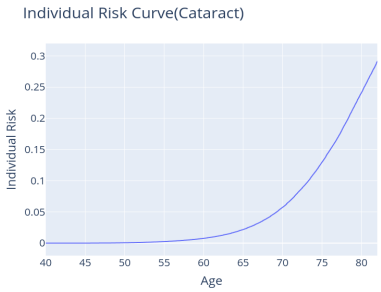

Figure 1: Individual risk curve.

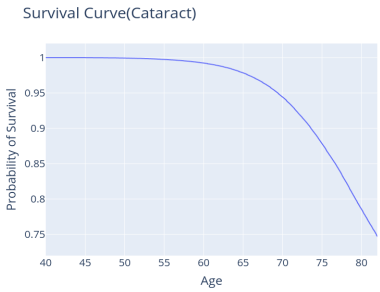

Figure 2: Survival curve.

4. *Methodological details*

   The "enhanced" Figure 11 is still a high-level schematic. Unless I have missed it, the authors have not really provided an item-by-item report against the TRIPOD checklist which could be incorporated into Supplementary Information. For example, Item 4b iii) "The length of follow-up and prediction horizon/time frame are reported", is a checklist item where the authors would benefit from giving a more explicit response. It is unclear what the prediction horizons are for the framework. It is unclear how deaths post-enrolment are handled. The removal of disease occurrences prior to the covariate collection leads to significant biases and limitations in the prediction model particularly if pre-existing diseases highly correlate with others. The authors have not considered competing risk implications? The methodological description could be improved significantly further.

   Response: Thank you for your feedback and for pointing out areas where the methodological details could be improved. Below, we provide a detailed response addressing each of the concerns raised:

(a) **Enhancements to Fig. 11**: We acknowledge that the current Fig. 11 on page 23 of the manuscript provides only a high-level overview. To address this, we have incorporated additional details on time alignment and provided a more detailed description of the time alignment process. Additionally, we have introduced the loss functions into the workflow diagram, illustrating how backpropagation (BP) is used to update network weights for downstream analyses. The updated figure offers a comprehensive explanation of the time alignment process and prediction data construction. For the revised version, please refer to Figure 3.

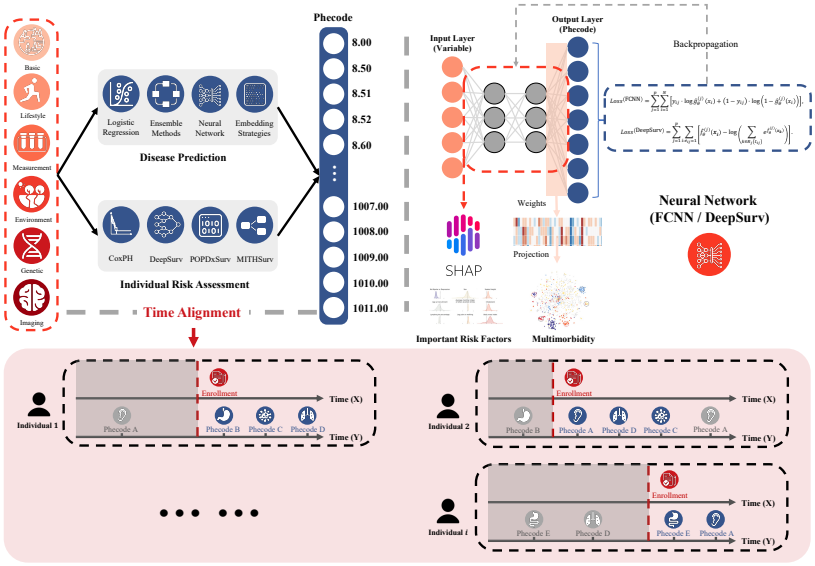

Figure 3: **Pipeline for constructing UKB-MDRMF.** First, a time alignment process is applied to ensure that disease occurrences post-date the baseline data. The red-shaded section at the bottom illustrates the alignment process for different individuals, where the red dashed line represents the enrollment time. Features such as basic, lifestyle, and other characteristics are collected at the time of enrollment. Phecodes recorded before enrollment are marked in gray and treated as missing values during model training, ensuring they are not used for training purposes. After integrating Phecode data from multiple sources, only the earliest occurrence of the same Phecode post-enrollment is retained. Next, various multi-disease prediction and risk assessment models are applied for comprehensive evaluation. These models are trained separately using distinct loss functions. Finally, model interpretability analysis is performed, incorporating associations between different diseases and risk factors for integrated analysis. Variable importance results are derived from all model weights, while multimorbidity relationships are inferred from the embeddings in the penultimate network layer.

(b) **TRIPOD Checklist**: To ensure transparency and reproducibility, we have conducted an item-by-item assessment following the TRIPOD checklist for the methodology section. The detailed responses are outlined below. Specifically:

- For Item 4b iii in TRIPOD Checklist regarding the length of follow-up and prediction horizon, we now explicitly report the time frames used in our models. The follow-up period in our study starts from the initial UKB data collection and extends until September 2023. In this revised manuscript, we have updated the dataset to incorporate more recent records, thereby expanding the number of diseases analyzed and re-evaluating all results accordingly. Despite these updates, the overall findings remain consistent with our previous results. For survival analysis, time-to-event outcomes (e.g., disease onset or death) are modeled continuously within the follow-up period. To ensure meaningful survival times, we use age as the time metric rather than the time since enrollment, a common approach in large-scale biomedical datasets. Additionally, events occurring at the end of the follow-up period (September 2023) are treated as right-censored. The relevant content has been supplemented and described in Section 4.5.1, Model Construction Process, on page 22 of the manuscript.
- Detailed TRIPOD Checklist Responses:
  - 4a. Source of Data: Our study is based on the UK Biobank (UKB), a large-scale biomedical dataset that includes longitudinal health data from approximately 500,000 participants aged 37–72 at recruitment. The detailed dataset description is in Section 4.1, Data Description, on page 16 of the manuscript. For external validation, we utilized data from the All of Us Research Program, which provides diverse demographic and clinical records from a broader U.S. population. A comprehensive description of this dataset and its application in our study is provided in Section 2.7, Multicenter Validation Using All of Us Data, on page 7.
  - 4b. Study Dates and Follow-up Period: The data collection period for UK Biobank began in 2006, with continuous updates and follow-ups. For this study, we used the most recent data available as of September 2023. The follow-up period spans from each participant's enrollment up to their last recorded update. Further details on follow-up duration and methodology can be found in Section 4.5.1, Model Construction Process, on page 22 of the manuscript.
  - 5a. Study Setting: The study includes data from both UKB and All of Us, covering multiple healthcare settings such as primary care, secondary care (specialist and hospital visits),

and general population-based registries. A detailed descriptions of these datasets is provided in Section 4.1 and Section 4.3 on page 16 and Section 4.5.1 on page 22.

- 5b. Eligibility Criteria for Participants: Participants were included in the study if they met the following criteria: availability of baseline data, including demographic, lifestyle, environmental, genetic, and imaging information; complete longitudinal health records post-enrollment, allowing for disease onset prediction; and no missing critical predictor variables (Phecodes) required for model training. In addition to the dataset descriptions in Section 4.1, further details on eligibility and preprocessing procedures are provided in Section 4.3, Data Preprocessing, on page 17 of the manuscript.
- 6a. Outcome Definition: The primary outcomes are disease diagnoses, defined using Phecodes extracted from UKB's hospital inpatient records, death registries, primary care data, and self-reported medical histories. For disease prediction and risk assessment (survival models), we consider all occurrences of the disease post-enrollment. Further details can be found in Section 4.3.3, Process of Phecodes, on page 19 and Section 4.5.1, Model Construction Process, on page 22.
- 7a. Predictors and Measurement Timing: Predictor variables include demographic factors, lifestyle habits, various clinical measurements, environmental exposures, genetic risk scores, and imaging-derived features. Except for imaging data, which is considered separately, these variables were collected at baseline (enrollment), ensuring that disease occurrence happens after predictor collection. Additional details are available in Section 2.2, Data Preprocessing, on page 3, and Section 4.3, Data Preprocessing, on page 17.
- 8. Sample Size Determination: Given the large-scale nature of UKB, we included all available participants who met the eligibility criteria to maximize model generalizability. Similarly, in the All of Us dataset, we utilized as many samples as possible, comprising a population of over 200,000 individuals.
- 9. Handling of Missing Data: We performed multiple imputation for missing continuous variables and category-specific imputations for missing categorical data. Further details on missing data processing are provided in Section 4.4, Missing Data Processing, on page 20.
- 10a. Predictor Handling in the Model: Predictors were processed using standardization for continuous variables, categorical encoding for non-numeric features, and feature selection based on relevance to disease outcomes. Additional methodological details can be found in Section 4.3, Data Pre-

processing, on page 17.

- 10b. Model Development and Validation: We developed and validated two types of models. For disease prediction (classification models), we used logistic regression, fully connected neural networks (FCNN), and ensemble methods. For risk assessment (survival models), we implemented Cox Proportional Hazards, DeepSurv, and multi-task survival models with the embedding matrix. The variable importance and multimorbidity analysis for disease prediction were conducted using FCNN, while DeepSurv was used for survival analysis. Validation was performed using both the UKB test set (for direct model transfer validation on the same feature set) and the All of Us dataset (to assess the generalizability of the modeling process). Further methodological details can be found in Section 2.3, Model Construction, on page 4, and Section 2.7, Multi-Center Validation Using All of Us Data, on page 7.
- 10d. Performance Metrics for Model Evaluation: For disease prediction models, we used the Area Under the Curve (AUC) as the primary performance metric, as described in Section 2.3.1, Disease Prediction, on page 4. For survival models, we used the concordance index (C-index) to evaluate time-to-event predictions, as detailed in Section 2.3.2, Risk Assessment, on page 5.

(c) **Prediction Horizons for the Framework.** Our framework considers all diseases that occur after enrollment and after the collection of baseline covariates as part of the prediction horizon. Since the target disease Y occurs after individual characteristics X are collected at enrollment, we can establish both prediction and survival models. This setup ensures that the framework supports forward-looking disease risk assessment rather than merely identifying retrospective associations between features and diseases. Further details can be found in Section 4.5.1, Model Construction Process, on page 22.

(d) **Handling Deaths Post-Enrolment**: In our previous manuscript version, we recorded disease-specific deaths but did not account for the censoring effect of death on other diseases. In this revised version, we have comprehensively incorporated this aspect. Specifically, deaths occurring post-enrollment are treated as censored events in the survival models. If an individual dies without developing a specific disease of interest, their time-to-event data for other diseases is censored at the time of death. This adjustment has been updated in Section 4.5.1, Model Construction Process, on page 22. Given that these changes impact data processing, we have updated all model results accordingly. The revised results remain consistent with our previous findings, confirming the robustness of our approach.

(e) **Bias Due to Removal of Pre-Baseline Diseases**: To ensure that our predictions focus on post-enrollment disease onset, our framework excludes disease occurrences before baseline covariate collection. While this approach enhances the validity of our predictive models and aligns with real-world clinical applications, we acknowledge that it may introduce biases, particularly when pre-existing conditions are strongly correlated with future disease risks. As a future direction, we plan to incorporate pre-baseline disease occurrences as additional covariates to capture potential interactions and correlations with post-enrollment diseases. This discussion has been included in Section 3, Discussion, on page 10, where we outline study limitations and future directions.

(f) **Competing Risks**: We appreciate the reviewer's concerns regarding competing risk implications. While our current framework does not explicitly model competing risks, our joint learning approach, which simultaneously trains on all Phecodes, allows the neural network to leverage shared representations across diseases. This implicit feature-sharing mechanism enables the model to capture interdependencies among diseases. However, we recognize that explicitly modeling competing risks using targeted survival models could further enhance the interpretability and precision of our predictions. As part of our future work, we plan to extend our survival models to incorporate competing risk frameworks to better account for interactions between multiple disease outcomes. In the revised manuscript, we have explicitly discussed this limitation and its potential implications in Section 3, Discussion, on page 10.

All these updates related to the methodology, discussion, and diagram refinements have been integrated into the revised manuscript. We believe these revisions effectively address the reviewer's concerns and significantly improve the methodological transparency and rigor of the study. We sincerely appreciate your valuable feedback, which has helped us enhance the clarity, precision, and completeness of our work.

5. *Software repository*

   The files on the Github have been updated and documentation improved.

   Response: Thank you! We will continue to optimize and update the code to ensure its long-term availability and usefulness for the research community.

6. *Overall comments*

While there are some improvements to the presentations. I found some substantial details missing, in particular, those related to methods. This continues to prevent me from a more comprehensive understanding of the paper, the appropriateness of the analyses and the results. While the change from a foundation model to predictive framework results in a less novel concept, the multi-disease approach would be potentially interesting. However, it is not clear as stated above, that there is any dependence introduced between diseases.

Response: Thank you for your valuable feedback. We appreciate the opportunity to clarify the methodological aspects of our study and address the concerns regarding disease dependence in our multi-disease framework.

We have expanded the methodological section to provide a more comprehensive and detailed explanation of our approach, including improvements in the description of time alignment, predictor processing, and survival modeling. Specifically, we have incorporated an item-by-item assessment based on the TRIPOD checklist to enhance transparency and reproducibility. Further refinements, such as detailed explanations of prediction horizons, handling of deaths post-enrollment, and a discussion of the limitations related to competing risks, have also been explicitly addressed in the revised manuscript.

Regarding the concern about disease dependence, while our current framework does not explicitly model dependencies between diseases using disease network-based approaches, our joint learning paradigm inherently captures shared representations across diseases. By simultaneously training on over 1,500 Phecodes, the model learns common features that link related diseases, allowing interdependencies to be reflected in the learned embeddings. Additionally, our multimorbidity analysis, which leverages the learned weights in the penultimate network layer, provides further insights into potential disease-disease relationships, as detailed in Section 2.6, Multimorbidity and Trend Analysis of Disease Risks, on page 6, and Appendix D.3.1, Multimorbidity Network. To further enhance disease dependence modeling, we plan to explore competing risk frameworks and multi-task learning strategies in future iterations of our work.

All these updates and refinements have been incorporated into the revised manuscript, specifically in Sections 2.3, 3, and 4.5, as well as in supplementary materials where additional methodological details have been provided. We believe these revisions significantly improve the clarity, completeness, and rigor of our study, and we sincerely appreciate your insightful suggestions that have helped enhance our work.

7. *The code repository has been substantially revised and is significantly more usable.*

Response: We sincerely appreciate your positive recognition. Moving forward, we are committed to continuously updating and refining the code to ensure its long-term accessibility and utility for researchers.

Response to Reviewer 3

Thank you very much for your constructive comments and your encouragement! We have revised our paper carefully to address the concerns. The responses to your comments are given in the following.

1. *Thank you for your response and revised manuscript. The revised submission has improved significantly and a greater deal of clarity has been added to the analysis and subsequent presentation. However, the two crucial points I raised in the previous review have not been addressed.*

   Firstly, the baseline health status has not been considered in the analysis which is very likely to have a hugely significant impact on the health outcomes studied.

   Response: Thank you for pointing out the importance of baseline health status. We would like to clarify that in our modeling approach (both prediction and survival models), only health outcomes that occurred after the baseline enrollment time were considered for model construction. This ensures that individuals included in the analysis were considered healthy with respect to the corresponding diseases at baseline. Specifically, as shown in Fig. 11 on Page 22 of the manuscript, the response variables were carefully aligned with their corresponding recording times, and only post-enrollment occurrences of ICD codes or health outcomes were included in the model. By excluding pre-baseline records, we minimize potential confounding effects of pre-existing health conditions, ensuring that the predictions reflect the natural progression of diseases starting from a healthy baseline. This design addresses concerns about the impact of baseline health status on the studied health outcomes.

2. *Secondly, a very wide variety of clinical outcomes are considered. Authors argue why this approach is relevant for their analysis. However, a sensitivity analysis with much smaller number but more clinically relevant outcomes will be more useful.*

   Response: Thank you for your thoughtful suggestion regarding sensitivity analysis with a smaller number of clinically relevant outcomes.

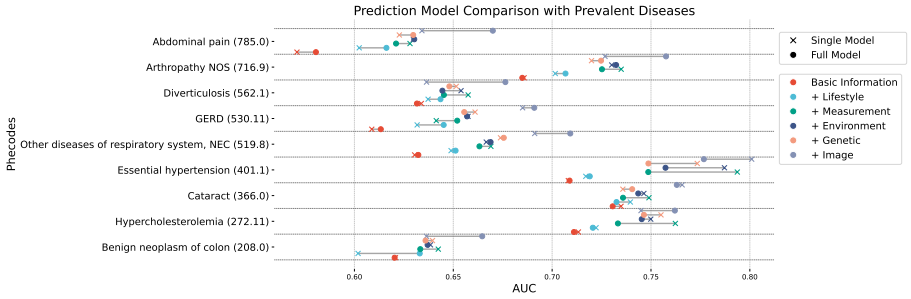

Figure 1: **Comparison of AUC performance for disease prediction in Single vs. Full Model (UKB-MDRMF) on the test set.** We evaluated nine high-prevalence diseases by comparing a model trained solely on individual diseases (Single Model) with the full model trained jointly on 1,560 Phecodes (UKB-MDRMF). Dots represent the results of the Full Model, while crosses indicate the performance of the Single Model. Different colors correspond to the sequential addition of different data categories.

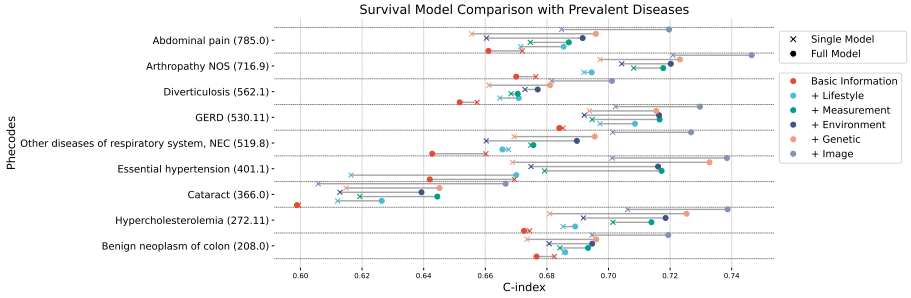

Figure 2: **Comparison of AUC performance for individual risk assessment in Single vs. Full Model (UKB-MDRMF) on the test set.** We evaluated nine high-prevalence diseases by comparing a model trained solely on individual diseases (Single Model) with the full model trained jointly on over 1,560 Phecodes (UKB-MDRMF). Dots represent the results of the Full Model, while crosses indicate the performance of the Single Model. Different colors correspond to the sequential addition of different data categories.

While our framework (UKB-MDRMF) simultaneously models over 1,500 Phecodes, the results for each disease can also be independently extracted and analyzed, allowing for a more detailed assessment of specific clinical outcomes. As shown in Figures 1 and 2, we compared the predictive performance of the full model, UKB-MDRMF (joint modeling of all Phecodes), with single-disease models trained independently for nine high-prevalence diseases, considering both disease prediction and survival anal-

ysis tasks. The results indicate that for disease prediction (Figure 1), joint modeling generally performs comparably to or better than single-disease modeling, except for Essential Hypertension (Phecode: 401.1) and Hypercholesterolemia (Phecode: 272.11), where the single-disease models showed slightly better performance. In contrast, for risk assessment (Figure 2), nearly all solid dots are positioned to the right of the "x" marks, demonstrating that the UKB-MDRMF framework significantly improves C-index performance through joint modeling. This suggests that incorporating shared information across diseases and risk factors enhances overall predictive accuracy, particularly in survival analysis. These additional analyses have been incorporated into the supplementary materials of the manuscript.

Additionally, we conducted analyses focusing on individual disease categories to evaluate the performance of our framework in a more targeted manner, as shown in Fig. 2c and Fig. 2f on Page 12 of the manuscript. These analyses highlight the robustness of our approach and demonstrate that the collective modeling framework does not diminish the ability to focus on smaller, clinically relevant subsets of diseases.

Beyond individual disease prediction, an important advantage of our framework is its ability to model multimorbidity, capturing inter-disease relationships that single-disease models cannot. By training on a diverse set of Phecodes simultaneously, UKB-MDRMF provides valuable insights into disease co-occurrence patterns and shared risk factors, allowing for a more holistic understanding of disease progression and interaction mechanisms, as discussed in Section 2.6, Multimorbidity and Trend Analysis of Disease Risks, in the manuscript.

As part of our ongoing research, we will continue to explore clinically relevant outcomes in greater depth and conduct further analyses to enhance the clinical utility of our findings.