DeepEXPOKE: A Deep Learning Framework with Polygenic Risk Scores as Knockoffs for Deconvoluting Genetic and Non-Genetic Exposure Risks in Sepsis and Coronary Heart Disease

Aditya Sriram[1], Rebecca Bohn[1], Kate Kernan[2], Joseph Carcillo[2], Soyeon Kim[3], Hyun Jung Park[1]

1. Department of Human Genetics, University of Pittsburgh, PA 15213, USA
2. Division of Pediatric Critical Care Medicine, Department of Critical Care Medicine, Children's Hospital of Pittsburgh, University of Pittsburgh, Pittsburgh, PA 15260, USA
3. Division of Pediatric Pulmonary Medicine, Children's Hospital of Pittsburgh, Pittsburgh, PA 15224, USA

ABSTRACT

The exposome refers to the totality of environmental, behavioral, and lifestyle exposures an individual experiences throughout one's lifetime. Due to the modifiability of exposures, identifying the risk exposures on a disease is crucial for effective intervention and prevention of the disease. However, traditional analytical methods struggle to capture the complexities of exposome data: nonlinear effects, correlated exposures, and potential interplay with genetic effects. To address these challenges and accurately estimate exposure effects on complex diseases, we developed DeepEXPOKE, a deep learning framework integrating two types of knockoff features: statistical knockoffs (statKO) and polygenic risk score as knockoffs (PRSKO). DeepEXPOKE-statKO controls exposure correlation and DeepEXPOKE-PRSKO isolates genetic effects, while both can capture nonlinear effects. We applied DeepEXPOKE to predict outcomes of two significant diseases with distinct etiology and clinical presentation: sepsis and coronary heart disease (CHD), demonstrating its performance in comparison to existing machine learning methods. Furthermore, both DeepEXPOKE-PRSKO and DeepEXPOKE-statKO identified metabolites such as glucose and triglycerides as risk factors for sepsis and suggested that their effects are primarily at the non-genetic level, consistent with the role of metabolites in responding to environmental factors. Additionally, DeepEXPOKE-PRSKO uniquely identified asthma as a sepsis risk factor and suggested its effect is partially at the genetic level, offering insights into the conflicting associations observed between the genome data studies and patient data analysis regarding asthma and sepsis risk. Overall, DeepEXPOKE offers a novel DNN approach for identifying and interpreting exposure risk factors, advancing our understanding of complex diseases.

INTRODUCTION

The exposome refers to the totality of environmental, behavioral, and lifestyle exposures an individual experiences throughout their lifetime[1]. Thus, they are modifiable primarily through changes in those aspects of one's life. Due to the modifiability, identifying the risk exposures is crucial to effectively intervene and prevent complex diseases[2-4]. Recently, population-based health databases have been organized to provide a wide range of exposures collected from many subjects, together with their genome data. For example, the UK Biobank (UKBB) provides a diverse range of exposures of approximately 500,000 participants from the United Kingdom[5,6]. Thus, an emerging task is to effectively explore the large databases and identify risk exposures linked to complex diseases. Especially, to estimate the modifiable portion of the risk separated from genetic influences, several analytical

approaches considering both genetic variants and exposures can be applied[7]. First, regression models are used to assess interaction terms pairing genes and environmental factors[8-10]. For example, Liu et al. explored how the interactions between particular genetic variants and environmental exposures shapes brain structure and function[8]. Second, polygenic risk score (PRS), combined with environmental data, improves disease risk prediction and examines the interactions between genetic and environmental factors. For example, Khera et al. demonstrated that integrating PRS with lifestyle factors, such as diet and exercise, better stratify cardiovascular disease patients for Cox proportional-hazard models [11]. Third, Mendelian Randomization (MR) leverages genetic variants as instrumental variables to infer causal relationships between exposures and outcomes. This approach has been particularly useful in distinguishing causality from correlation in observational studies[7, 12, 13].

While the studies highlight the significant contributions of the interactions between genetic factors and exposures, multiple limitations have been discussed of the current studies for accurate estimation of exposure risk[14, 15] (**Fig. 1A**). First, exposures affect outcomes in highly nonlinear fashions due to complex interaction patterns that involve the threshold effect, inverse relationship, or saturation effect[16]. However, previous studies have been based on linear models, such as regression[17-20]. Second, exposures are often correlated with each other[5], making it difficult to attribute observed health outcomes to specific exposures. For example, if both poor diet and lack of exercise are both linked to a disease, current methods are not designed to separate their effects, making it difficult to determine whether to focus on diet or exercise[6]. Third, genetic effects often influence a person's likelihood of engaging in certain exposures, such as dietary choices or physical activity[21]. If these exposures impact health, it is challenging for existing methods to separate their direct effects from genetic influences. MR methods help separate effects, but mostly based only on linear models, and thus not able to capture the nonlinear effects. The only deep neural network (DNN) model in MR approach, DeepMR[22], is to link mutagenesis marks in a small genomic region to a related trait. It is not suitable for genome-exposome data analysis, so it was not considered in this manuscript.

To address each of the limitations, we propose a novel deep learning for exposome analysis using knockoff estimation (deepEXPOKE) (**Fig. 1B**). First, to model the nonlinear interaction patterns between disease outcomes and their exposures, we will employ multiple layers of nonlinear activators in the deep neural network (DNN) framework. We previously showed that this DNN approach successfully validates known nonlinear interactions and identifies novel therapeutic targets for complex diseases, such as sepsis and breast cancer[23]. Second, to control correlations between exposures, we propose to employ knockoff features in the DNN approach. Knockoff features are synthetic and noisy copies of the input variables, which 1) resemble the correlation structure of the input variables but 2) are conditionally independent of the outcome, given the input variables[24]. Due to these two mathematical conditions that control the correlation structure of the input variables, knockoff features have been effective in dealing with variable correlation issues in DNN approaches[23, 25], although they have not been applied to exposure data yet. Third, to separate exposure effects from genetic influence, we propose a novel type of knockoff based on polygenic risk scores (PRS). Standard knockoffs are statistically generated and thus do not represent the genetic influence on the exposures. On the other hand, PRS, the aggregate effect of multiple genetic variants to a specific trait, represents a baseline level of genetic influence while satisfying the two mathematical conditions to be knockoff given above. First, if two exposures share genetic factors, they will correlate in PRS, satisfying the first mathematical condition to be knockoff. Second, PRS or a subset of genetic variants in the PRS model are used to infer causal relationships

between exposures and health outcomes[26]. This is because PRS is independent of the outcomes conditioned on the exposures, which demonstrates that PRS satisfies the second mathematical condition to be a knockoff variable.

Based on these ideas, we developed deepEXPOKE with two types of knockoffs: statistical knockoff (statKO) and PRS knockoff (PRSKO). We tested deepEXPOKE on two complex diseases with distinct clinical presentation, sepsis and coronary heart disease (CHD) (**S. Fig. 1**). Sepsis, accounting for 20% of all deaths worldwide[27], is initiated by viral/bacterial infection and involves dysfunctional immune system. Exposures such as nutritional deficiencies or stress can raise sepsis risk[28, 29], while genetic factors can also influence exposures. On the other hand, coronary heart disease (CHD), which involves blockage of the coronary arteries, is influenced by exposures, such as diet, smoking, and physical activity[30]. Using the disease data downloaded from UKBB, we will validate PRS as a type of knockoff based on the mathematical properties. Then, we will evaluate DeepEXPOKE in predicting sepsis and CHD incidence and mortality in comparison to other machine learning methods. In the process, we will also identify individual exposure risk factors and interpret their roles in clinical outcomes. DeepEXPOKE also offers alternative explanations for previous inconsistencies, potentially arising from biases such as exposure correlation or genetic effects. This includes complex relationships like the 'hypertension paradox[31], where hypertension is paradoxically associated with better outcomes in certain contexts. Together, deepEXPOKE provides the first deep-learning framework to elucidate the exposure risk to complex diseases by addressing collinearity and genetic factors.

RESULTS

DeepEXPOKE identifies exposure risk factors in the DNN framework with novel PRS knockoff features.

To control for correlation with other exposures and underlying genetic effects when estimating exposure risk for complex diseases, deepEXPOKE incorporates a DNN framework with two different types of knockoff features: statistical knockoffs (statKO) and PRS as knockoffs (PRSKO) (**Fig. 1B**). The statistical knockoffs are constructed to mimic the correlation structure between input variables[24], while PRS values are used to estimate genetic contributions to the outcome. Since both correlation and genetic effects can influence disease risk through complex, nonlinear pathways across multiple biological layers—such as the genomic, transcriptomic, and proteomic levels—DeepEXPOKE integrates these knockoff features into a DNN model. DNNs utilize multiple layers of nonlinear activators to capture nonlinear patterns from data[23]. Specifically, DeepEXPOKE generates both types of statistical and PRS knockoff for each input feature as a novel type of knockoff and pairs them with the input feature to feed into the DNN. After training on these input-knockoff pairs to predict outcomes, DeepEXPOKE estimates the nonlinear effect sizes using the trained knockoff values, as the trained knockoff values capture the contribution of collinearity or underlying genetic effects to the prediction, respectively (**Fig. 1C**). Specifically, risk factors identified both by deepEXPOKE-PRSKO and deepEXPOKE-statKO exert effect not through underlying genetic influences and collinearity, and thus are exposure-level risk factors. In the same sense, those only by deepEXPOKE-PRSKO exert effect not through underlying genetic influences, affecting by collinearity and/or at the exposure level. However, if they are at the exposure level, then deepEXPOKE-PRSKO should also identify their risk, suggesting they likely exert their effect by collinearity (correlation with other exposures). Lastly, in the same sense, those only by deepEXPOKE-statKO exert effect by genetic influence.

While statistical knockoffs have been applied in DNN frameworks for variable selection and causal inference[23, 25], we are the first to propose their use in addressing exposure collinearity to our knowledge. More importantly, the implementation of PRS as a knockoff within the DNN framework is a novel approach that, to our knowledge, has not been explored in any prior research. Overall, DeepEXPOKE is a scientifically innovative and clinically valuable deep learning framework designed to distinguish how exposure effects contribute to phenotypes. This model not only deepens our understanding of complex traits but also holds significant clinical potential by identifying more precise biomarkers and therapeutic targets to address disease management and treatment.

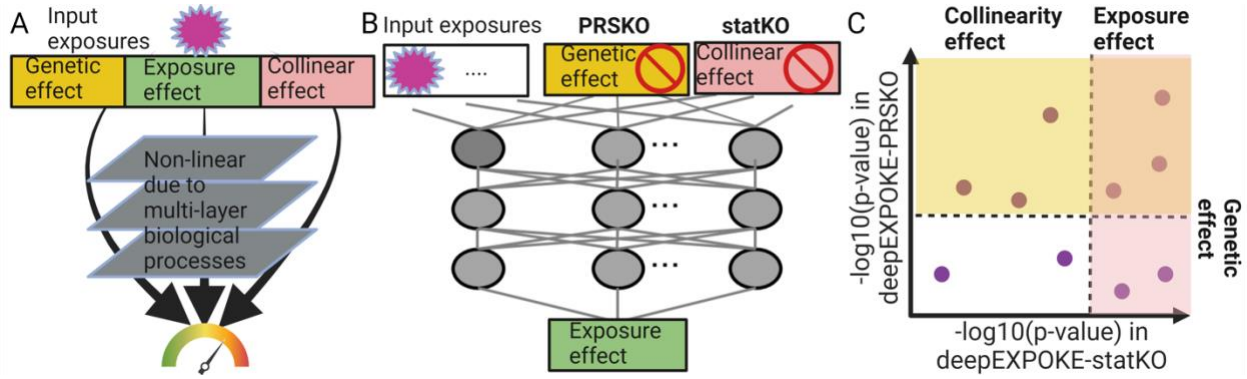


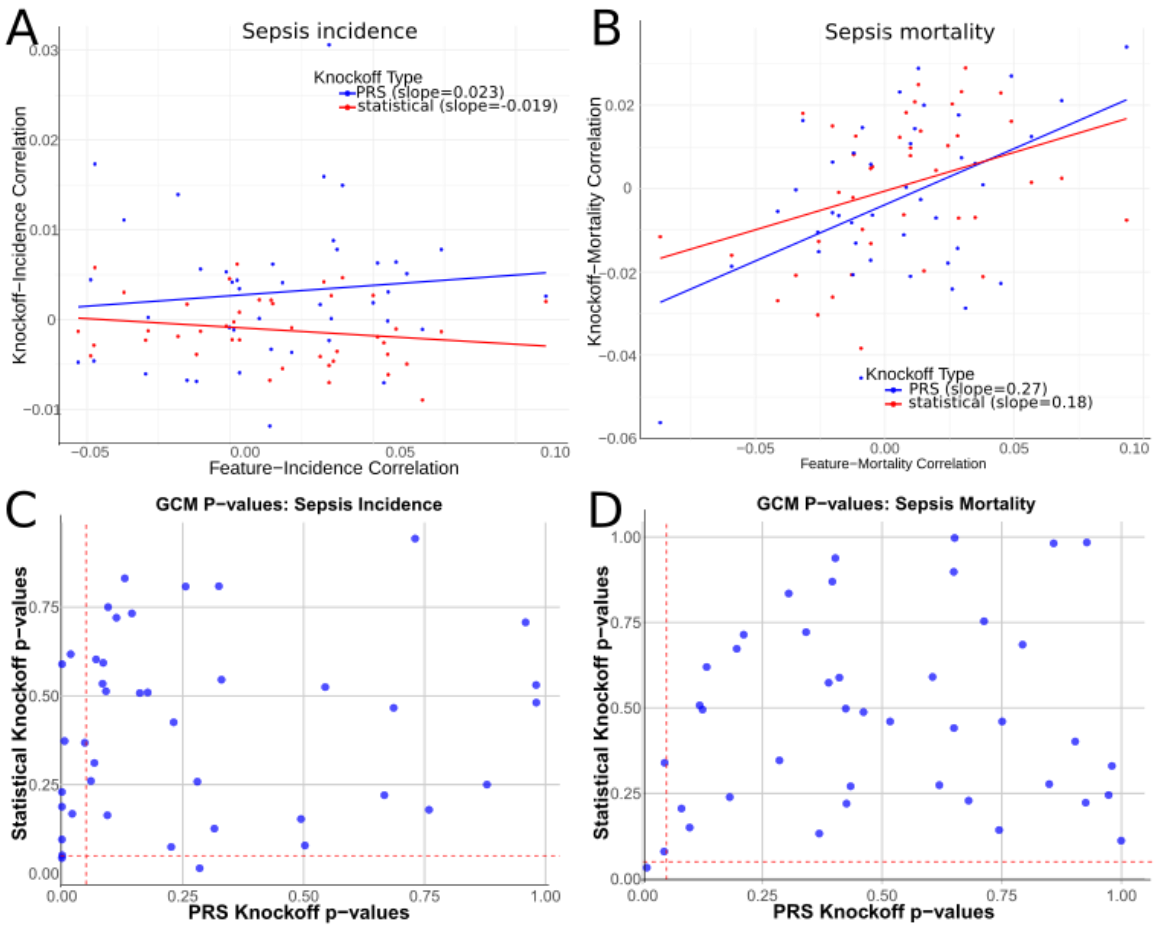
Figure 1 - DeepEXPOKE identifies exposure risk factors in the DNN framework with novel PRS knockoff features. A) Illustration showing how an exposure exerts effect by passing the genetic, exposure, and collinear effect through multi-layer biological processes. **B)** deepEXPOKE architecture that estimates the exposure effect by controlling the genetic and/or collinear effect via PRS knockoffs (KO) and statistical (stat) knockoffs, respectively. **C)** Illustration showing the effect size in PRS and/or stat KO indicates collinearity, genetic, or exposure effect. Note that if an exposure effect is identified significant by both PRSKO and statKO, then the effect is estimated to be at the exposure level.

PRS satisfies the mathematical properties to be a knockoff variable.

To validate PRS as a novel type of knockoff, we selected 3,103 participants with history of sepsis, based on established ICD-10 codes (see Methods), and 79,791 healthy controls from UKBB. We then identified 42 clinical and epidemiological input exposures reflecting general health status unaffected by temporal factors (see Methods, **S. Fig. 1**). For each exposure, we estimated a PRS for each input exposure using the pgsc_calc[32] (see Methods) method and generated statistical knockoffs using the 'model-X' framework across all UKBB samples[24]. To assess whether the first mathematical condition in the PRS values was satisfied—correlation with the input values for representing an outcome—we evaluated the correlation between the PRS values and the outcome and compared it with the correlation between the input exposure values and the outcome (**Fig. 2A**). The positive relationship (0.023) observed supports the expectation that knockoffs should reflect the correlation between input exposures and the outcome. To compare it with an established statistical knockoff, we performed the same analysis with the model-X statistical knockoffs to find that the PRS values show a more conducive correlation structure between knockoff and outcome for sepsis incidence than the statistical knockoff values (-0.019). To further validate PRS, we performed the same comparison analysis on the sepsis mortality with 163 non-survivors from 3,181 survivors with sepsis. While 'model-X' knockoffs did show a positive correlation with the

outcome this time (0.18), the PRS knockoffs showed a stronger positive trend (0.27), demonstrating that PRS satisfies the first mathematical condition as a knockoff, often in an improved manner than the statistical knockoff. To ensure its generalizability beyond sepsis, we performed the same analysis on CHD incidence with 75,184 healthy controls and 7,710 UKBB subjects with CHD history and on CHD mortality with 122 non-survivors from 7,588 CHD cases (**S. Fig. 2A, B**). The PRS knockoffs demonstrate a strong positive correlation trend for both CHD incidence and mortality (0.002 and 0.098, respectively), similar to the statistical knockoff values (0.015 and 0.014, respectively), indicating comparable performance in these contexts.

The second condition posits that knockoff features should be independent of the outcome when conditioned on their corresponding input features. To assess this for PRS, we ran generalized covariance measure (GCM), a statistical test for conditional independence that evaluates the normalized empirical covariance between residuals from regression models. Running GCM on the PRS knockoffs with an input exposure and sepsis incidence outcome (**Fig. 2C**), we find that most exposures (80%, 34 out of 42) did not reject the null hypothesis of conditional independence (p-value > 0.05), indicating conditional independence and satisfying the second condition. In the same experiment with sepsis mortality outcome (**Fig. 2D**), even more exposures (95.2%, 40 out of 42) demonstrated the expected conditional independence, comparable to those obtained with statistical knockoffs (95.2% and 97.6%, respectively, **S. Table 2**). Further, the PRS knockoffs exhibit a consistent performance of conditional independence in CHD (**S. Fig. 2C, 2D**), supporting their generalizability.



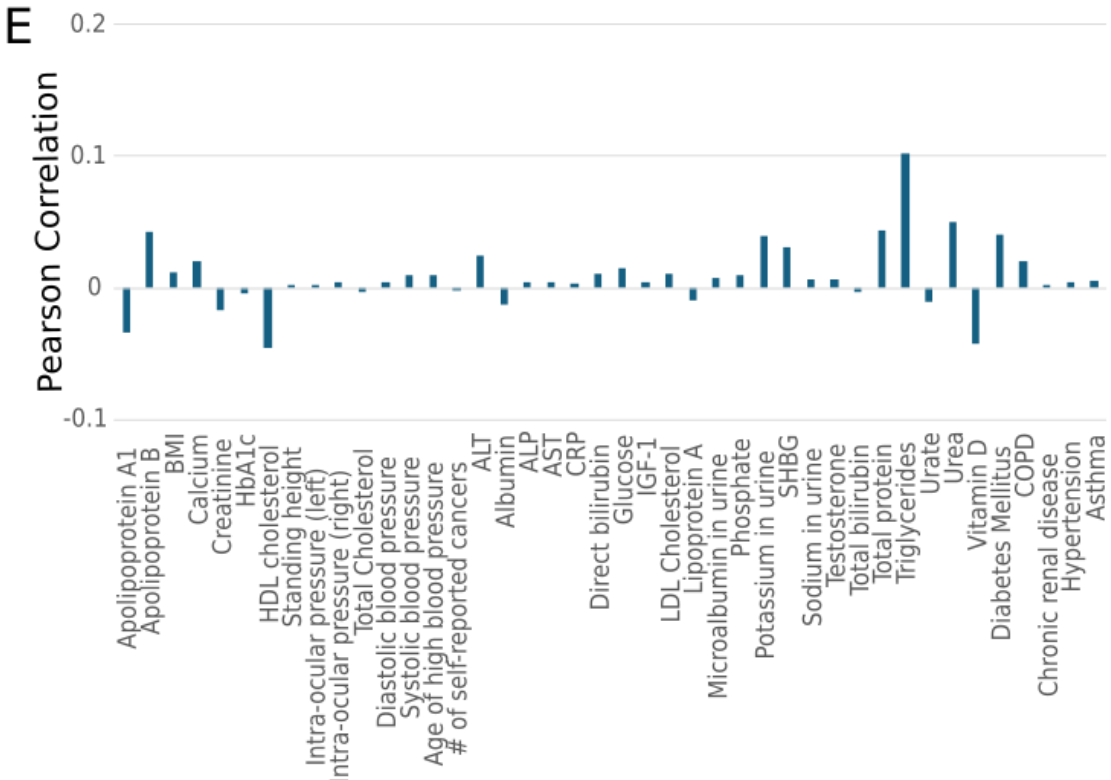


Figure 2 – PRS satisfies the mathematical properties to be knockoff. A) Scatterplot of correlation values for knockoff variables and sepsis incidence plotted against the correlation values for input exposures and sepsis incidence. **B)** Scatterplot of correlation values for knockoff variables and sepsis mortality plotted against the correlation values for input exposures and sepsis mortality. In A) and B), the blue points and regression line represent the data for PRS knockoffs, and the red points and regression line represent the data for statistical knockoffs. **C)** P-values for the GCM test for conditional independence for the sepsis incidence phenotype for PRS knockoffs (x-axis) and statistical knockoffs (y-axis). **D)** P-values for the GCM test for conditional independence for the sepsis mortality phenotype for PRS knockoffs (x-axis) and statistical knockoffs (y-axis). In C) and D), red dashed lines represent significance levels of 0.05. Points above the horizontal red dashed line exhibit conditional independence in the construct of PRS knockoffs, and points to the right of the vertical red dashed line exhibit conditional independence in the construct of statistical knockoffs. **E)** Pearson correlation coefficient values were calculated between the statistical and PRS knockoff variables for each of our 42 input exposures.

While PRS serves as a legitimate knockoff as statistical KO for complex diseases, PRS is also expected to capture distinct aspects of the exposure relationships with genetic predisposition, differing from what is represented by statKO. To explore this distinction, we calculated Pearson's correlation coefficient between the PRSKO and statKO across all the UKBB subjects we had available for each exposure (**Fig. 2E**). The generally low correlation (all below 0.12) indicates that PRSKO and statKO provide distinct and non-redundant information regarding exposure effects. Altogether, these results show that PRS is a valid type of knockoff with inherent biological context across sepsis and CHD, complex diseases with varying etiologies.

DeepEXPOKE accurately predicts outcomes by selecting important exposures.

To comprehensively evaluate the performance, we developed both DeepEXPOKE-statKO and DeepEXPOKE-PRSKO models using the 42 input exposures and applied each respective method to the UKBB sepsis incidence data with 5-fold cross validation. We compared these models to other established machine learning methods of different algorithms, XGBoost and Random Forest to compare sepsis case prediction performance. XGBoost (Extreme Gradient Boosting) is an ensemble learning method based on decision trees[33]. It builds multiple decision trees sequentially, where each tree corrects the errors of the previous one. While Random Forest is also an ensemble method[34], it consists of multiple decision trees constructed independently from different random subsets of the data. The final prediction is made by majority voting. We avoided developing machine learning models that are not straightforward to perform variable selection with, such as support vector classifier, as our interest is to further identify exposures significantly associated with the outcome as potential risk factors. DeepEXPOKE-PRSKO and DeepEXPOKE-statKO models significantly outperform the other machine learning methods (p-value= 3.4e-01, 9.9e-05, and 8.1e-02 from two-sample t tests when DeepEXPOKE-PRSKO is compared with DeepEXPOKE-statKO, XGBoost, and Random Forest, respectively). In terms of selected variables, all four machine learning methods identify unique sets of exposures with unique findings being most prevalent (**Fig. 3B**). Interestingly, the findings from DeepEXPOKE-PRSKO and DeepEXPOKE-statKO diverged substantially, similar to how they differed from the other machine learning methods. Given that both models share the same underlying deep neural network (DNN) architecture, with the only variation being the use of knockoff variables, this suggests that the DNN is both sensitive enough for robust predictions and flexible enough to effectively address various challenges in exposome data analysis, such

as genetic predispositions or collinearity. To further explore how the choice of knockoff affects model performance, we compared DeepEXPOKE-PRSKO and DeepEXPOKE-statKO across multiple performance metrics, including accuracy, sensitivity, and specificity. In all metrics, DeepEXPOKE-PRSKO consistently outperformed DeepEXPOKE-statKO (**Fig. 3C, D, E**), indicating that accounting for genetic predisposition is as crucial in identifying significant risk exposures for sepsis as accounting for exposure collinearity.

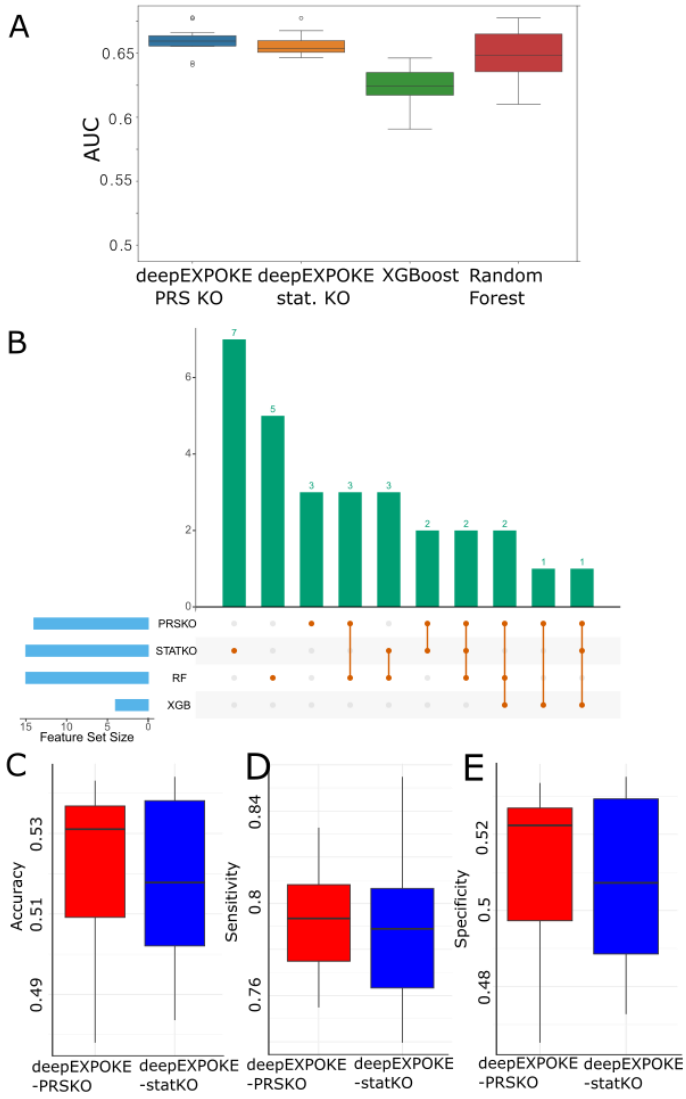


Figure 3: DeepEXPOKE accurately predicts outcomes by selecting important exposures. **A)** Boxplot showing the distribution and comparison of AUC values for models using DeepEXPOKE-PRS, DeepEXPOKE-STATKO, XGBoost, and Random Forest. Horizontal lines inside the bars denote median values. **B)** UpSet plot of identified selected features across DeepEXPOKE-PRS, DeepEXPOKE-STATKO, XGBoost, and Random Forest models. **C)** Boxplot comparison of accuracy between DeepEXPOKE-PRSKO (red) and DeepEXPOKE-STATKO (blue). **D)** Boxplot comparison of sensitivity between DeepEXPOKE-PRSKO (red) and DeepEXPOKE-STATKO (blue), **E)** Boxplot comparison of specificity between DeepEXPOKE-PRSKO (red) and DeepEXPOKE-STATKO (blue).

DeepEXPOKE identifies risk factors that affect outcomes at different biological levels.

To demonstrate how DeepEXPOKE-PRSKO and DeepEXPOKE-statKO differentiate risk exposures across biological layers, we estimated the effect size of the 14 and 15 exposures, respectively, that are strongly associated with sepsis incidence (**Fig. 4A, 4B**, see Methods). Both models identified five common exposures as risk factors (**Fig. 4C**), suggesting that these five shared exposures likely reflect the exposure risk independent of the genetic effect and collinearity (**Fig. 1C**). For example, glucose, triglycerides, and diabetes are known to exert the effect primarily at the metabolomic level[35-37]. Since metabolite levels and diabetes are highly responsive to environmental factors[38], the results support our findings as independent of genetic effect and collinearity.

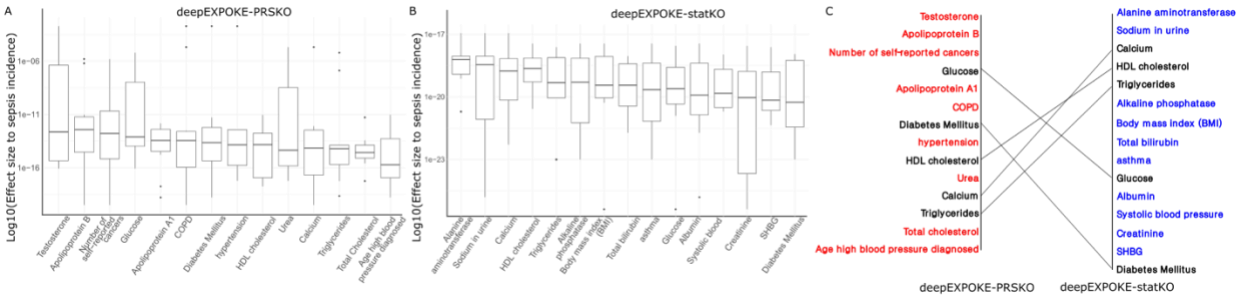


Figure 4: DeepEXPOKE identifies risk factors that affect outcomes at different biological levels. A) Boxplot showing the log-transformed effect sizes (W-statistics) of input risk factors on sepsis incidence using deepEXPOKE-PRSKO. Boxes are sorted by descending mean effect size value. **B)** Boxplot showing the log-transformed effect sizes (W-statistics) of input risk factors on sepsis incidence using deepEXPOKE-STATKO. Boxes are sorted by descending mean effect size value. **C)** Bump chart comparing significant risk factors between deepEXPOKE-PRSKO (red text) and deepEXPOKE-STATKO (blue text). Overlapping factors contributing to sepsis incidence in both models are noted in black text.

Separately, some exposures are identified only by DeepEXPOKE-statKO, such as asthma. Since risk factors only by DeepEXPOKE-statKO are likely driven mainly by genetic factors (**Fig. 1C**), this can partially explain the discrepancy where genome-wide association studies link asthma to increased sepsis risk[39] but patient data often shows a negative association[40], suggesting that asthma itself is not a direct sepsis risk factor, but the genetic variants associated with asthma may contribute to the risk. Similarly, the unique findings of DeepEXPOKE-PRSKO can be attributed to other factors than the genetic or exposome effects, such as collinearity. For example, blood pressure is found as a sepsis risk factor only by DeepEXPOKE-PRSKO. Despite many excellent antihypertensive drugs designed for controlling blood pressure, it remains difficult to achieve the desired target blood pressure; this phenomenon has been described as the hypertension paradox[31]. Our results suggest that this may be because hypertension is operated in association with other exposure associations, and it is not effective to control the internal biological system with the antihypertensive drugs unless the associations are controlled.

To assess generalizability, we estimated the effect sizes of 13 and 15 exposures significantly associated with CHD incidence by DeepEXPOKE-PRSKO and DeepEXPOKE-statKO, respectively (**S. Fig. 3A, B, C**) (see Methods). Consistently, exposures identified by both models, such as hypertension and total cholesterol, indicate that the risk is at the exposure level neither at the genetic level nor by collinearity. Hypertension and total cholesterol, are well-established topics for its genetic-independent contributions in

cardiovascular research, as listed in a recent review paper[41], support our finding. Overall, these findings demonstrate that DeepEXPOKE, using both statistical and PRS knockoffs, can effectively delineate the biological layers through which exposures influence disease phenotypes.

DeepEXPOKE reveals potential mechanisms of sepsis.

To demonstrate the clinical relevance of the DeepEXPOKE findings, we applied one-sample Mendelian Randomization using the 2-Stage Least Squares (2SLS) method[42] to the sepsis incidence data using the PRS scores generated by pgsc_calc as instrumental variables for each exposure (**Fig. 5A, B**). Comparing these results with those from DeepEXPOKE-PRSKO and DeepEXPOKE-statKO highlights several potential sepsis mechanisms. First, DeepEXPOKE-PRSKO and DeepEXPOKE-statKO identified several exposures related to lipoprotein and lipid metabolism, such as hypertension, Apolipoprotein B, and BMI, consistently with 2SLS, further supporting their involvement in sepsis. Second, DeepEXPOKE-PRSKO and DeepEXPOKE-statKO uncovered additional significant exposures related to metabolism not captured as highly prioritized features by 2SLS, such as testosterone, total cholesterol, and Apolipoprotein A1[43, 44], providing a more comprehensive view of how the metabolism pathways are related to sepsis. Third, DeepEXPOKE-PRSKO and DeepEXPOKE-statKO allow us to differentiate the exposure effects by the genetic, exposure, or collinearity-related influences. While both models emphasize the role of lipoprotein and lipid metabolism in sepsis risk, DeepEXPOKE-PRSKO uniquely identified Apolipoprotein B, among others. As DeepEXPOKE-PRSKO controls for genetic influences, this exposure likely impacts sepsis risk through collinearity with other exposures. Indeed, recent genetic studies suggest that lower low-density lipoprotein (LDL) levels, primarily Apolipoprotein B, tend to be more associative than causal for increased mortality risk in sepsis[45], further supporting and validating our findings. Fourth, by distinguishing between genetic and exposure influences, DeepEXPOKE can inform clinical trial designs for interventions targeting lipoprotein cholesterol, such as cholesteryl ester transfer protein (CETP) inhibitors or testosterone replacement therapy (TRT)[46] for sepsis management. For instance, if Apolipoprotein B is only associatively linked to sepsis risk, patient selection for treatment could focus on other lipid metabolism features with strong effect sizes identified by both DeepEXPOKE-PRSKO and DeepEXPOKE-statKO, such as urea, testosterone, total cholesterol, and Apolipoprotein A1.

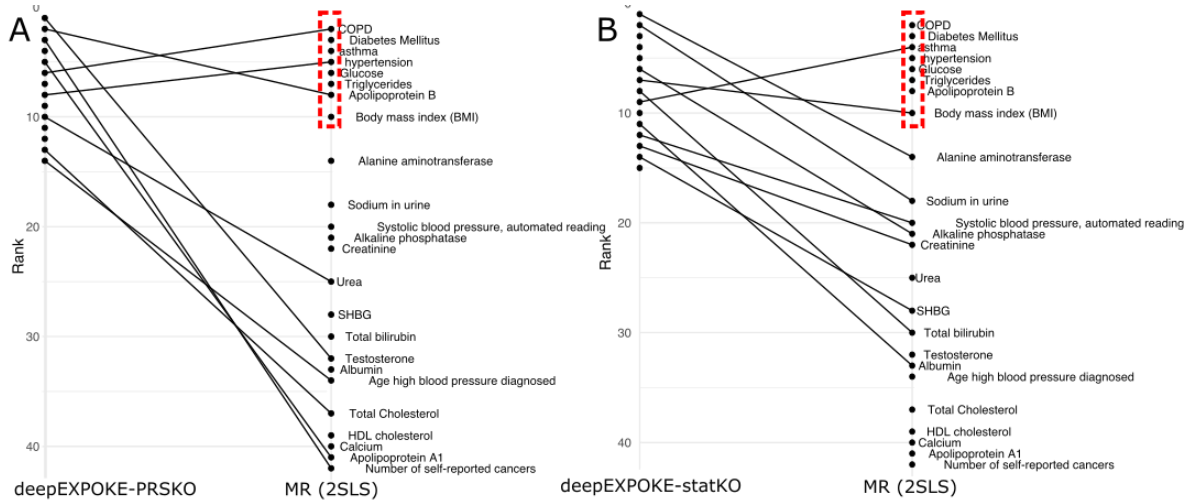


Figure 5. DeepEXPOKE reveals potential mechanisms of sepsis. A) Bump chart of selected risk factors between deepEXPOKE-PRSKO and MR with 2SLS. The red dashed box represents the eight highest

ranking risk factors identified via MR. **B)** Bump chart of selected risk factors between deepEXPOKE-STATKO and MR with 2SLS. The red dashed box represents the eight highest ranking risk factors identified via MR.

DISCUSSION

In this study, we developed DeepEXPOKE, a novel deep neural network (DNN) framework, to identify exposure risk factors for complex diseases by both employing statistical knockoff features and integrating polygenic risk score (PRS) as a novel type of biologically-driven knockoff. Our approach leverages the mathematical properties of PRS to validate it as a valid knockoff, thereby maintaining control over the genetic effects in exposome data analysis. Using these knockoffs, DeepEXPOKE successfully predicts disease outcomes by selecting and prioritizing important exposure variables that contribute to sepsis and coronary heart disease (CHD).

Further, DeepEXPOKE allows us to distinguish the genetic and exposome-level effects of the exposures, enabling a mechanistic understanding of sepsis and CHD. For sepsis, the exposome-level impacts of key metabolites, such as lipoproteins and glucose, were highlighted by the fact that both DeepEXPOKE-PRSKO and DeepEXPOKE-statKO identified them as risk factors. Also, deepEXPOKE helps explain the inconsistency of asthma genomics and patient data studies regarding its sepsis risk by estimating its risk only with statKO, but not with PRSKO. Since the risk estimated only by DeepEXPOKE-statKO indicates risk strictly from genetic factors, not from exposure collinearity nor exposure effect, the inconsistency confirms the identification of asthma only by our DeepEXPOKE-statKO method highlights how the separation of genetic and non-genetic effects helps elucidate the ongoing paradox in understanding the totality of effects[47, 48], such as those related to asthma in our analysis. Similarly, for CHD, DeepEXPOKE-statKO and DeepEXPOKE-PRSKO distinctively identify a strong non-genetic effect of both hypertension and total cholesterol, and a strong genetic effect related to testosterone, specifying the biological layer at which these effects occur. Altogether, DeepEXPOKE clearly demonstrates that genetic risk alone does not fully capture an individual's susceptibility to sepsis and CHD, aligning with findings that emphasize the multifactorial nature of sepsis and cardiovascular events[49]. This underscores the need for prevention and management strategies that account for both genetic and non-genetic risk factors.

DeepEXPOKE is also advantageous in its use of deep neural networks to capture nonlinear exposure effects. For example, DeepEXPOKE-PRSKO identifies testosterone as having a significant effect on sepsis risk while 2SLS, a linear model-based MR analysis, fails to highlight its importance. Testosterone is known to support immune functions, such as enhancing neutrophil differentiation and certain inflammatory responses, while also promoting an immunosuppressive phenotype that can inhibit bactericidal properties[28]. This contradictory behavior suggests a non-linear relationship where varying levels of testosterone may differentially influence immune cell behavior in sepsis.

Despite its advantages in risk factor identification and prediction capabilities, DeepEXPOKE is limited in a few aspects. First, while PRS can be useful, some data complexities exist in the use of PRS as instrumental variables, such as pleiotropy, population stratifications, and linkage disequilibrium (LD). Pleiotropy occurs when a single genetic variant influences multiple traits through different biological pathways. In the context of PRS, this means that the score may be associated with the outcome through pathways involving other than the exposure of interest, thus violating the exclusion restriction assumption of instrumental variables. Population stratification, which involves differences in allele

frequencies and trait distributions between subpopulations due to ancestry, can confound PRS-outcome associations if not properly controlled. Additionally, LD, the non-random association of alleles at different loci, can cause the PRS to overestimate the genetic component of the exposure when constructed from closely linked variants. Although we employed the PRS method developed to address ancestry structure and LD[32], unmeasured portion of the data complexities may still persist. Second, the model assumes that the causal direction flows from exposure to outcome, whether through genetic, non-genetic, or collinearity effects. However, in the case of collinearity, this assumption may not hold true. It is important to distinguish correlation due to causation from other forms of association, such as reverse causation or confounding. Reverse causality occurs when an outcome influences the exposure, rather than the exposure influencing the outcome. For instance, if disease severity (outcome) affects lifestyle behaviors (exposure), you would observe a correlation between lifestyle and disease, which might misleadingly suggest that lifestyle changes lead to changes in disease severity. Finally, bidirectional relationships can complicate causal relationship interpretation. For example, obesity and depression may influence each other: obesity can lead to depression, and depression can lead to obesity. This reciprocal relationship can be detected through the presence of correlation, but it complicates the interpretation of causality[50]. Thus, careful interpretation is warranted to understand the collinear effect of exposures.

The findings of this study have significant implications for understanding and managing complex diseases such as sepsis and CHD. By developing DeepEXPOKE, a deep neural network framework incorporating both the well-established statistical knockoff and a novel polygenic risk score (PRS) knockoff, we have demonstrated a powerful tool for identifying critical exposure risk factors that contribute to disease outcomes at the genetic, non-genetic, or collinearity level.

Methods

Calculating polygenic risk scores

PRS for each of our input exposures in our UK Biobank patient data were calculated using pgsc_calc software. Pgsc_calc enables streamlined calculation of PRS using a structured workflow system called Nextflow. The tool is a bioinformatics best-practice analysis pipeline for calculating polygenic risk scores on samples with imputed genotypes using existing scoring files from the Polygenic Score (PGS) Catalog. In the first step of the calculations, we used existing score files with comprehensive variant information from two UK Biobank PRS evaluation studies[51, 52], and then matched these score files to target genomes downloaded from the UK Biobank in VCF format. Genomic data and score files are in the GRCh37 specified genome build. Variants in the scoring files were matched against variants in the target genome data. Finally, we performed a scoring step that calculated PRS for each of our samples across our 42 features as a linear sum of weights and dosages. Pgsc_calc software leverages utilities from pgscatalog_utils and PLINK2 as part of its scoring pipeline, enabling us to compare PRS calculation differences between conventional methods (PLINK2) and the complete pgsc_calc pipeline.

Exposures for sepsis and CHD in the UKBB

We identified sepsis cases in the UK Biobank using The International Classification of Diseases, Tenth Revision (ICD-10) codes. The codes we included were A02.1, A22.7, A26.7, A32.7, A40, A40.0, A41.1, A41.2, A41.3, A41.4, A41.5, A41.6, A41.9, A42.7, B37.7, and O85, corresponding to septicemia subtypes ‘Salmonella Septicemia,’ ‘Anthrax Septicemia,’ ‘Erysipelothrix Septicemia,’ ‘Listerial Septicemia,’ ‘Streptococcal Septicemia,’ ‘Other Septicemia,’ ‘Unspecified Septicemia,’ ‘Actinomycotic Septicemia,’

'Candidal Septicemia,' and 'Puerperal Sepsis.' Sepsis deaths were confirmed using the corresponding ICD-10 codes from death cause records. For our control subjects, we selected a sample size of 79,791 individuals with no recorded hospital diagnoses of any sepsis or sepsis subtype.

42 unique features in the UK Biobank were selected as input variables for our PRS scoring framework and subsequent causal analyses. These variables fit into the following categories of measurements: routine measures of cardiac activity, respiratory condition, blood and urine biomarkers, anthropometry, lifestyle and environment, family history, complex disease, and health and medical history. All complex diseases were categorically coded with 0 representing "no disease" and 1 representing "diseased" states.

DeepEXPOKE Causal Association Estimation

To identify causal associations between the selected 42 features and the derived phenotypes, we built the DAG-deepVASE DNN architecture consisting of an input layer with twice the number of neurons as input features, accommodating both the original features and knockoff features in a pairwise manner. The network has two hidden layers, each with the same number of neurons as the input variables and uses the rectified linear unit (ReLU) activation function. The weights in the network are initialized using the Glorot normal initializer, and L1 (LASSO) regularization is applied to prevent overfitting. The model is optimized using the Adam algorithm with a learning rate of 0.001, and the mean squared error (MSE) serves as the loss function during training[23].

For identifying linear associations, we implemented Lee and Hastie's log-likelihood model to evaluate all possible pairs of variables. Lee and Hastie's log-likelihood model provides a framework for learning the structure of mixed graphical models that include both continuous and discrete variables[53]. It evaluates conditional dependencies between variables using Gaussian regression for continuous variables and multiclass logistic regression for discrete variables. By applying a log-likelihood approach with a sparsity penalty, this model allows for the identification of important variable associations and, by eliminating non-significant edges, results in a simpler interpretable model[23]. The sparsity penalty parameter was set to 0.3, allowing us to select the most important variables. The variable pairs that remained significant after applying the sparsity penalty were identified as linear associations.

To identify nonlinear associations between input variables x_i, DAG-deepVASE constructs a series of perceptron layers between $X_{\setminus i} = \{x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_M\}$ and x_i. This allows us to estimate the effect size of the association between $x_j \in X_{\setminus i}$ and x_i using model-X knockoffs. For the input variables x_i and x_j, the method ensures exchangeability $(x_i, x_j, \tilde{x}_j) \overset{d}{=} (x_i, \tilde{x}_j, x_j)$, where $\overset{d}{=}$ denotes equality in distribution. This property allows the identification of causal relationships, differentiating them from simple observed correlations. For instance, if (x_i, x_k) is a correlation without a causal link, then the feature exchangeability $(x_i, x_k, \tilde{x}_k) \overset{d}{=} (x_i, \tilde{x}_k, x_k)$ will hold, making the relationship measures $|R_{ik}|$ and $|\tilde{R}_{ik}|$ symmetric and exchangeable around zero where R_{ik} is the correlation between the original feature X_i and the outcome after training, and $\tilde{R}_{ik}$ is the correlation between the knockoff $\tilde{X}_i$ and the outcome after training. On the other hand, if (x_i, x_j) is a causal relationship, $S_{ij} = |R_{ik}| - |\tilde{R}_{ik}|$ will capture the deviation of this relationship from the null hypothesis. The knockoff matrix $\tilde{X}$ is designed to match the correlation structure within X while minimizing cross-correlation with the outcome variable Y[23].

The knockoff variables $\tilde{X} = (\tilde{x}_1, \dots, \tilde{x}_p)^T$ follow two main properties: exchangeability, where $(X, \tilde{X})_{\text{swap}(S)} \overset{d}{=} (X, \tilde{X})$, where swapping denotes exchanging any subset $S \subseteq \{1, \dots, M\}$ of the variables x_j, and independence, meaning $\tilde{X} \perp Y|X$, ensuring that $\tilde{X}$ is independent of X given the outcome Y.

To construct knockoffs, we assume $X \sim N(0, \Sigma)$, with $\Sigma \in \mathbb{R}^{M \times M}$ as the covariance matrix. A valid knockoff construction for $\tilde{X}$ is given by:

$$\tilde{X}|X \sim N(X - \text{diag}(S)\Sigma^{-1}X, 2\text{diag}(S) - \text{diag}(S)\Sigma^{-1}\text{diag}(S))$$

The effect size S_{ik} is computed as the difference in correlation strength between the original feature and the knockoff feature:

$$S_{ik} = |R_{ik}| - |\tilde{R}_{ik}|$$

If the original feature has a stronger correlation than its knockoff counterpart, it is called as a selected feature for continued analysis. Effect size calculations for each identified association involved pairing each input variable with its knockoff counterpart and using matrix operations across the DNN's layers to quantify the importance of each original variable relative to its knockoff. The effect size determining process utilizes weight matrices to compute feature importance scores. These importance scores are then used to calculate an effect size for each feature that reflects the strength of each feature's association with the outcome of interest, calculated as the difference between the importance of the original feature and its knockoff variable[23]. We parameterized *q* as a user-defined nominal false discovery rate (FDR), and subsequently set FDR to a level *q* = 0.05 to enable balance estimation of associations.

Comparison of DeepEXPOKE to Other Machine Learning Methods

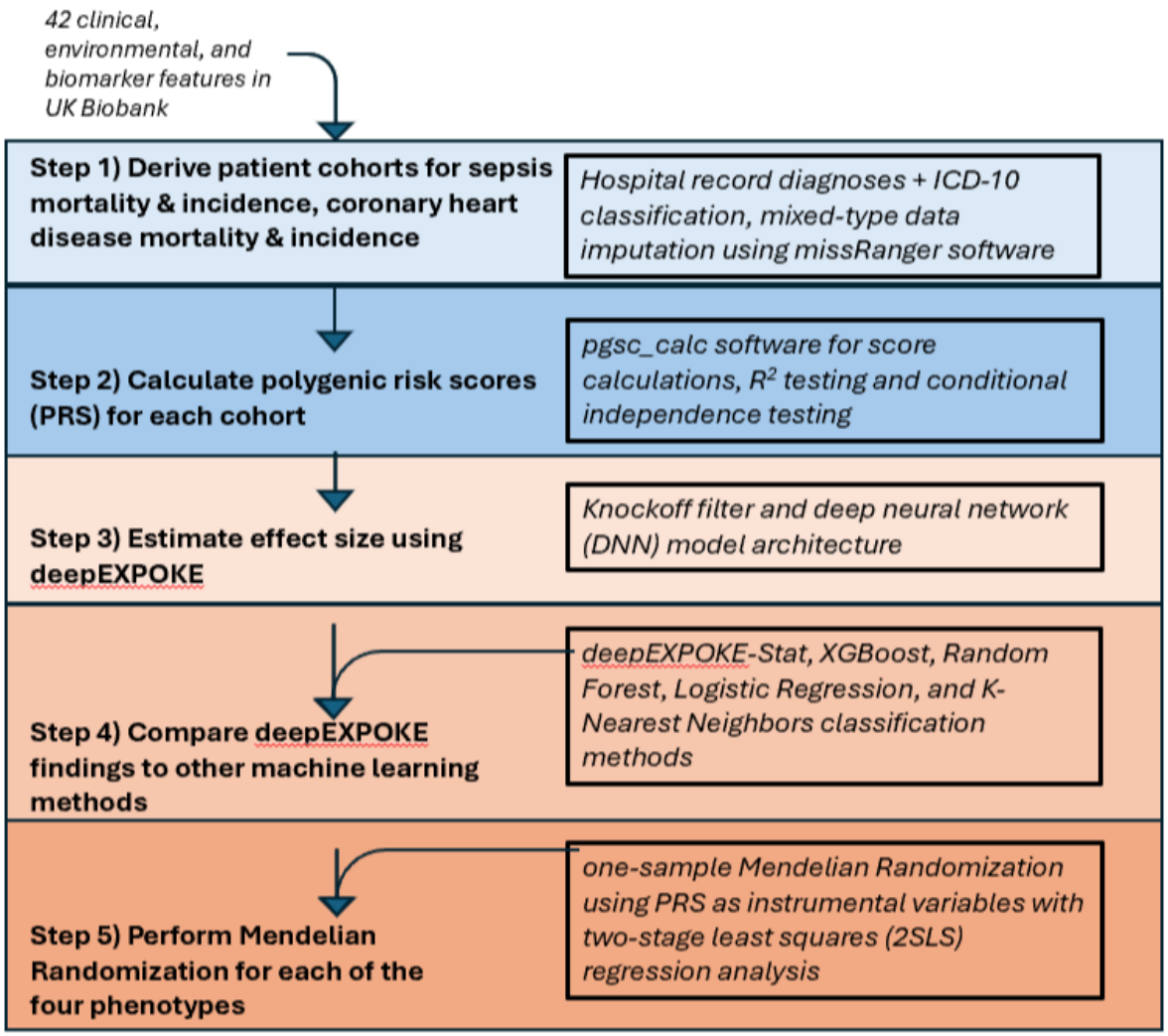
To evaluate DeepEXPOKE-PRS's performance in a comparative manner, we first tested it against DeepEXPOKE-statKO and compared accuracy, sensitivity, and specificity metrics between the two. A two-sample t-test was performed to compare the difference in mean values of each of the three metrics. Next, we compared model area under the receiver operating curve (AUC) values for DeepEXPOKE-PRS DeepEXPOKE-statKO, XGBoost, and Random Forest. XGBoost builds a series of decision trees sequentially, where each new tree recursively corrects errors made by previous trees [33]. Random Forest, rather than sequentially building trees, builds several independent decision trees and then aggregates their outputs to yield one resultant outcome[54]. XGBoost, Random Forest, and KNN parameter assignment and model setup were conducted using Python 3.8.5 and the *scikit-learn* and *xgboost* software packages[55].

For comparing the overlap of selected features across different machine learning methods, we used the *UpSetR* package to visualize the intersecting sets of features between DeepEXPOKE-PRS, DeepEXPOKE-Stat, XGBoost, Random Forest, and linear regression machine learning methods [56]. XGBoost's feature selection was performed automatically within the framework of the algorithm. Random Forest feature selection was done via a permutation feature importance threshold model. The mean value was the threshold for which selected features were identified. After the model was fit, features that had non-zero coefficients were extracted and identified as selected features.

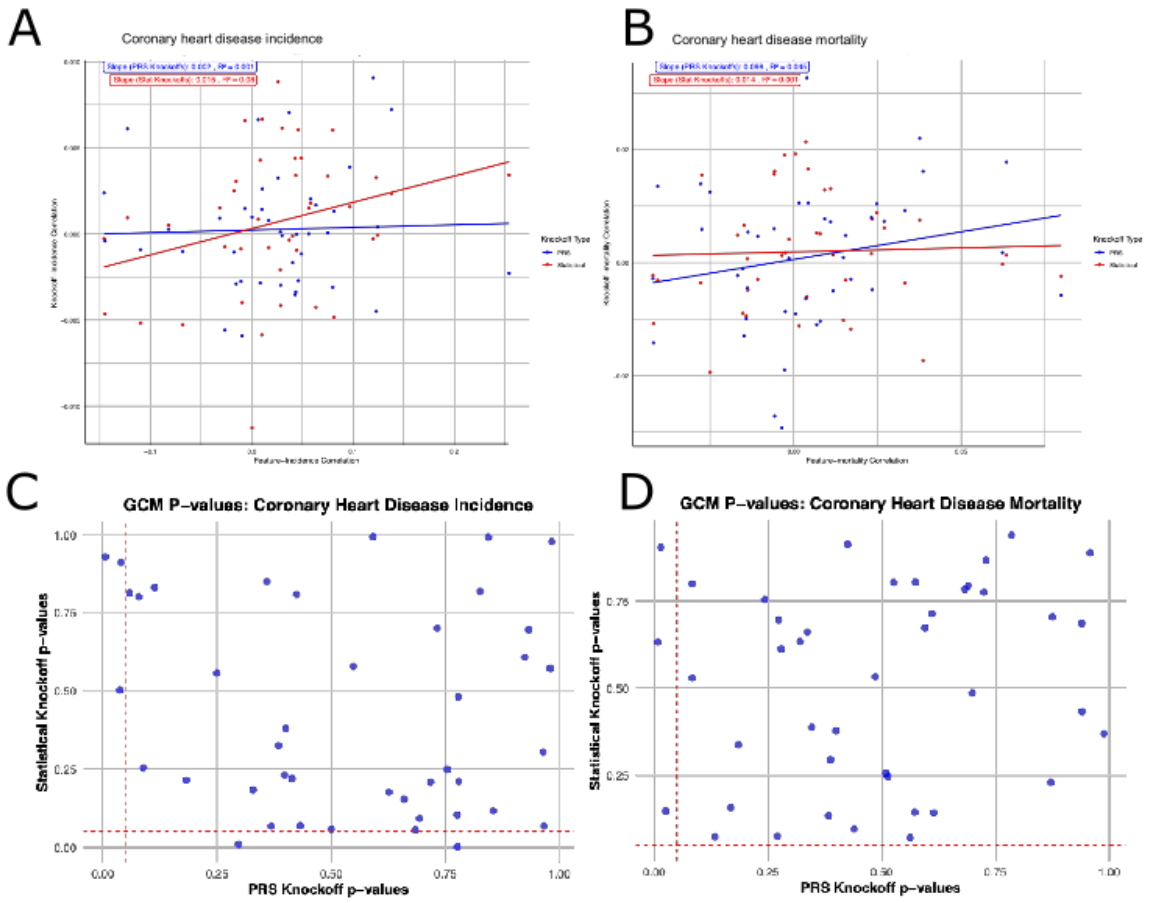
One-Sample Mendelian Randomization (MR) Analysis

For the primary Mendelian Randomization (MR) analysis, we used the calculated PRS generated from the pgsc_calc software as instrumental variables (IVs) to estimate relationships between our set of 42 exposures and the derived phenotypes of interest[32]. We first mathematically validated that PRS were appropriate instrumental variables to use as part of the MR step. The PRS generated by pgsc_calc for each feature serve as IVs for each exposure in our analysis, and PRS were used to capture the genetic predisposition to the exposures in a similar manner to the way single nucleotide polymorphisms (SNPs) are conventionally used in MR analysis[13].

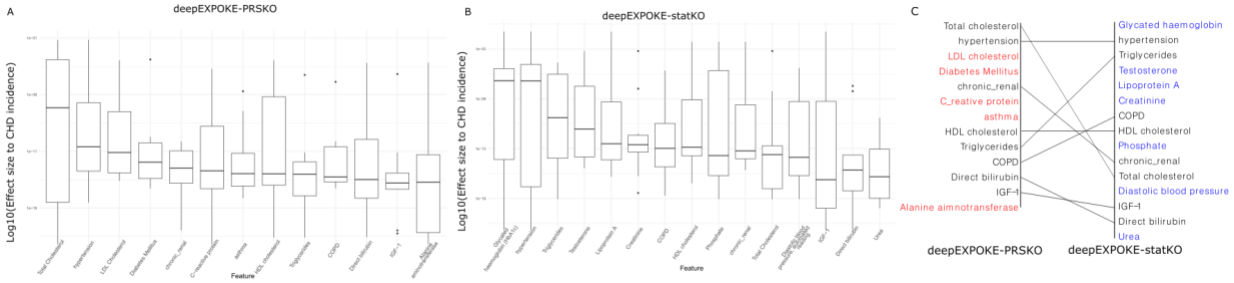
We used the two-stage least squares (2SLS) method for MR, a standard approach for one-sample MR experiments. The 2SLS approach involves two primary steps. First, the exposure (one of the 42 features) is regressed on the IV (feature PRS). This stage aims to estimate the genetically driven predicted values of the exposure. Second, the outcome is then regressed on the predicted values of the exposure calculated from the first stage. The causal effect of the exposure on the outcome is given by the coefficient of the genetically predicted exposure from the stage-two regression. We implemented the 2SLS approach using *ivreg* from the AER package in R. For each exposure, we fit separate models, using the exposure's corresponding PRS as the IV. Robust standard errors were used to account for any heteroscedasticity. All statistical analyses were conducted using R version 4.2.1.



S. Figure 1. Overview of experimental design: cohort establishment, polygenic risk score calculations, DeepEXPOKE model setup, performance evaluation and comparison to other machine learning methods, and Mendelian Randomization setup.



S. Figure 2. A) Scatterplot of correlation values for knockoff variables and CHD incidence plotted against the correlation values for input exposures and CHD incidence. **B)** Scatterplot of correlation values for knockoff variables and CHD mortality plotted against the correlation values for input exposures and CHD mortality. In A) and B), the blue points and regression line represent the data for PRS knockoffs, and the red points and regression line represent the data for statistical knockoffs. **C)** P-values for the GCM test for conditional independence for the sepsis incidence phenotype for PRS knockoffs (x-axis) and statistical knockoffs (y-axis). **D)** P-values for the GCM test for conditional independence for the sepsis mortality phenotype for PRS knockoffs (x-axis) and statistical knockoffs (y-axis). In C) and D), red dashed lines represent significance levels of 0.05. Points above the horizontal red dashed line exhibit conditional independence in the construct of PRS knockoffs, and points to the right of the vertical red dashed line exhibit conditional independence in the construct of statistical knockoffs.



S. Figure 3. A) Boxplot showing the log-transformed effect sizes (W-statistics) of input risk factors on sepsis incidence using deepEXPOKE-PRSKO. Boxes are sorted by descending mean effect size value. **B)** Boxplot showing the log-transformed effect sizes (W-statistics) of input risk factors on sepsis incidence using deepEXPOKE-STATKO. Boxes are sorted by descending mean effect size value. **C)** Bump chart comparing significant risk factors between deepEXPOKE-PRSKO (red text) and deepEXPOKE-STATKO (blue text). Overlapping factors contributing to sepsis incidence in both models are noted in black text.

References

1. Wild CP. Complementing the Genome with an "Exposome": The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology. Cancer Epidemiology, Biomarkers & Prevention. 2005;14(8):1847-50. doi: 10.1158/1055-9965.Epi-05-0456.
2. Wild CP. The exposome: from concept to utility. International Journal of Epidemiology. 2012;41(1):24-32. doi: 10.1093/ije/dyr236.
3. Vermeulen R, Schymanski EL, Barabási A-L, Miller GW. The exposome and health: Where chemistry meets biology. Science. 2020;367(6476):392-6. doi: doi:10.1126/science.aay3164.
4. Jiang C, Wang X, Li X, Inlora J, Wang T, Liu Q, Snyder M. Dynamic Human Environmental Exposome Revealed by Longitudinal Personal Monitoring. Cell. 2018;175(1):277-91.e31. doi: 10.1016/j.cell.2018.08.060.
5. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, Cortes A, Welsh S, Young A, Effingham M, McVean G, Leslie S, Allen N, Donnelly P, Marchini J. The UK Biobank resource with deep phenotyping and genomic data. Nature2018. p. 203-9.
6. Burkett JP, Miller GW. Using the exposome to understand environmental contributors to psychiatric disorders. Neuropsychopharmacology. 2021;46(1):263-4. doi: 10.1038/s41386-020-00851-0.
7. Lin BD, Pries L-K, Sarac HS, van Os J, Rutten BPF, Luykx J, Guloksuz S. Nongenetic Factors Associated With Psychotic Experiences Among UK Biobank Participants: Exposome-Wide Analysis and Mendelian Randomization Analysis. JAMA Psychiatry. 2022;79(9):857-68. doi: 10.1001/jamapsychiatry.2022.1655.
8. Liu F, Xu J, Guo L, Qin W, Liang M, Schumann G, Yu C. Environmental neuroscience linking exposome to brain structure and function underlying cognition and behavior. Molecular Psychiatry. 2023;28(1):17-27. doi: 10.1038/s41380-022-01669-6.
9. Zhou X, Lee SH. An integrative analysis of genomic and exposomic data for complex traits and phenotypic prediction. Scientific Reports. 2021;11(1):21495. doi: 10.1038/s41598-021-00427-y.
10. Di Scipio M, Khan M, Mao S, Chong M, Judge C, Pathan N, Perrot N, Nelson W, Lali R, Di S, Morton R, Petch J, Paré G. A versatile, fast and unbiased method for estimation of gene-by-environment interaction effects on biobank-scale datasets. Nature Communications. 2023;14(1):5196. doi: 10.1038/s41467-023-40913-7.
11. Khera AV, Emdin CA, Drake I, Natarajan P, Bick AG, Cook NR, Chasman DI, Baber U, Mehran R, Rader DJ, Fuster V, Boerwinkle E, Melander O, Orho-Melander M, Ridker PM, Kathiresan S. Genetic Risk, Adherence to a Healthy Lifestyle, and Coronary Disease. New England Journal of Medicine. 2016;375(24):2349-58. doi: doi:10.1056/NEJMoa1605086.
12. Ding E, Wang Y, Liu J, Tang S, Shi X. A review on the application of the exposome paradigm to unveil the environmental determinants of age-related diseases. Human Genomics. 2022;16(1):54. doi: 10.1186/s40246-022-00428-6.
13. Bond TA, Richmond RC, Karhunen V, Cuellar-Partida G, Borges MC, Zuber V, Couto Alves A, Mason D, Yang TC, Gunter MJ, Dehghan A, Tzoulaki I, Sebert S, Evans DM, Lewin AM, O'Reilly PF, Lawlor DA, Järvelin M-R. Exploring the causal effect of maternal pregnancy adiposity on offspring adiposity: Mendelian randomisation using polygenic risk scores. BMC Medicine. 2022;20(1):34. doi: 10.1186/s12916-021-02216-w.
14. Chung M-K, House J, Akhtari F, Makris K, Langston M, Islam T, Holmes P, Chadeau-Hyam M, Smirnov A, Du X, Thessen A, Cui Y, Zhang K, Manrai A, Motsinger-Reif A, Patel C, Bisson W. Decoding the exposome: data science methodologies and implications in exposome-wide association studies (ExWASs). Exposome. 2024;4.
15. Moore TM, Visoki E, Argabright ST, Didomenico GE, Sotelo I, Wortzel JD, Naeem A, Gur RC, Gur RE, Warrier V, Guloksuz S, Barzilay R. Modeling environment through a general exposome factor in two

independent adolescent cohorts. Exposome. 2022;2(1):osac010. Epub 2023/01/07. doi: 10.1093/exposome/osac010. PubMed PMID: 36606125; PMCID: PMC9798749.
16. Adami G, Pontalti M, Cattani G, Rossini M, Viapiana O, Orsolini G, Benini C, Bertoldo E, Fracassi E, Gatti D, Fassio A. Association between long-term exposure to air pollution and immune-mediated diseases: a population-based cohort study. RMD Open. 2022;8(1). Epub 2022/03/17. doi: 10.1136/rmdopen-2021-002055. PubMed PMID: 35292563; PMCID: PMC8969049.
17. Reilly JP, Zhao Z, Shashaty MGS, Koyama T, Jones TK, Anderson BJ, Ittner CA, Dunn T, Miano TA, Oniyide O, Balmes JR, Matthay MA, Calfee CS, Christie JD, Meyer NJ, Ware LB. Exposure to ambient air pollutants and acute respiratory distress syndrome risk in sepsis. Intensive Care Med. 2023;49(8):957-65. Epub 2023/07/20. doi: 10.1007/s00134-023-07148-y. PubMed PMID: 37470831; PMCID: PMC10561716.
18. Naidoo S, Zwane AM, Paruk A, Hardcastle TC. Diagnosis and Management of Severe Water-Related Skin and Soft Tissue Sepsis: A Summative Review of the Literature. Diagnostics (Basel). 2023;13(13). Epub 2023/07/14. doi: 10.3390/diagnostics13132150. PubMed PMID: 37443543; PMCID: PMC10340249.
19. Stensrud VH, Gustad LT, Damås JK, Solligård E, Krokstad S, Nilsen TIL. Direct and indirect effects of socioeconomic status on sepsis risk and mortality: a mediation analysis of the HUNT Study. J Epidemiol Community Health. 2023;77(3):168-74. Epub 2023/01/28. doi: 10.1136/jech-2022-219825. PubMed PMID: 36707239.
20. Rudd KE, Kissoon N, Limmathurotsakul D, Bory S, Mutahunga B, Seymour CW, Angus DC, West TE. The global burden of sepsis: barriers and potential solutions. Crit Care. 2018;22(1):232. Epub 2018/09/24. doi: 10.1186/s13054-018-2157-z. PubMed PMID: 30243300; PMCID: PMC6151187.
21. Bojarczuk A, Egorova ES, Dzitkowska-Zabielska M, Ahmetov, II. Genetics of Exercise and Diet-Induced Fat Loss Efficiency: A Systematic Review. J Sports Sci Med. 2024;23(1):236-57. Epub 20240301. doi: 10.52082/jssm.2024.236. PubMed PMID: 38455434; PMCID: PMC10915602.
22. Malina S, Cizin D, Knowles DA. Deep mendelian randomization: Investigating the causal knowledge of genomic deep learning models. PLOS Computational Biology. 2022;18(10):e1009880. doi: 10.1371/journal.pcbi.1009880.
23. Fan Z, Kernan KF, Sriram A, Benos PV, Canna SW, Carcillo JA, Kim S, Park HJ. Deep neural networks with knockoff features identify nonlinear causal relations and estimate effect sizes in complex biological systems. GigaScience. 2023;12. doi: 10.1093/gigascience/giad044.
24. Candès E, Fan Y, Janson L, Lv J. Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. Journal of the Royal Statistical Society Series B: Statistical Methodology2018. p. 551-77.
25. Lu YY, Fan Y, Lv J, Noble WS. Deeppink: Reproducible feature selection in deep neural networks. Advances in Neural Information Processing Systems2018. p. 8676-86.
26. Cornish AJ, Tomlinson IPM, Houlston RS. Mendelian randomisation: A powerful and inexpensive method for identifying and excluding non-genetic risk factors for colorectal cancer. Mol Aspects Med. 2019;69:41-7. Epub 2019/02/03. doi: 10.1016/j.mam.2019.01.002. PubMed PMID: 30710596; PMCID: PMC6856712.
27. Rudd KE, Johnson SC, Agesa KM, Shackelford KA, Tsoi D, Kievlan DR, Colombara DV, Ikuta KS, Kissoon N, Finfer S, Fleischmann-Struzek C, Machado FR, Reinhart KK, Rowan K, Seymour CW, Watson RS, West TE, Marinho F, Hay SI, Lozano R, Lopez AD, Angus DC, Murray CJL, Naghavi M. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. The Lancet: The Author(s). Published by Elsevier Ltd. This is an Open Access Article under the CC BY 4.0 licence; 2020. p. 200-11.
28. Ben-Batalla I, Vargas-Delgado ME, von Amsberg G, Janning M, Loges S. Influence of Androgens on Immunity to Self and Foreign: Effects on Immunity and Cancer. Front Immunol. 2020;11:1184. Epub 20200702. doi: 10.3389/fimmu.2020.01184. PubMed PMID: 32714315; PMCID: PMC7346249.

29. Oz HS. Nutrients, Infectious and Inflammatory Diseases. Nutrients. 2017;9(10). Epub 20170930. doi: 10.3390/nu9101085. PubMed PMID: 28973995; PMCID: PMC5691702.
30. Bhatnagar A. Environmental Determinants of Cardiovascular Disease. Circ Res. 2017;121(2):162-80. doi: 10.1161/circresaha.117.306458. PubMed PMID: 28684622; PMCID: PMC5777598.
31. Chobanian AV. The Hypertension Paradox — More Uncontrolled Disease despite Improved Therapy. New England Journal of Medicine. 2009;361(9):878-87. doi: doi:10.1056/NEJMsa0903829.
32. Lambert SA, Wingfield B, Gibson JT, Gil L, Ramachandran S, Yvon F, Saverimuttu S, Tinsley E, Lewis E, Ritchie SC, Wu J, Canovas R, McMahon A, Harris LW, Parkinson H, Inouye M. The Polygenic Score Catalog: new functionality and tools to enable FAIR research. medRxiv. 2024:2024.05.29.24307783. doi: 10.1101/2024.05.29.24307783.
33. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; San Francisco, California, USA: Association for Computing Machinery; 2016. p. 785–94.
34. Breiman LEO. Random Forests. Machine Learning2001. p. 5-32.
35. Sachwani GR, Jaehne AK, Jayaprakash N, Kuzich M, Onkoba V, Blyden D, Rivers EP. The association between blood glucose levels and matrix-metalloproteinase-9 in early severe sepsis and septic shock. Journal of Inflammation. 2016;13(1):13. doi: 10.1186/s12950-016-0122-7.
36. Lu Z, Tao G, Sun X, Zhang Y, Jiang M, Liu Y, Ling M, Zhang J, Xiao W, Hua T, Zhu H, Yang M. Association of Blood Glucose Level and Glycemic Variability With Mortality in Sepsis Patients During ICU Hospitalization. Frontiers in Public Health. 2022;10. doi: 10.3389/fpubh.2022.857368.
37. Rivas AM, Nugent K. Hyperglycemia, Insulin, and Insulin Resistance in Sepsis. The American Journal of the Medical Sciences. 2021;361(3):297-302. doi: https://doi.org/10.1016/j.amjms.2020.11.007.
38. Rattray NJW, Deziel NC, Wallach JD, Khan SA, Vasiliou V, Ioannidis JPA, Johnson CH. Beyond genomics: understanding exposotypes through metabolomics. Hum Genomics. 2018;12(1):4. Epub 20180126. doi: 10.1186/s40246-018-0134-x. PubMed PMID: 29373992; PMCID: PMC5787293.
39. Luo J, Liu P, Luo Y. Genetic prediction of asthma increases multiple sepsis risks: A Mendelian randomization study. World Allergy Organization Journal. 2024;17(8):100937. doi: https://doi.org/10.1016/j.waojou.2024.100937.
40. Zein JG, Love TE, Erzurum SC. Asthma Is Associated with a Lower Risk of Sepsis and Sepsis-related Mortality. Am J Respir Crit Care Med. 2017;196(6):787-90. Epub 2017/05/23. doi: 10.1164/rccm.201608-1583LE. PubMed PMID: 28530491; PMCID: PMC5620674.
41. Enkhmaa B, Berglund L. Non-genetic influences on lipoprotein(a) concentrations. Atherosclerosis. 2022;349:53-62. doi: 10.1016/j.atherosclerosis.2022.04.006. PubMed PMID: 35606076; PMCID: PMC9549811.
42. Windmeijer F. Two-stage least squares as minimum distance. The Econometrics Journal. 2019;22(1):1-9. doi: 10.1111/ectj.12115.
43. Liu G, Jiang L, Kerchberger VE, Oeser A, Ihegword A, Dickson AL, Daniel LL, Shaffer C, Linton MF, Cox N, Chung CP, Wei WQ, Stein CM, Feng Q. The relationship between high density lipoprotein cholesterol and sepsis: A clinical and genetic approach. Clin Transl Sci. 2023;16(3):489-501. Epub 20230116. doi: 10.1111/cts.13462. PubMed PMID: 36645160; PMCID: PMC10014701.
44. De Geest B, Mishra M. Impact of High-Density Lipoproteins on Sepsis. Int J Mol Sci. 2022;23(21). Epub 20221026. doi: 10.3390/ijms232112965. PubMed PMID: 36361756; PMCID: PMC9655469.
45. Walley KR, Boyd JH, Kong HJ, Russell JA. Low Low-Density Lipoprotein Levels Are Associated With, But Do Not Causally Contribute to, Increased Mortality in Sepsis. Crit Care Med. 2019;47(3):463-6. doi: 10.1097/ccm.0000000000003551. PubMed PMID: 30394916.

46. Corona G, Monami M, Rastrelli G, Aversa A, Tishova Y, Saad F, Lenzi A, Forti G, Mannucci E, Maggi M. Testosterone and metabolic syndrome: a meta-analysis study. J Sex Med. 2011;8(1):272-83. Epub 20100830. doi: 10.1111/j.1743-6109.2010.01991.x. PubMed PMID: 20807333.
47. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, Bellomo R, Bernard GR, Chiche J-D, Coopersmith CM, Hotchkiss RS, Levy MM, Marshall JC, Martin GS, Opal SM, Rubenfeld GD, van der Poll T, Vincent J-L, Angus DC. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016;315(8):801-10. doi: 10.1001/jama.2016.0287.
48. Gotts JE, Matthay MA. Sepsis: pathophysiology and clinical management. Bmj. 2016;353:i1585. Epub 20160523. doi: 10.1136/bmj.i1585. PubMed PMID: 27217054.
49. Yusuf S, Hawken S, Ounpuu S, Dans T, Avezum A, Lanas F, McQueen M, Budaj A, Pais P, Varigos J, Lisheng L. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. Lancet. 2004;364(9438):937-52. doi: 10.1016/s0140-6736(04)17018-9. PubMed PMID: 15364185.
50. Milano W, Ambrosio P, Carizzone F, De Biasio V, Di Munzio W, Foia MG, Capasso A. Depression and Obesity: Analysis of Common Biomarkers. Diseases. 2020;8(2):23. PubMed PMID: doi:10.3390/diseases8020023.
51. Tanigawa Y, Qian J, Venkataraman G, Justesen JM, Li R, Tibshirani R, Hastie T, Rivas MA. Significant sparse polygenic risk scores across 813 traits in UK Biobank. PLOS Genetics. 2022;18(3):e1010105. doi: 10.1371/journal.pgen.1010105.
52. Sinnott-Armstrong N, Tanigawa Y, Amar D, Mars N, Benner C, Aguirre M, Venkataraman GR, Wainberg M, Ollila HM, Kiiskinen T, Havulinna AS, Pirruccello JP, Qian J, Shcherbina A, Rodriguez F, Assimes TL, Agarwala V, Tibshirani R, Hastie T, Ripatti S, Pritchard JK, Daly MJ, Rivas MA, FinnGen. Genetics of 35 blood and urine biomarkers in the UK Biobank. Nature Genetics. 2021;53(2):185-94. doi: 10.1038/s41588-020-00757-z.
53. Lee JD, Hastie TJ. Learning the Structure of Mixed Graphical Models. J Comput Graph Stat. 2015;24(1):230-53. doi: 10.1080/10618600.2014.900500. PubMed PMID: 26085782; PMCID: PMC4465824.
54. Touw WG, Bayjanov JR, Overmars L, Backus L, Boekhorst J, Wels M, van Hijum SA. Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? Brief Bioinform. 2013;14(3):315-26. Epub 20120710. doi: 10.1093/bib/bbs034. PubMed PMID: 22786785; PMCID: PMC3659301.
55. Taunk K, De S, Verma S, Swetapadma A. A Brief Review of Nearest Neighbor Algorithm for Learning and Classification. 2019 International Conference on Intelligent Computing and Control Systems (ICCS). 2019:1255-60.
56. Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics. 2017;33(18):2938-40. doi: 10.1093/bioinformatics/btx364.