

Article

Machine Learning-Directed Discovery and Statistical Validation of Post-COVID-19 Condition Sequelae Using Military Health System Data

Jed Shakarji ^{1,2,*} , Apryl Susi ^{1,2,*} , Zella Berill ^{1,2}, Remle Scott ^{1,2} , Dominic Nathan ^{3,4} 
and Cade M. Nylund ^{1,5} 

- ¹ Department of Pediatrics, Uniformed Services University, Bethesda, MD 20814, USA
 - ² The Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc., Bethesda, MD 20817, USA
 - ³ Department of Preventive Medicine and Biostatistics, Uniformed Services University, Bethesda, MD 20814, USA; dominic.nathan@usuhs.edu
 - ⁴ National Institutes of Neurological Diseases and Stroke, National Institutes of Health, Bethesda, MD 20824, USA
 - ⁵ Department of Preventative and Restorative Care, Brigham Young University School of Medicine, Provo, UT 84602, USA
- * Correspondence: jed.shakarji.ctr@usuhs.edu (J.S.); apryl.susi.ctr@usuhs.edu (A.S.)

Abstract

Background: Post-COVID-19 conditions (PCCs) present a significant public health challenge due to a vast array of new or persistent health symptoms across subjects. The complex, multi-systemic nature of PCCs makes these conditions difficult to differentiate from other non-COVID-19 related medical conditions. While the Military Health System Data Repository (MDR) provides a robust supply of population-level encounter data, its high-dimensional structure poses challenges for knowledge discovery and outcome research. **Objectives:** The primary aim of this study was to identify novel manifestations of PCCs among active-duty service members, and model the probabilistic relationships between PCC-related diagnoses. We propose a machine learning workflow as an effective tool for knowledge discovery to statistically validate candidate PCCs from large datasets. **Methods:** We conducted a retrospective cohort study using MDR records from July 2018 to June 2023. From an initial pool of 311,367 eligible Active-Duty Tricare beneficiaries, we isolated 101,789 COVID-19 infections and matched them 1:1 with uninfected controls (N = 203,578 total) based on age, sex, and propensity for COVID-19. Encounter data was mapped to 392 clinical categories using the Healthcare Cost and Utilization Project (HCUP) Clinical Classification Software Refined (CCSR). Candidate PCC categories were isolated using a cross-validated lasso regression model optimized with a Tree of Parzen Estimators algorithm. A consensus Bayesian Network structure was fitted to model potential probabilistic dependency structures between identified PCCs and prior COVID-19 diagnosis. Finally, conditional Cox proportional hazards models were used to statistically validate selected novel conditions using larger cohorts drawn from the same initial eligible pool by matching cases 1:2 with controls. **Results:** Feature selection reduced the diagnosis set by 97.96%, isolating 8 clinical categories from the initial 392. The model confirmed known PCCs, such as respiratory symptoms and malaise, and identified two potentially novel candidate PCCs: tinnitus and personality disorders. Survival analysis validated the selection of tinnitus, showing a significant association with COVID-19 (HR: 1.17, 95% CI: 1.12–1.22). No significant association was found between COVID-19 infection and personality disorders (HR: 1.11, 95% CI: 0.97–1.26). **Conclusions:** This study demonstrates an effective analytical pathway for addressing the limitations of analyzing complex, high-dimensional healthcare billing data. The methodology successfully generated testable hypotheses, identifying



Received: 6 April 2026

Revised: 20 June 2026

Accepted: 24 June 2026

Published: 30 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

tinnitus as a relevant sequela, and is generalizable to future research involving unknown health outcomes related to prior infection.

Keywords: COVID-19; post-COVID-19 conditions; machine learning; feature selection; lasso regression; statistical validation; Bayesian Networks; survival analysis

1. Introduction

The clinical understanding of what is now formally classified as post-COVID-19 conditions (PCCs) has undergone a significant paradigm shift since the early stages of the pandemic [1–4]. Initially, the term “Long COVID” was coined in the spring of 2020 by the patients experiencing it to describe a condition where they failed to recover for several weeks or months following the onset of their illness [2,5,6]. In this early phase, the condition was largely viewed through a negative definition, simply a lack of full recovery from the initial COVID-19 infection [2,5–7]. This protracted recovery period was primarily characterized by ongoing symptoms such as breathlessness, cough, and fatigue, which were very common among previously hospitalized patients [3,4,7,8].

However, as the COVID-19 pandemic progressed and longitudinal patient data accumulated, a more complex understanding of PCCs emerged. More delayed-onset symptoms and medical conditions began to be described, often appearing weeks or months after the initial viral clearance in patients who had otherwise appeared to recover [2–4,7–9]. These sequelae often involve systems entirely distinct from the primary respiratory site, spanning neurological [2,7,9–11], cardiovascular [3,4,7,8], gastrointestinal and other organ systems [2–4,7,8]. PCCs are complex and multi-systemic, making them difficult to differentiate from other non-COVID-19-related medical conditions, such as myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) [9,11]. In the wake of the COVID-19 pandemic, there is a pressing need to develop a better understanding of the virus’s long-lasting side effects, particularly given its staggering global scale [2–4,7,8].

With COVID-19 serving as a model for future epidemiological research, there is a critical need for efficient analytical frameworks to identify candidate associated conditions. While large healthcare billing datasets offer abundant longitudinal data on clinical encounters, symptoms, and diagnoses, they inherently introduce significant challenges regarding sparsity and high dimensionality. Traditional statistical methods remain foundational to epidemiology, particularly for rigorous hypothesis testing; however, they are often challenged by the sheer scale of modern repositories like the Military Health System Data Repository (MDR). Furthermore, while hypothesis testing is a highly formalized scientific activity, hypothesis generation has historically remained a largely informal process. Relying solely on conventional epidemiological techniques to manually sift through thousands of potential diagnosis codes is time- and resource-inefficient and highly vulnerable to missing subtle, unhypothesized clinical signals. Machine learning offers a complementary approach that is less constrained by prior assumptions, assisting in the modeling of complex relationships to uncover hidden patterns. Applying machine learning algorithms to high-dimensional data allows researchers to systematically generate novel, interpretable hypotheses that might otherwise go unnoticed.

Using the large MDR, this study proposes a machine learning-directed discovery workflow which moves from dimensionality reduction to interpretable candidate discovery, followed by traditional epidemiological confirmation. We propose this sequential, integrated approach to serve as a template for knowledge discovery, outcome research, and

investigating novel and unknown outcomes following any significant health exposure, with an emphasis on efficiency for large and disparate datasets.

2. Methods

We conducted a retrospective cohort study utilizing data from the MDR from July 2018 to June 2023. The timeline began in July 2018 to allow for the enforcement of a rigorous baseline washout period for the earliest waves of pandemic infections, enabling the complete capture of pre-existing chronic conditions and baseline healthcare utilization trends prior to the emergence of COVID-19. To be included, individuals were required to be active-duty service members with at least 12 months of continuous enrollment and at least one encounter or prescription in the prior year. COVID-19 infection was determined during the months of July 2020 to June 2021 using either ICD-10 diagnosis codes [12] or positive antigen or polymerase chain reaction (PCR) laboratory test results. ICD-10 codes used to identify cases of COVID-19 included U07.1 (coronavirus disease-2019 (COVID-19)), J12.82 (pneumonia due to coronavirus disease 2019), and M35.81 (multisystem inflammatory syndrome). Uninfected controls were identified from the remaining eligible pool as individuals who lacked any documented COVID-19 diagnosis codes or positive laboratory test results. The World Health Organization defines PCCs as new or continuing symptoms 3 months after infection that last for at least 2 months [1]. However, our machine learning workflow required a broader, data-driven operational definition. We defined candidate PCCs as novel incident conditions recorded during a prolonged 1-year (and 1.5-year for validation) follow-up window after COVID-19 diagnosis. This methodology prioritizes identifying conditions statistically associated with prior infection rather than enforcing strict symptom-duration criteria. To accommodate the distinct requirements of hypothesis generation and statistical validation, we derived two separate analytical cohorts from the overarching eligible pool of eligible individuals. This overarching eligible pool of individuals consisted of controls matched 2:1 with COVID-19 infections on a month-by-month basis based on age, sex, and propensity for COVID-19. The propensity score was calculated using a logistic regression model incorporating over 1800 variables reflecting each patient's comprehensive medical history in the preceding year [Supplemental Table S2]. These variables included inpatient and outpatient diagnoses and procedures (mapped to clinically relevant categories using SAS 9.4 programs from the Agency for Healthcare Research and Quality), as well as dispensed outpatient medications (grouped by American Hospital Formulary Services therapeutic class codes) [13–16]. To ensure a fair timeline comparison, each matched control began their 1-year follow-up period on the exact same calendar date as their infected counterpart. Anchoring the uninfected controls to the infection date of their paired case allowed the analysis to track both groups across identical timelines. For the hypothesis generation cohort, we utilized a 1-year washout period leading up to the COVID-19 diagnosis to exclude pre-existing conditions and to isolate candidate PCCs while minimizing baseline clinical noise. Infections meeting these criteria were matched 1:1 with uninfected controls utilizing the same matched groups from the overarching eligible pool. For the statistical validation phase, we returned to the broader pool of eligible individuals and constructed a larger cohort by utilizing the 2:1 match between uninfected controls and COVID-19 infections to ensure sufficient statistical power for survival analysis. Specifically, utilizing this expanded comparison group optimizes the precision of our hazard ratio estimates and narrows the resulting confidence intervals for less frequent incident outcomes. To ensure a more robust exclusion of chronic conditions for this validation, we applied wider criteria consisting of a two year washout period and a follow-up period of one and a half years.

Propensity score matching was performed using the “MatchIt” package in R [17]. Individuals in the initial cohort were followed through a 1-year observation window. To isolate potential PCCs from the acute phase of initial viral infection, a 30 day washout period immediately following the COVID-19 diagnosis was enforced. Any incident clinical categories recorded during this timeframe were excluded from the analysis. Consequently, the novel diagnoses captured and aggregated over the remainder of the follow-up window represent delayed-onset or persistent conditions, rather than early transient symptoms. To address the high dimensionality of the data, raw International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) diagnosis codes were mapped into clinically homogeneous categories defined by the Healthcare Cost and Utilization Project (HCUP) Clinical Classification Software Refined (CCSR) [18]. This mapping functioned not only as a dimensionality reduction step, but also as a form of representation abstraction, reducing surface-level coding variability while preserving broader clinically meaningful disease structure. Similar abstraction-based approaches in other domains have demonstrated that collapsing highly variable observations into more generalized representations can retain stable latent patterns while improving interpretability; for example, handwriting recognition studies have shown that abstracting stylistic variability can preserve latent identity across diverse samples [19]. In the present study, CCSR categorization served an analogous purpose by grouping sparse ICD-10 observations into clinically coherent categories suitable for downstream hypothesis generation and network modeling. This process collapsed all ICD-10-CM diagnosis codes into 543 clinical categories. These categories were further reduced through a manual filtering process, which was reviewed and finalized by a clinician, to exclude codes with no plausible biological or physiological association with COVID-19, such as external injuries from vehicle accidents or unrelated mechanical complications. This deliberate filtering step ensured obvious background noise was eliminated before downstream steps.

We employed a hyperparameter-tuned lasso regression model to perform feature selection utilizing the hypothesis generation cohort [20]. The dependent variable was the prior COVID-19 diagnosis, and the independent variables were the clinical categories. Because our outcome was dichotomous, the application of the lasso algorithm effectively functioned as an L1-penalized linear probability model. While linear models are unconventional for final probability forecasting of binary outcomes, this approach was intentionally chosen for its superior speed and computational efficiency during dimensionality reduction. Lasso was explicitly chosen for its L1 regularization penalty, which forces the coefficients of non-informative variables to exactly zero [21]. Given that individual patient vectors in the MDR are often sparsely populated, standard unpenalized regressions would struggle with multicollinearity and overparameterization [21]. Lasso mathematically overcomes this sparsity by efficiently shrinking the high-dimensional feature space into a clinically interpretable subset of predictors [22]. The model was trained using 10-fold cross-validation to ensure stability. For our model we specifically used the “lassoCV” estimator from the Python (3.12.10) package Scikit-learn (version 1.7.1) [23]. This estimator selects an optimal regularization strength by computing a coordinate descent path and using 10-fold cross-validation to select the lambda that minimizes average Mean Squared Error (MSE). We emphasize that this linear estimator was used strictly as an automated feature selector, explicitly reserving formal epidemiological risk estimation for the downstream conditional Cox proportional hazards models. Additional hyperparameter tuning was performed to select the maximum number of iterations the model can take to reach convergence. This tuning was performed with the Tree of Parzen Estimators (TPE) hyperparameter tuning algorithm using the Python package Hyperopt (version 0.2.7) [24]. Unlike exhaustive grid searches or undirected random searches, TPE is a Bayesian optimization algorithm that

builds a probabilistic model of the objective function [25]. By utilizing past evaluations to direct the search toward the most promising hyperparameter spaces, TPE ensures that the optimal configuration is found with high computational efficiency [25]. The resulting model was then used to prune out redundant categories, leaving a collection of clinical categories with potential relationships to prior COVID-19.

We then selected the clinical categories assigned positive coefficients in the lasso regression as candidate PCCs. We retained only these categories because our goal was to identify conditions positively associated with prior COVID-19 diagnosis as opposed to those assigned negative coefficients which indicate a protective effect. Using these candidate PCCs identified we sought to characterize the probabilistic dependency structures between the identified conditions; we fitted a Bayesian Network to the selected clinical groups. We utilized the Hill-Climbing structure learning algorithm [26], penalized by the Bayesian Dirichlet equivalent (BDe) score [27]. To ensure model stability, the algorithm was run on 1000 bootstrap samples [28]. A consensus Directed Acyclic Graph (DAG) structure was ascertained by averaging the network structures generated from the bootstrap samples [28]. An arc blacklist was used to restrict the model from making impossible connections, such as arcs indicating COVID-19 being caused by a post-COVID outcome. This was performed using the “bnlearn” package in the R programming language [29].

In order to evaluate the effectiveness of our variable selection process, we utilized a conditional Cox proportional hazards regression model to calculate hazard ratios and 95% confidence intervals for the novel clinical categories identified by the lasso model. For this validation task, we disaggregated the novel CCSR categories to isolate their constituent ICD-10 codes, explicitly selecting the individual codes that accounted for the majority of each category’s overall incidence. To ensure sufficient statistical power for this survival analysis, the statistical validation cohort consisting of the broader pool of eligible MDR records, 1:2 match, was utilized. Models accounted for the matched nature of the data by stratifying on matched, in turn accounting for an individual’s propensity for a COVID-19 infection [30]. Individuals were censored if they became ineligible for TRICARE for two consecutive months, after the passage of 548 days, or (controls only) if they developed COVID-19 during the follow-up period. The Cox models did not account for competing risks or repeated diagnoses. The statistical significance level was set at 0.05, and the proportional hazards assumption was assessed using Schoenfeld residuals. This final validation step was performed using SAS 9.4 (Cary, NC) [31]. This study was reviewed and approved by the Uniformed Services University of the Health Sciences IRB.

3. Results

Application of the study inclusion criteria and a 2:1 match between uninfected controls and COVID-19 infections yielded an overarching eligible pool of 311,367 Active-Duty Tricare beneficiaries [Supplemental Table S1]. For the feature selection phase, the application of a 1-year washout period and 1:1 propensity matching yielded an initial feature-selection cohort of 203,578 individuals, consisting of 101,789 individuals with a confirmed COVID-19 diagnosis and an equal number of uninfected controls. Following the initial mapping of the encounter data within this cohort, and the removal of codes with no plausible association with COVID-19, a baseline of 392 clinical categories was established for subsequent machine learning evaluation.

Following hyperparameter tuning, the lasso regression model converged optimally at 4400 iterations with a lambda (regularization penalty) of 0.015 [Figure 1]. As illustrated in Figure 1, the ten colored dotted lines trace the model’s performance across individual cross-validation folds, demonstrating tight convergence and minimal variance between data splits, while the solid black line tracks the overall average Mean Squared Error (MSE). At

this optimal threshold, the model minimized the MSE to 0.244 and achieved an R-squared of 0.014 or 1.4%. For comparison, at the worst lambda, the MSE was only minimized to 0.406, representing a substantial increase in error. While an R-squared value of 0.014 or 1.4% appears low by traditional predictive applications, it is entirely typical for population-level electronic health records where viral infection is modeled against vast background noise, driving down the variance measured by the R-squared value. Within this context, the value of the model rests on its performance as an automated, regularizing feature selector rather than a clinical forecasting tool. Applying these optimized parameters reduced the initial diagnosis set from 392 categories to 26 that were assigned lasso coefficients, with only 8 assigned positive coefficients. The eight selected categories primarily consisted of well-documented post-COVID-19 sequelae, including respiratory issues, malaise and fatigue, headache, joint pain, and nervous system disorders [Figure 2] [7,10]. Because the association of these six conditions with prior COVID-19 infection is already well established in the literature, they were excluded from further statistical validation. Additionally, two categories stood out as novel potential PCCs: disorders of the ear and personality disorders. Because the association of these two specific conditions with prior COVID-19 has not been as extensively established, they were deliberately selected for our confirmatory survival analysis to evaluate the pipeline's capacity to discover novel conditions.

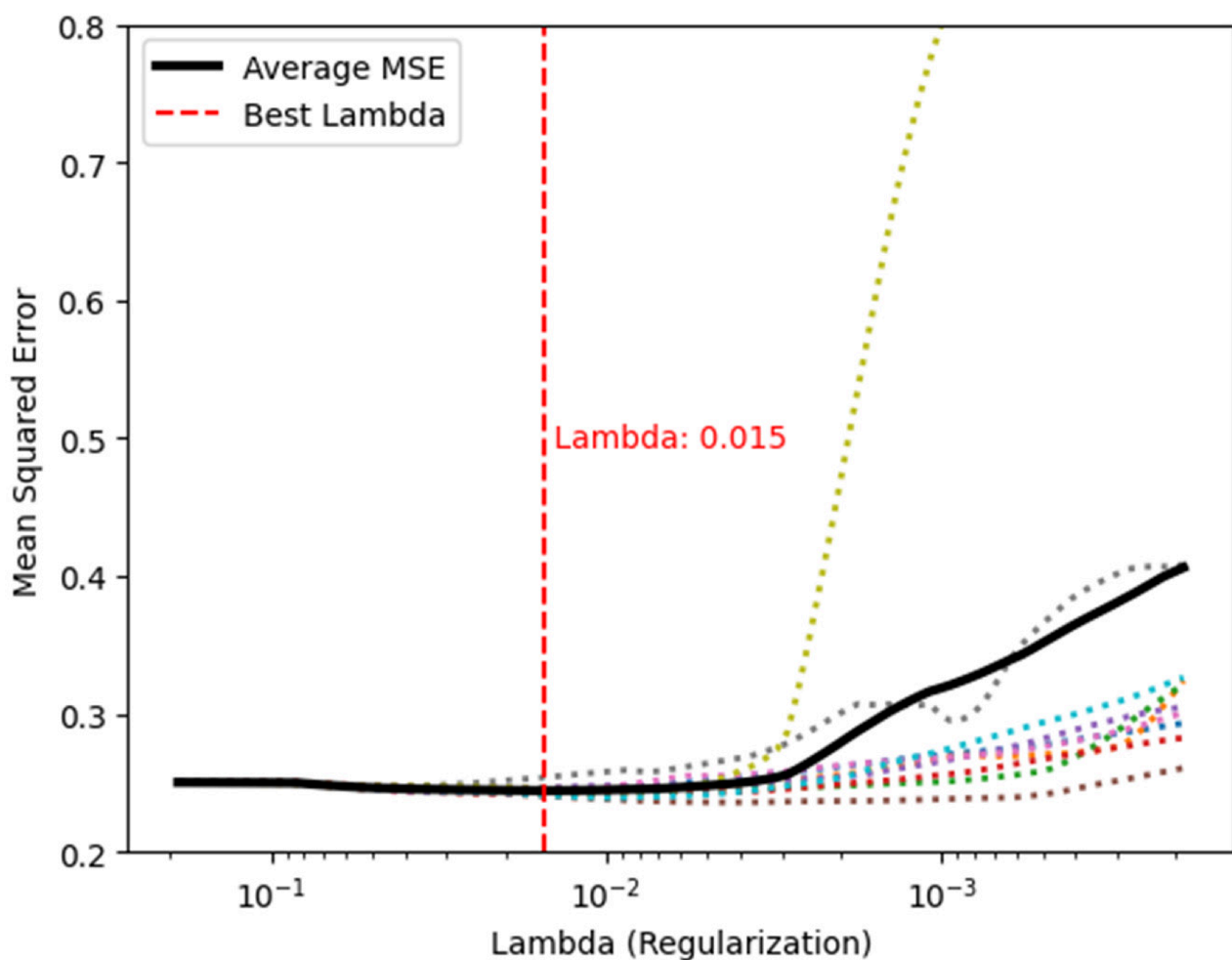


Figure 1. Optimal Lambda selection via 10-Fold Cross-Validation in lasso Regression.

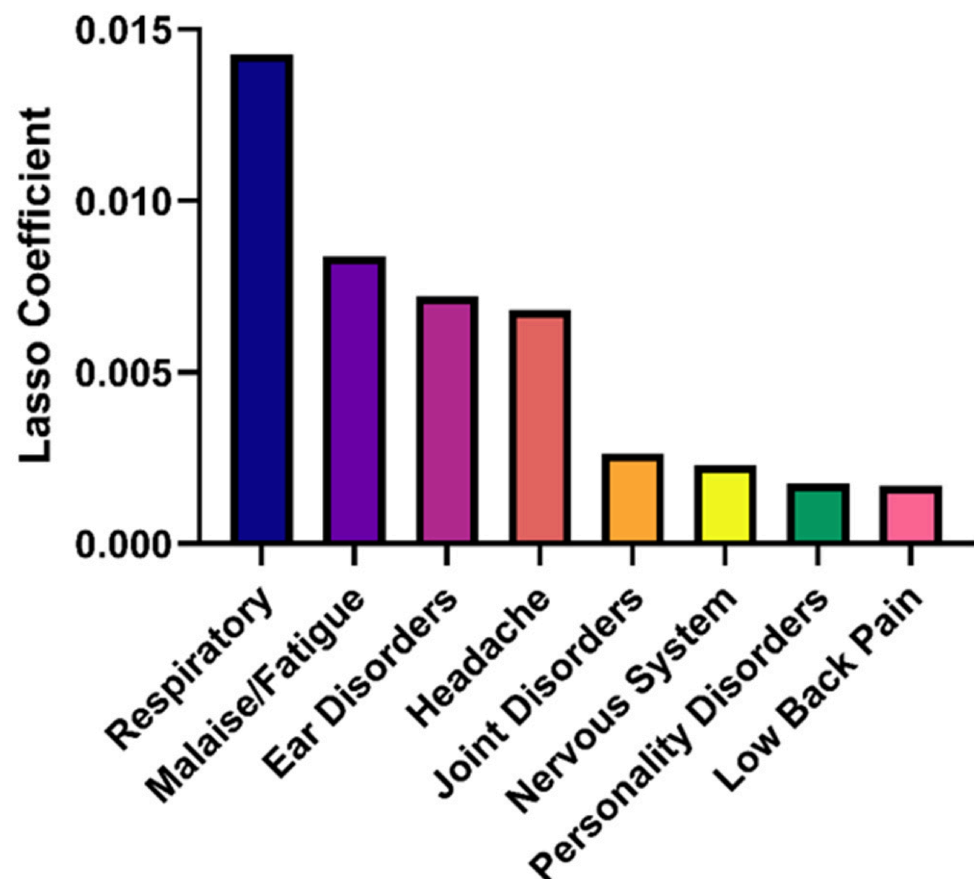


Figure 2. Positive lasso Regression Coefficients assigned to HCUP CCS Categories.

Within the ‘disorders of the ear’ category, 48.62% of the incidence was attributable to tinnitus-related ICD-10 codes (H9313, H9312, H9319). Within the ‘personality disorders’ category, 91.43% of the incidence was attributable to borderline personality disorder (F603) and personality disorder, unspecified (F609).

The consensus DAG structure [Figure 3] revealed potential probabilistic dependency structures and diagnostic association pathways between the identified conditions. The model showed a Bayesian Information Criterion (BIC) improvement of -3248 over the baseline network, confirming a strong fit for the available data. The structure suggested a cascade of symptoms, moving from COVID-19 to headaches, nervous system signs, respiratory symptoms, and subsequently malaise/fatigue and disorders of the ear.

To conduct the confirmatory survival analysis on the novel candidate PCCs identified by this process, we returned to the overarching eligible pool of 311,367 individuals in the form of the statistical validation cohort. From this cohort, the analysis revealed that individuals with a prior COVID-19 infection had a 1.17 times higher risk of developing tinnitus compared to uninfected controls (HR: 1.17, 95% CI: 1.12–1.22, $p < 0.0001$), consistent with the higher cumulative incidence visualized in Figure 4. This significant association solidified the legitimacy of tinnitus as a PCC. In contrast, there was no significantly elevated risk for developing personality disorders following a COVID-19 infection (HR: 1.11, 95% CI: 0.97–1.26, $p = 0.1319$), reflecting the lack of significant divergence in the cumulative incidence shown in Figure 5.

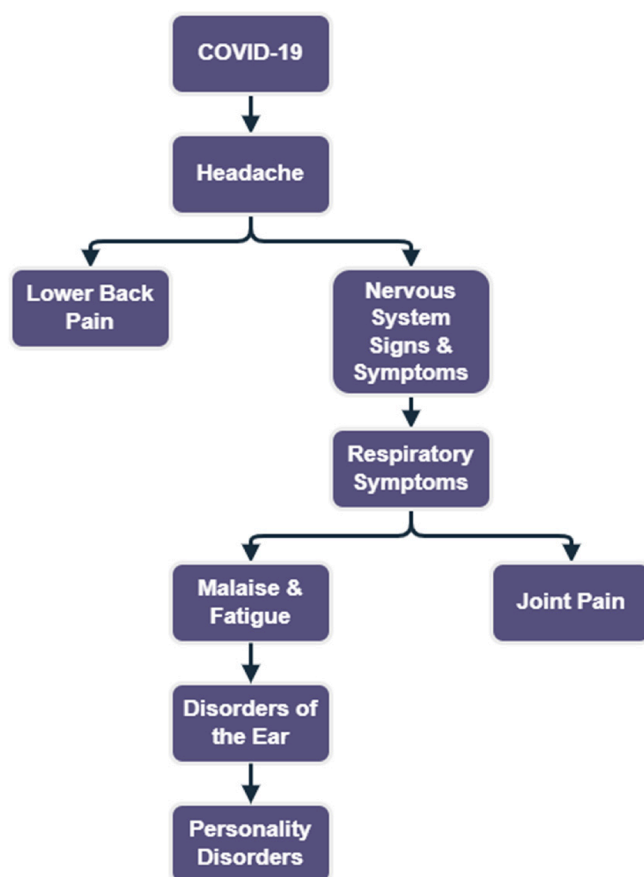


Figure 3. Directed Acyclic Graph of Candidate Probabilistic Dependency Structures between PCCs.

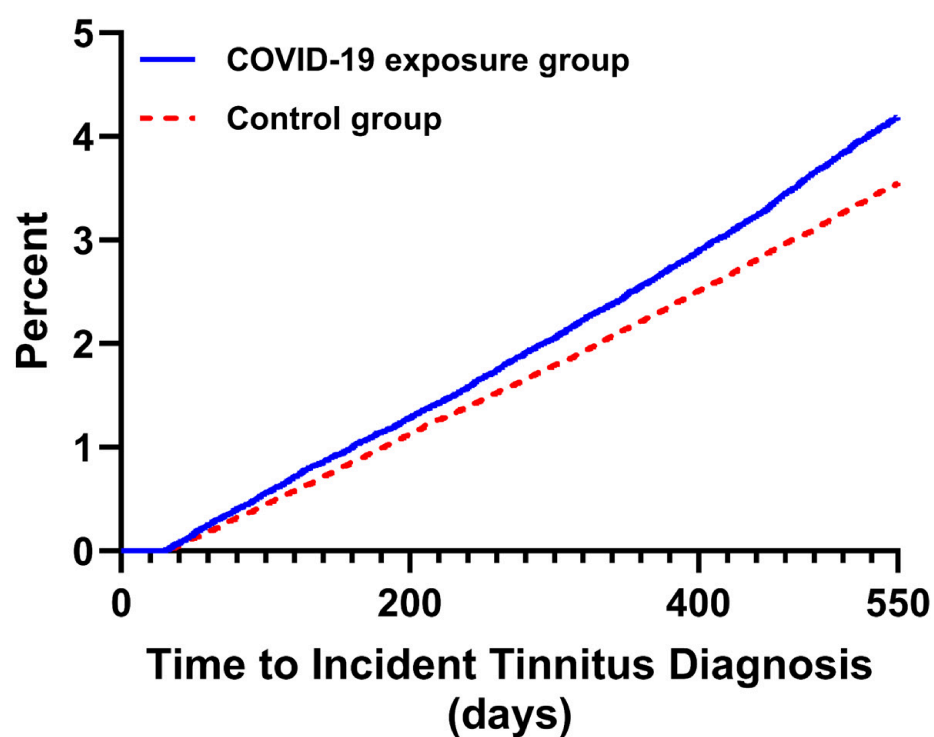


Figure 4. Cumulative Incidence Curve for Tinnitus by COVID-19 Infection Status.

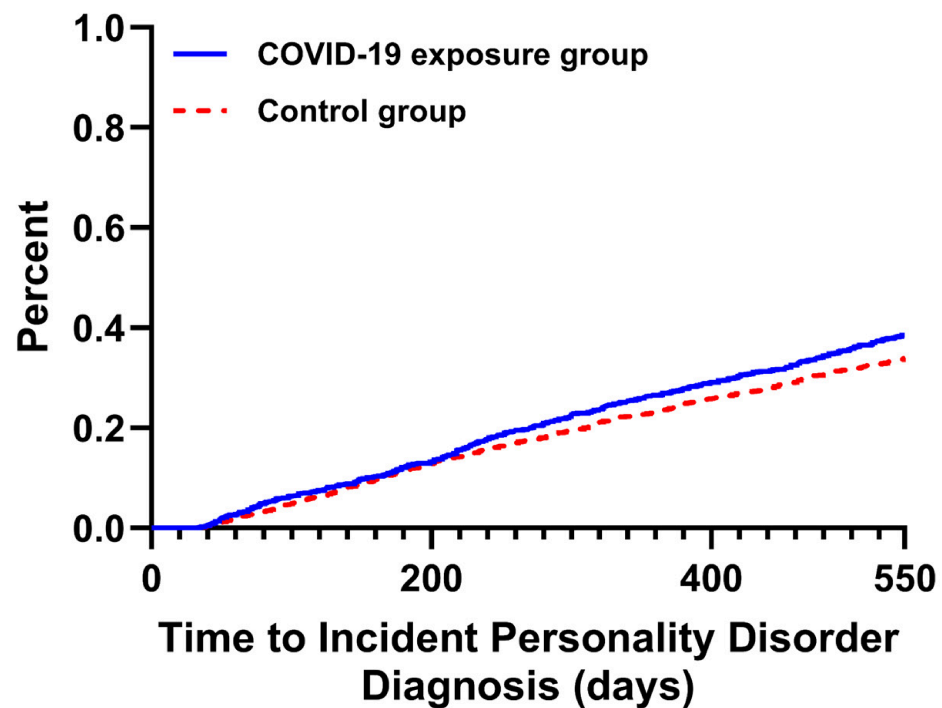


Figure 5. Cumulative Incidence Curve for Personality Disorders by COVID-19 Infection Status.

4. Discussion

This study demonstrates the effectiveness of a sequential machine learning-guided analytical workflow for addressing the limitations of analyzing complex, sparse, high-dimensional healthcare encounter data. The primary strength of this approach lies in its ability to overcome high dimensionality and extreme sparsity that characterize healthcare encounter data. Rather than representing a routine application of standard machine learning methodologies, the value of this framework rests upon its deliberate structural sequence, functioning as an operational bridge between unconstrained automated data discovery and strict statistical validation. While the MDR contains millions of records, individual patient vectors are often sparsely populated with only a few diagnosis codes among tens of thousands of possibilities. The model's robust performance is evidenced by its automated selection of well-documented post-COVID-19 conditions, including respiratory symptoms, malaise, fatigue, and nervous system disorders. These selected categories align strongly with the World Health Organization's international consensus definition, which identifies debilitating fatigue, shortness of breath, and cognitive dysfunction as hallmark symptoms of the condition [1]. Furthermore, physiological evaluations frequently demonstrate underlying cardiopulmonary abnormalities in these patients, revealing prolonged fatigue, ventilatory inefficiency, and impaired oxygen delivery [7]. Concurrently, the model's selection of nervous system disorders is supported by extensive evidence of post-COVID-19 neuroinflammation, microvascular injury, and neural cell dysregulation, which are known to drive persistent neurological and neuropsychiatric manifestations [10]. This serves as a strong internal validation, confirming that the feature selection process was capable of isolating relevant biological signals without manual curation. As introduced in our results, while an R-squared value of 0.014 is low for the traditional predictive applications of lasso regression, it accurately reflects that a prior COVID-19 diagnosis accounts for only a minor 1.4% variance in the cohort's overall clinical profile. More importantly, it demonstrates the model's effectiveness in achieving our goal of filtering out pervasive background noise to successfully isolate critical predictors. The necessity of automated hyperparameter tuning is underscored by the model's high sensitivity to the regularization

penalty. Applying the optimal lambda minimized the mean squared error (MSE) to 0.244, whereas the worst-performing alpha yielded an MSE of 0.406, representing a substantial 66.39% increase in error.

The extended washout and follow up periods used in the confirmatory survival analysis helped to ensure maximum diagnostic clarity while capturing late-emerging post-COVID sequelae. Similarly, during our initial hypothesis generation phase, while we aggregated diagnoses over a prolonged window rather than strictly enforcing the WHO's 3-month onset and 2-month duration criteria, the 30-day post-infection washout period effectively excluded early, transient symptoms related to the acute phase of COVID-19. Therefore, the clinical categories captured across both our feature selection and validation phases are highly likely to reflect true delayed-onset or prolonged sequelae rather than early acute symptoms. Additionally, our 1 to 2 year pre-infection washout periods and the use of propensity matched controls functionally align with the WHO principle of excluding symptoms due to other background diseases [1]. However, as a limitation, our data-driven operational definition cannot strictly verify an uninterrupted 2-month symptom duration or definitively rule out other underlying conditions on an individual case-by-case basis. Utilizing a 1:2 case-control ratio for confirmatory survival analysis allowed us to include a greater proportion of the uninfected population to evaluate longitudinal person-time at risk. This expanded comparison group increased the baseline event pool, which can yield higher precision for our hazard ratio estimates and help narrow confidence intervals for less frequent incident outcomes like tinnitus.

The discovery of tinnitus as a PCC validates the model's capacity to identify less studied, clinically relevant sequelae that are frequently overlooked in traditional hypothesis-driven research. This finding aligns with and strengthens the existing literature by providing large-scale statistical validation of early clinical signals. For instance, early case reports by Chirakkal et al. first flagged the damaging impact of the COVID-19 virus on the inner ear, urging clinicians to be aware of auditory-vestibular symptoms as novel peripheral nervous system manifestations [11]. Subsequently, a comprehensive systematic review by Beukes et al. confirmed this broader trend, hypothesizing that the condition is driven by direct viral damage to cochlear outer hair cell functioning or neuroinflammatory ischemic damage [32]. By statistically validating this association at a population level through survival analysis (HR: 1.17), this study provides robust confirmation of these prior clinical observations. Although a hazard ratio of 1.17 represents a relatively modest effect size at the individual level, the unprecedented global scale of the COVID-19 pandemic implies that even a slight increase in relative risk translates to a massive absolute number of tinnitus cases, highlighting the significant public health relevance of this finding. Ultimately, this reinforces the utility of machine learning as an effective hypothesis-generation tool for novel disease manifestations.

The findings from the model pertaining to Borderline Personality Disorder (BPD) that were not different from the control group could indicate transient neuropsychiatric symptoms that mimic BPD that resolve as neuroinflammation improves [33]. However, prolonged neuroinflammation is thought to disrupt neurotransmitter production and affect brain regions involved in emotional process and impulse control [34], leading to symptoms such as severe mood swings [35], dissociation and brain fog [34], emotional dysregulation [36] and perceived instability [37]. In addition, the presence of social distancing, movement control, changes in availability of daily essential goods, and job stability could also contribute towards BPD-type symptoms [36]. However, more testing to validate BPD in chronic PCCs is needed.

While traditional statistical methods remain foundational to epidemiology, they can be challenged by the high dimensionality of modern healthcare datasets [38]. Additionally,

while hypothesis testing is a highly formalized scientific activity, hypothesis generation has historically remained a largely informal process [38]. Machine learning offers a complementary approach that is less constrained by prior assumptions, assisting in the modeling of complex relationships to uncover hidden patterns [38]. Applying machine learning algorithms to high-dimensional data may help researchers systematically generate novel, interpretable hypotheses that might otherwise go unnoticed [39]. In this study, using machine learning to assist in hypothesis generation facilitated the extraction of unexpected clinical sequelae from an initial pool of hundreds of categories, which were subsequently evaluated through rigorous statistical validation.

This study also encountered specific hurdles and technical limitations inherent to large scale encounter data analysis. The identification of personality disorders as a candidate PCC, which was subsequently ruled out by survival analysis (HR: 1.11, $p = 0.13$), highlights the limitations imposed by feature sparsity. The initial selection of this category was likely a false positive driven by the model attempting to force convergence on sparse data vectors. This discrepancy underscores the necessity of the multi-step validation process. While machine learning models are powerful for reducing dimensionality, their outputs must be subjected to rigorous epidemiological verification. However, it is important to note that this downstream statistical validation phase also carries inherent constraints. Specifically, our conditional Cox proportional hazards models did not account for competing risks or repeated diagnoses, which may influence the precise estimation of longitudinal risk.

Additionally, the Bayesian Network modeling faced constraints related to the depth of the available data. While the consensus DAG showed a significant BIC improvement of -3248 over the baseline, indicating a strong fit for the observed data, the model's interpretation is inherently limited by the assumption of causal sufficiency [40]. In causal inference frameworks, causal sufficiency dictates that a dataset must contain all common causes of the variables being modeled to accurately infer causal relationships [40]. Because the network could only map relationships among the specific clinical categories present in the encounter data, it remained blind to unmeasured confounders such as patient lifestyle, genetics, or environmental stressors. Furthermore, because our uninfected control cohort was defined by the absence of documented COVID-19, encounters or laboratory results, it is possible that controls experienced asymptomatic or unreported infections, though we do expect the latter to be minimal as active-duty service members are required to seek medical care for missed work days and have universal access to healthcare. Furthermore, administrative billing data does not allow us to ascertain the patient's primary subjective reason for seeking care, limiting our ability to differentiate whether follow-up encounters were explicitly driven by perceived COVID-19 sequelae or unrelated general illnesses. This potential misclassification bias is an inherent limitation of utilizing observational healthcare encounter data during a widespread pandemic. This potential misclassification bias is an inherent limitation of utilizing observational healthcare encounter data during a widespread pandemic. These unobserved variables could act as unseen common causes, potentially influencing the strength and direction of the observed causal arcs or creating spurious associations [40]. Consequently, while the DAG provides a robust probabilistic map of diagnostic relationships, the presence of these unmeasured mediators makes it difficult to definitively determine biological causality from this encounter data alone.

To properly interpret the topology of the consensus DAG [Figure 3], it is critical to evaluate these links as probabilistic dependency structures learned from observational data matrices rather than a biological model of direct clinical causation. The directed arrows signify conditional independence constraints determined by the score-based algorithm. For example, an arrow extending from 'Headache' to 'Lower Back Pain' does not imply that headaches physically precipitate spinal pathology. Instead, it demonstrates that within this

observational cohort, knowledge of a patient's medical encounter for a headache significantly alters the conditional probability distribution of recording a lower back pain code within the electronic health record. This pattern may be capturing a latent, or unmeasured healthcare-seeking behavior common to a highly specific subpopulation of post-COVID patients, rather than a chronological or biological pathway.

Another critical limitation of our structural learning approach is the reliance on a static Bayesian Network. A static network is inherently limited in its ability to capture temporal dynamics [41]. In our methodology, diagnoses were aggregated over a 1 year follow-up window; therefore, an arc from one CCSR category to another (e.g., from headaches to nervous system signs) indicates a strong conditional probability of co-occurrence rather than a verified chronological sequence. This limitation is pronounced when observing the order of the cascade of conditions within the resulting DAG, as headaches for example are not typically seen as an early defining symptom of COVID-19. The current model cannot discern whether these conditions present simultaneously or in a strict temporal cascade. To fully elucidate the temporal evolution of Post-COVID-19 conditions, future research should transition from static to Dynamic Bayesian Networks (DBNs), which can incorporate time-series data to model the state of the clinical network at distinct intervals [41]. Despite these limitations, this study provides a generalizable framework that successfully filters high-dimensional noise to generate and validate testable hypotheses for complex disease sequelae.

5. Conclusions

As healthcare datasets grow in volume and complexity, traditional statistical methods alone may be insufficient for uncovering hidden clinical patterns [38]. By pairing machine learning-assisted feature selection with statistical validation, this study explored a hybrid analytical pipeline aimed at bridging the gap between informal hypothesis generation and formal hypothesis testing [39]. This framework aided in validating tinnitus as a post-COVID-19 condition, suggesting that machine learning can be a useful tool for extracting interpretable hypotheses from sparse, high-dimensional encounter data. This methodology may serve as a helpful supplement in modern epidemiological research, offering a generalizable approach to exploring complex, unknown health outcomes related to prior infections. To position this work as a foundation for broader discovery, future research should prioritize external validation in civilian populations and incorporate subgroup analyses based on age, sex, military rank, and the severity of acute infection. Additionally, future works should transition from the use of static Bayesian Networks toward dynamic Bayesian Networks to more accurately capture progression patterns. Similarly, the incorporation of sensitivity analyses that restrict the follow-up window to specific periods (such as the latter half of the year) may help further isolate late-emerging clinical trajectories. Targeted clinical investigations should also be conducted to explore the exact mechanisms of these identified sequelae, such as investigating whether incident tinnitus is a direct consequence of COVID-19 infection or secondary to other COVID-19-associated manifestations.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/sci8070153/s1>, Supplemental Table S1. Demographic Breakdown of Eligible Active-Duty Service Members, Stratified by COVID-19 Infection Status. Supplemental Table S2. Variables used in Propensity Score Model.

Author Contributions: Conceptualization, A.S., D.N. and C.M.N.; Methodology, A.S., Z.B., D.N. and C.M.N.; Formal analysis, J.S. and Z.B.; Data curation, J.S. and Z.B.; Writing—original draft, J.S.; Writing—review & editing, J.S., A.S., Z.B., Remle Scott and D.N.; Supervision, A.S., R.S. and C.M.N.;

Project administration, C.M.N.; Funding acquisition, C.M.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by USU Intramural Research Program, grant number HU0001-22-2-0066.

Institutional Review Board Statement: The study protocol (2020-065) was reviewed and deemed no-more-than-minimal-risk human-subject research by the institutional review board of the Uniformed Services University of the Health Sciences, Bethesda, MD, USA. Approval Date: 2 October 2021.

Informed Consent Statement: Informed consent and HIPAA consent are not available as the Institutional Review Board granted waivers due to the nature of the data being accessed.

Data Availability Statement: All data were accessed from the Military Health System database, which requires a valid data sharing agreement due to the nature of personal health information and cannot be shared without period approval from the U.S. Department of Defense, Defense Health Agency.

Conflicts of Interest: Authors Jed Shakarji, Apryl Susi, Zella Berill and Remle Scott were employed by the company Henry M. Jackson Foundation for the Advancement of Military Medicine. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Soriano, J.B.; Murthy, S.; Marshall, J.C.; Relan, P.; Diaz, J.V.; WHO Clinical Case Definition Working Group on Post-COVID-19 Condition. A Clinical Case Definition of Post-COVID-19 Condition by a Delphi Consensus. *Lancet Infect. Dis.* **2022**, *22*, e102–e107. [CrossRef] [PubMed]
2. Vu, T.; McGill, S.C. An Overview of Post-COVID-19 Condition (Long COVID). *Can. J. Health Technol.* **2021**, *1*, 1–31. [CrossRef]
3. Nalbandian, A.; Sehgal, K.; Gupta, A.; Madhavan, M.V.; McGroder, C.; Stevens, J.S.; Cook, J.R.; Nordvig, A.S.; Shalev, D.; Sehrawat, T.S.; et al. Post-Acute COVID-19 Syndrome. *Nat. Med.* **2021**, *27*, 601–615. [CrossRef] [PubMed]
4. Lopez-Leon, S.; Wegman-Ostrosky, T.; Perelman, C.; Sepulveda, R.; Rebolledo, P.A.; Cuapio, A.; Villapol, S. More than 50 Long-Term Effects of COVID-19: A Systematic Review and Meta-Analysis. *Sci. Rep.* **2021**, *11*, 16144. [CrossRef] [PubMed]
5. Alwan, N.A.; Johnson, L. Defining Long COVID: Going Back to the Start. *Med* **2021**, *2*, 501–504. [CrossRef] [PubMed]
6. Munblit, D.; Nicholson, T.; Akrami, A.; Apfelbacher, C.; Chen, J.; De Groote, W.; Diaz, J.V.; Gorst, S.L.; Harman, N.; Kokorina, A.; et al. A Core Outcome Set for Post-COVID-19 Condition in Adults for Use in Clinical Practice and Research: An International Delphi Consensus Study. *Lancet Respir. Med.* **2022**, *10*, 715–724. [CrossRef] [PubMed]
7. Astin, R.; Banerjee, A.; Baker, M.R.; Dani, M.; Ford, E.; Hull, J.H.; Lim, P.B.; McNarry, M.; Morten, K.; O'Sullivan, O.; et al. Long COVID: Mechanisms, Risk Factors and Recovery. *Exp. Physiol.* **2023**, *108*, 12–27. [CrossRef] [PubMed]
8. Xie, Y.; Choi, T.; Al-Aly, Z. Postacute Sequelae of SARS-CoV-2 Infection in the Pre-Delta, Delta, and Omicron Eras. *N. Engl. J. Med.* **2024**, *391*, 515–525. [CrossRef] [PubMed]
9. Komaroff, A.L.; Lipkin, W.I. ME/CFS and Long COVID Share Similar Symptoms and Biological Abnormalities: Road Map to the Literature. *Front. Med.* **2023**, *10*, 1187163. [CrossRef] [PubMed]
10. Monje, M.; Iwasaki, A. The Neurobiology of Long COVID. *Neuron* **2022**, *110*, 3484–3496. [CrossRef] [PubMed]
11. Chirakkal, P.; Al Hail, A.N.; Zada, N.; Vijayakumar, D.S. COVID-19 and Tinnitus. *Ear Nose Throat J.* **2021**, *100*, 160S–162S. [CrossRef] [PubMed]
12. Centers for Medicare & Medicaid Services (U.S.); National Center for Health Statistics (U.S.); American Hospital Association (AHA); American Health Information Management Association (AHIMA). ICD-10-CM Official Guidelines for Coding and Reporting: FY 2019: (October 1, 2018–September 30, 2019). 2023. Available online: <https://stacks.cdc.gov/view/cdc/133289> (accessed on 24 March 2023).
13. Francke, D.E. Uses of AHFS Classification System. *Am. J. Health-Syst. Pharm.* **1963**, *20*, 119–120. [CrossRef]
14. Sexton, K.W.; Berill, Z.; Susi, A.; Coene, J.; Madison, K.E.; Nylund, C.M. The Metabolic Aftershock: COVID-19 and Metabolic Disease Risk Among U.S. Active-Duty Military Personnel. *Metabolites* **2025**, *15*, 795. [CrossRef] [PubMed]
15. Healthcare Cost and Utilization Project (HCUP). Clinical Classifications Software Refined (CCSR) for ICD-10-CM Diagnoses, V2022.1. Available online: <https://hcup-us.ahrq.gov/toolssoftware/ccsr/dxcsr.jsp> (accessed on 24 March 2023).
16. Healthcare Cost and Utilization Project (HCUP). Clinical Classifications Software Refined (CCSR) for ICD-10-CM Diagnoses, V2023.1. Available online: <https://hcup-us.ahrq.gov/toolssoftware/ccsr/dxcsr.jsp> (accessed on 24 March 2023).

17. Ho, D.E.; Imai, K.; King, G.; Stuart, E.A. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *J. Stat. Softw.* **2011**, *42*, 1–28. [\[CrossRef\]](#)
18. Healthcare Cost and Utilization Project (HCUP). Clinical Classifications Software Refined (CCSR) for ICD-10-CM Diagnoses, V2024.1. Available online: <https://hcup-us.ahrq.gov/toolssoftware/ccsr/dxccsr.jsp> (accessed on 24 March 2023).
19. Huang, J.; Feng, Y.; Cui, F.-Q.; Zhang, X.; Liu, Z.; Liu, X.; Liu, J.; Zhang, F.; Li, M. Identifying Who You Are No Matter What You Write Through Abstracting Handwriting Style. *IEEE Trans. Dependable Secur. Comput.* **2026**, *23*, 6890–6905. [\[CrossRef\]](#)
20. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1996**, *58*, 267–288. [\[CrossRef\]](#)
21. Sudjai, N.; Duangsaphon, M.; Chandhanayingyong, C. Relaxed Adaptive Lasso for Classification on High-Dimensional Sparse Data with Multicollinearity. *Int. J. Stat. Med. Res.* **2023**, *12*, 97–108. [\[CrossRef\]](#)
22. Avalos, M.; Pouyes, H.; Grandvalet, Y.; Orriols, L.; Lagarde, E. Sparse Conditional Logistic Regression for Analyzing Large-Scale Matched Data from Epidemiological Studies: A Simple Algorithm. *BMC Bioinform.* **2015**, *16*, S1. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Müller, A.; Nothman, J.; Louppe, G.; et al. Scikit-Learn: Machine Learning in Python. *arXiv* **2012**. [\[CrossRef\]](#)
24. Bergstra, J.; Komer, B.; Eliasmith, C.; Yamins, D.; Cox, D.D. Hyperopt: A Python Library for Model Selection and Hyperparameter Optimization. *Comput. Sci. Discov.* **2015**, *8*, 014008. [\[CrossRef\]](#)
25. Ahmed Arafa, A.; Radad, M.; Badawy, M.; El-Fishawy, N. Logistic Regression Hyperparameter Optimization for Cancer Classification. *Menoufia J. Electron. Eng. Res.* **2022**, *31*, 1–8.
26. Jacobson, S.H.; Yücesan, E. Analyzing the Performance of Generalized Hill Climbing Algorithms. *J. Heuristics* **2004**, *10*, 387–405. [\[CrossRef\]](#)
27. Heckerman, D.; Geiger, D.; Chickering, D.M. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Mach. Learn.* **1995**, *20*, 197–243. [\[CrossRef\]](#)
28. Friedman, N.; Goldszmidt, M.; Wyner, A. Data Analysis with Bayesian Networks: A Bootstrap Approach. *arXiv* **2013**. [\[CrossRef\]](#)
29. Scutari, M. Learning Bayesian Networks with the Bnlearn R Package. *J. Stat. Softw.* **2010**, *35*, 1–22. [\[CrossRef\]](#)
30. Faries, D.; Leon, A.; Haro, J.M.; Obenchain, R.L. *Analysis of Observational Health Care Data Using SAS*; SAS Institute Inc.: Cary, NC, USA, 2010.
31. SAS Institute Inc. *SAS/STAT® 9.3 User's Guide*; SAS Institute Inc.: Cary, NC, USA, 2011.
32. Beukes, E.; Ulep, A.J.; Eubank, T.; Manchaiah, V. The Impact of COVID-19 and the Pandemic on Tinnitus: A Systematic Review. *J. Clin. Med.* **2021**, *10*, 2763. [\[CrossRef\]](#) [\[PubMed\]](#)
33. Choi-Kain, L.W.; Sahin, Z.; Traynor, J. Borderline Personality Disorder: Updates in a Postpandemic World. *FOCUS* **2022**, *20*, 337–352. [\[CrossRef\]](#) [\[PubMed\]](#)
34. Orfei, M.D.; Porcari, D.E.; D'Arcangelo, S.; Maggi, F.; Russignaga, D.; Ricciardi, E. A New Look on Long-COVID Effects: The Functional Brain Fog Syndrome. *J. Clin. Med.* **2022**, *11*, 5529. [\[CrossRef\]](#) [\[PubMed\]](#)
35. Preti, E.; Di Pierro, R.; Fanti, E.; Madeddu, F.; Calati, R. Personality Disorders in Time of Pandemic. *Curr. Psychiatry Rep.* **2020**, *22*, 80. [\[CrossRef\]](#) [\[PubMed\]](#)
36. Chong, S.C. Psychological Impact of Coronavirus Outbreak on Borderline Personality Disorder from the Perspective of Mentalizing Model: A Case Report. *Asian J. Psychiatry* **2020**, *52*, 102130. [\[CrossRef\]](#) [\[PubMed\]](#)
37. Starcevic, V.; Janca, A. Personality Dimensions and Disorders and Coping with the COVID-19 Pandemic. *Curr. Opin. Psychiatry* **2022**, *35*, 73–77. [\[CrossRef\]](#) [\[PubMed\]](#)
38. Atias, D.; Ashri, S.; Goldbourt, U.; Benyamini, Y.; Gilad-Bachrach, R.; Hasin, T.; Gerber, Y.; Obolski, U. Machine Learning in Epidemiology: An Introduction, Comparison with Traditional Methods, and a Case Study of Predicting Extreme Longevity. *Ann. Epidemiol.* **2025**, *110*, 23–33. [\[CrossRef\]](#) [\[PubMed\]](#)
39. Ludwig, J.; Mullainathan, S. Machine Learning as a Tool for Hypothesis Generation. *Q. J. Econ.* **2024**, *139*, 751–827. [\[CrossRef\]](#)
40. Pearl, J. *Causality: Models, Reasoning, and Inference*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2009. [\[CrossRef\]](#)
41. Alves, J.M.; Martins, T.; Esteves, S.; Dias, C.C.; Rodrigues, P.P. Temporal and Dynamic Bayesian Networks for Prognosis and Diagnosis in Clinical Settings: A Scoping Review. *Comput. Biol. Med.* **2025**, *198*, 111193. [\[CrossRef\]](#) [\[PubMed\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.