UNIVERSITÄT ZU LÜBECK

From the Institute of Experimental Dermatology
of the University of Luebeck
Director: Prof. Dr. Hauke Busch

Comprehensive Genetic and Comorbidity Profiling of Autoimmune Diseases: Integrating UK Biobank, TriNetX, and Global Data

Dissertation
for Fulfillment of
Requirements for the
Doctoral Degree
of the University of Luebeck
from the Department of Natural Sciences

Submitted by
Rochi Saurabh
From India
2025

First referee: Prof. Dr. rer. nat. Hauke Busch

Second referee: PD Dr. hum. biol. Zouhair Aherrahrou

Date of oral examination: September 26th, 2025

Approved for printing: October 6th, 2025

Acknowledgments

I express my profound gratitude to my supervisor, Prof. Dr. rer. nat. Hauke Busch, whose unwavering guidance, support, and encouragement have been indispensable throughout my doctoral studies. I am equally indebted to my co-supervisor, Prof. Dr. rer. biol. hum. Inke R. König, whose incisive insights and constructive counsel have been pivotal to the development of this dissertation. I extend my deepest gratitude to my mentors, Prof. Dr. rer. nat. Inken Wohlers and Dr. Marius Möller, for the substantial time and effort they invested and for their enduring mentorship and discerning guidance.

I am also indebted to Dr. rer. nat. habil. Anke Fähnrich and Dr. rer. nat. Misa Hirose for enabling me to undertake additional investigations beyond the scope of this thesis. I thank Dr. rer. hum. biol. Césaire J. K. Fouodo and Anikamila Cani for collaborating with me on separate projects. My sincere thanks go to Dr. rer. nat. Axel Künstner, whose invaluable guidance helped me overcome numerous technical challenges, and to Sen, whose steadfast support has accompanied me throughout my doctoral journey. Lastly, I am grateful to every member of the laboratory for cultivating a productive research environment and for the many valued memories we share.

I am sincerely grateful to the spokespersons, Prof. Dr. Ralf Ludwig, and Prof. Dr. Jennifer Hundt for providing me with the opportunity to participate in RTG2633 research project. I also wish to acknowledge the outstanding organizational efforts of Dr. Skadi Lange, Dr. Laura Kirchhoff, and Sina Jäschke.

Finally, I extend my profound gratitude to Mrs. Rashmi Verma, my mother, and Mr. Suresh Kr.Verma, my grandfather, whose steady faith in my goals has always inspired me. I am equally grateful to Dr. Shubham Ranjan, my husband, whose support and love have carried me through this journey. I extend my gratitude to my brothers and other family members for their unwavering encouragement and support at every step. I also recognize the effort and persistence required to bring this work to completion.

॥सर्वे योगदानकर्तारः तथा समर्थकाः स्मरणीयाः, मम प्रयासानां प्रेरकानां कृते हृदयङ्गमः धन्यवादः॥
All contributors and supporters are memorable, and heartfelt thanks to those who inspired my efforts

Abstract

Introduction: Autoimmune diseases share a general mechanism of autoantigens harming tissues. Still, they are phenotypically diverse, with genetic as well as environmental factors contributing to their etiology at varying degrees. Associated genomic loci and variants have been identified in numerous genome-wide association studies (GWAS), whose results are increasingly used for polygenic scores (PGS) that are used to predict disease risk. Current, publicly available GWAS and PGS data for autoimmune diseases are examined using information from the GWAS Catalog and PGS Catalog. Which summarizes the autoimmune conditions investigated, the individual studies conducted, and their reported findings. The UK Biobank (UKB), an academic repository of genetic and phenotypic data on autoimmune disease cases, is also evaluated. Further, study quantifies diagnostic overlap among autoimmune disease cases in the UKB using phenotype data, employs PGS to cluster individuals and assess heterogeneity in genetic risk profiles, and examines genetic variants in the unrelated White British subgroup guided by functional annotations. It then integrates common, rare, and high-impact variant data to identify genes shared across multiple autoimmune diseases and map them to relevant biological pathways. Additionally, study examines how frequently rare and common autoimmune diseases occur together, quantifying their comorbidity and co-occurrence by analysing patient records from both the UKB and the TriNetX (TNX) global health-research network. It also uses the TNX dataset alone to investigate how the rare autoimmune blistering disorder pemphigus is associated with other conditions across multiple ancestral groups, providing population-specific insights into its clinical profile.

Methods: Phenotypic data for all 502,371 UK Biobank participants underpin the overlap analysis, while an enhanced subset of 104,544 individuals spanning six autoimmune diseases is used to examine risk-score distributions. Genetic analyses apply a significance threshold of $-\log_{10}(5\times10^{-8})$ to identify salient variants, and shared genes are mapped to pathways via the Hallmark gene sets. Separately, a cohort of 126 million patients, including 18 000 with pemphigus is analysed to estimate the subsequent risk of 74 autoimmune diseases; for 26 conditions with adequate numbers, retrospective case-control analyses provide odds and hazard ratios across ethnic groups. Comorbidity in rare and common autoimmune diseases is further assessed in both the UKB and TNX using ICD-10 codes, focusing on White individuals (71,069,654 in TNX and 502,371 in UKB). Odds ratios for 15 autoimmune diseases are calculated, and cross-cohort comparisons assess reproducibility and highlight differences.

Results: Study find that only comparably prevalent autoimmune diseases are covered by the UKB and at the same time assessed by both GWAS and PGS catalogs. These are systemic (systemic lupus erythematosus) as well as organ specific, affecting the gastrointestinal tract (inflammatory bowel disease as well as specifically Crohn's disease and ulcerative colitis), joints (juvenile idiopathic arthritis, psoriatic arthritis, rheumatoid arthritis, ankylosing spondylitis), glands (Sjögren syndrome), the nervous system (multiple sclerosis), and the skin (vitiligo). Consistent sex-based differences are observed, including a predominance of women in diseases like multiple sclerosis and rheumatoid arthritis, and higher male prevalence in ankylosing spondylitis in UKB data. Pairwise analyses of autoimmune disease risk scores show shared and distinct genetic patterns, suggesting comorbidity from partial genetic overlap rather than complete overlap, also supporting poly-autoimmunity and the complex nature of genetic risk. Variant-level analyses identified both novel gene associations and genes previously reported in the literature as being linked to autoimmune diseases. Pathway analysis identifying systemic immune dysregulation in some diseases and localized effects in others. Ancestry related analysis reveals highly significant and generalizable associations between pemphigus and pemphigoid diseases, discoid lupus erythematosus, lichen planus, and undifferentiated connective tissue disease, among others. UKB and TNX resources highlight the impact of database-specific factors like healthcare delivery context, environmental exposures, and data granularity. Comparisons between TNX and UKB also replicated co-occurrence in certain disease pairings. It emphasizes on standardization is crucial for interpreting epidemiological signals and understanding autoimmune disease aetiology and comorbidity.

Abstrakt

Einleitung: Autoimmunerkrankungen beruhen grundsätzlich darauf, dass Autoantigene körpereigenes Gewebe schädigen. Dennoch zeigen sie eine große phänotypische Vielfalt, da genetische ebenso wie Umweltfaktoren in unterschiedlichem Ausmaß zu ihrer Ätiologie beitragen. Zahlreiche genomweite Assoziationsstudien (GWAS) haben damit verbundene Genomloci und Varianten identifiziert; deren Ergebnisse werden zunehmend für Polygenische Scores (PGS) genutzt, um das Krankheitsrisiko vorherzusagen. Aktuelle, öffentlich verfügbare GWAS- und PGS-Daten zu Autoimmunerkrankungen werden mithilfe von Informationen aus dem GWAS Catalog und dem PGS Catalog ausgewertet, die die untersuchten Krankheitsbilder, die einzelnen Studien und deren Befunde zusammenfassen. Zudem wird die UK Biobank (UKB), ein akademisches Register mit genetischen und phänotypischen Daten zu Autoimmunerkrankungen, herangezogen. Die Studie quantifiziert darüber hinaus die diagnostische Überlappung von Autoimmunerkrankungen in der UKB anhand von Phänotypdaten, nutzt PGS zur Clusterbildung von Personen und zur Bewertung der Heterogenität genetischer Risikoprofile und untersucht genetische Varianten in der Gruppe nicht verwandter weißer Britinnen und Briten auf Grundlage funktioneller Annotationen. Anschließend werden Daten zu häufigen, seltenen und hochwirksamen Varianten integriert, um Gene zu identifizieren, die mehreren Autoimmunerkrankungen gemeinsam sind, und diese relevanten biologischen Signalwegen zuzuordnen. Ebenso wird untersucht, wie häufig seltene und häufige Autoimmunerkrankungen gemeinsam auftreten; hierzu werden Komorbidität und gleichzeitiges Vorkommen durch Analyse von Patientendaten sowohl der UKB als auch des globalen Gesundheits forschungsnetzwerks TriNetX (TNX) quantifiziert. Schließlich wird das TNX-Datenset allein genutzt, um zu ermitteln, wie die seltene Autoimmunschleimhauterkrankung Pemphigus mit anderen Krankheitsbildern in verschiedenen Bevölkerungsgruppen zusammenhängt und um populationsspezifische Einblicke in ihr klinisches Profil zu gewinnen.

Methoden: Phänotypische Daten aller 502,371 UK-Biobank-Teilnehmenden bilden die Grundlage der Überlappungsanalyse, während ein erweiterter Teildatensatz von 104,544 Personen, die sechs Autoimmunerkrankungen abdecken, zur Untersuchung der Risikoscore-Verteilungen herangezogen wird. Genetische Analysen setzen eine Signifikanzschwelle von $-\log_{10}(5 \times 10^{-8})$ an, um relevante Varianten zu identifizieren, und gemeinsame Gene werden mithilfe der Hallmark-Gensätze Signalwegen zugeordnet. Unabhängig davon wird eine Kohorte von 126 Millionen Patienten, darunter 18,000 mit Pemphigus, analysiert, um das nachfolgende Risiko von 74 Autoimmunerkrankungen abzuschätzen; für 26 Erkrankungen mit ausreichender Fallzahl liefern retrospektive Fall-Kontroll-Analysen Odds Ratios und Hazard Ratios für unterschiedliche ethnische Gruppen. Die Komorbidität seltener und häufiger Autoimmunerkrankun-

gen wird außerdem sowohl in der UKB als auch in TNX anhand von ICD-10-Codes untersucht, wobei der Schwerpunkt auf weißen Personen liegt (71,069,654 in TNX und 502,371 in UKB). Odds Ratios für 15 Autoimmunerkrankungen werden berechnet; Kohortenübergreifende Vergleiche prüfen die Reproduzierbarkeit und heben Unterschiede hervor.

Ergebnisse: Die Studie stellt fest, dass nur relativ häufige Autoimmunerkrankungen sowohl in der UKB erfasst als auch gleichzeitig in den GWAS- und PGS-Katalogen bewertet werden. Dazu zählen systemische Erkrankungen (systemischer Lupus erythematodes) sowie organspezifische Formen, die den Gastrointestinaltrakt (entzündliche Darmerkrankungen, insbesondere Morbus Crohn und Colitis ulcerosa), die Gelenke (juvenile idiopathische Arthritis, Psoriasis-Arthritis, rheumatoide Arthritis, ankylosierende Spondylitis), Drüsen (Sjögren-Syndrom), das Nervensystem (Multiple Sklerose) und die Haut (Vitiligo) betreffen. Es zeigen sich konsistente geschlechtsspezifische Unterschiede, darunter eine Überrepräsentation von Frauen bei Erkrankungen wie Multipler Sklerose und rheumatoider Arthritis sowie eine höhere Prävalenz von Männern bei ankylosierender Spondylitis in den UKB-Daten. Paarweise Analysen von Autoimmunerkrankungs-Risikowerten zeigen sowohl gemeinsame als auch unterschiedliche genetische Muster, was auf Komorbiditäten aufgrund teilweiser statt vollständiger genetischer Überlappung hinweist und damit Polyautoimmunität sowie die komplexe Natur des genetischen Risikos unterstützt. Variantenbasierte Analysen identifizierten sowohl neue Genassoziationen als auch Gene, die bereits in der Literatur mit Autoimmunerkrankungen in Verbindung gebracht wurden. Pfadanalysen zeigten eine systemische Immun dysregulation bei einigen Erkrankungen und lokalisierte Effekte bei anderen. Abstammungsbezogene Analysen legen hochsignifikante und generalisierbare Assoziationen unter anderem zwischen Pemphigus- und Pemphigoiderkrankungen, diskoidem Lupus erythematodes, Lichen planus und undifferenzierter Bindegewebserkrankung offen. Die Ressourcen von UKB und TNX unterstreichen den Einfluss datenbankspezifischer Faktoren wie des Versorgungskontextes, von Umweltexpositionen und der Datengranularität. Vergleiche zwischen TNX und UKB reproduzierten außerdem das gemeinsame Auftreten bestimmter Krankheitskombinationen. Es wird betont, dass Standardisierung entscheidend ist, um epidemiologische Signale zu interpretieren und die Ätiologie sowie Komorbidität von Autoimmunerkrankungen zu verstehen.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview of Autoimmune Diseases

The concept of autoimmune diseases (AIDs) has undergone substantial evolution over time. In the mid-20th century, the scientific community gradually acknowledged their existence, dispelling earlier skepticism regarding the body's capacity to harm itself. Initially, autoimmunity is regarded as biologically implausible, with Ehrlich famously coining the term "horror autotoxicus" [1]. However, ground-breaking discoveries during this period, such as the identification of the systemic lupus erythematosus (SLE) cell and rheumatoid factor, paved the way for the acceptance of AIDs by the 1960s [2]. The recognition of antiphospholipid syndrome in 1983 marked a pivotal milestone in AID research [3]. Several studies have demonstrated that AIDs exhibit a heritable nature, with genetic factors accounting for approximately 42%-91% of the variance in occurrence, particularly in paediatric cases [4]. Among the more prevalent and impactful autoimmune conditions are rheumatoid arthritis (RA), SLE, type 1 diabetes (T1D), multiple sclerosis (MS), and psoriasis (PSO) [5]. These conditions exhibit a genetic component, which can significantly impact the quality of life. These conditions often cluster within families, suggesting a hereditary influence [6]. Autoantibodies serve as crucial diagnostic markers for various autoimmune disorders [7].

RA is a chronic, systemic inflammatory AID primarily affecting joints, but can also impact other organs [8] [9]. Its historical development has been challenging to trace, with the first clear description attributed to Augustin Jacob Landré-Beauvais in 1800 [10]. Similarly, as a chronic AID, SLE is characterized by multiorgan involvement and the production of autoantibodies against nuclear and cytoplasmic antigens [11]. The term "lupus" was first used in the middle ages, with the modern understanding of SLE emerging in the 19th century [12]. The historical development of MS spans centuries, with significant advancements occurring in the 19th and 20th centuries. Jean-Martin Charcot's work in 1868 established MS as a distinct neurological disease [13].

A chronic inflammatory skin disease, PSO was long confused with leprosy until the 1800s when it was finally recognized as a distinct condition. There is a rare autoimmune blistering disease known as pemphigus characterized by intraepidermal cleavage and loss of adhesion between cells affecting the skin and mucous membranes [14]; its research saw significant milestones in the 1960s [15]. Crohn's disease (CD) has a long history, with early descriptions dating back to the 17th century [16]. Ulcerative colitis (UC) was first described by Karl Rokytansky in 1842 [17]. CED an autoimmune enteropathy triggered by gluten consumption in genetically predisposed individuals [18] [19]; it has a long history, with possible origins dating back to ancient times [20].

AIDs can be diagnosed and comprehended through the identification of autoantibodies [21]. The activation of autoreactive T and B cells can lead to inflammation and tissue damage [22]. In addition to autoantibodies, B cells and plasma cells play a substantial role in AIDs through the regulation of inflammation [23]. Immune diseases are the result of intricate interactions between B and T cells. B cells contribute to autoimmunity by producing autoantibodies, presenting autoantigens, secreting inflammatory cytokines, and forming ectopic germinal centres [24]. B cells receive assistance from T cells, resulting in the production of pathogenic autoantibodies in certain diseases [25]. Inflammatory cytokines and dysregulation of T-helper lymphocytes are pivotal factors in the pathogenesis, along with intricate cellular and molecular mechanisms [26]. Targeting B cells may be a potential therapeutic approach for certain autoimmune disorders, but it can also exacerbate the condition in some patients [23]. Consequently, comprehending these mechanisms is essential for developing more effective therapeutic strategies for AIDs [27]. Recent studies suggest that novel regulatory subsets, such as Treg17 and Breg cells, may play a protective role in preventing AIDs [28].

SLE is diagnosed based on clinical findings and laboratory tests, often including autoantibodies against nuclear components [29]. MS diagnosis relies on demonstrating dissemination in space and time of central nervous system lesions, guided by the McDonald (a set of diagnostic guidelines) criteria [30]. In PSO, diagnosis is typically clinical, sometimes supported by skin biopsy [31]. Pemphigus diagnosis depends on clinical symptoms, histology, and immunochemical testing [32]. T1D is identified by hyperglycemia and the need for insulin replacement [33]. CD and UC employ endoscopic, histological, and serologic findings [34], while celiac disease (CED) involves serological testing, genetic screening, and intestinal biopsy [19].

RA progresses through several phases, beginning with genetic and environmental risk factors, followed by asymptomatic autoimmunity, and culminating in clinically apparent disease [35]. MS is characterized by inflammation, demyelination, and neurodegeneration [36]. PSO involves a hyperproliferation of keratinocytes and immune cell infiltration, creating distinctive skin lesions [37]. Pemphigus comprises a group of rare

autoimmune skin blistering diseases associated with a significantly elevated mortality risk [38] [39], featuring autoantibodies targeting desmosomal proteins [40] [41]. T1D is a chronic AID characterised by the destruction of pancreatic beta cells, leading to insulin deficiency [33] [42]. In autoimmunity-induced gastrointestinal diseases, T-cell mediated inflammation as well as regulatory T-cell dysfunction play essential roles [43], with CD presenting transmural inflammation [44] and UC primarily affecting the colon and rectum [45]. CED consists of a heterogeneous clinical presentation, ranging from classical gastrointestinal symptoms to atypical extraintestinal manifestations [46].

Common symptoms associated with autoimmune disorders include fatigue, joint pain, skin manifestations, and organ-specific dysfunction [47]. Additionally, fever, rashes, ulcers, and muscle weakness may be present. RA is characterized by symmetric polyarthritis, typically involving small joints of the hands and feet [48], stiffness, fatigue, and weight loss [49], and may also affect the skin, eyes, heart, lungs, and blood vessels [50]. SLE consists of periods of remission and relapse, with symptoms ranging from mild rashes to severe organ damage [11]. MS includes progressive, relapsing, and relapsing-remitting phenotypes [51]. PSO is characterized by reddish, scaly plaques on the skin surface [52]. Pemphigus vulgaris (PV) and pemphigus foliaceus (PF) involve blister formation due to autoantibodies against desmogleins [41]. T1D typically presents with polyuria, polydipsia, and weight loss [33]. CD and UC usually includes abdominal pain, diarrhea, and weight loss [34] [45]. CED exhibits a heterogeneous range of gastrointestinal and extraintestinal manifestations [46].

RA management includes lifestyle modifications and early disease-modifying antirheumatic drugs (DMARDs) [53]. SLE therapy often involves antimalarials, corticosteroids, and immunosuppressive agents [29], with rituximab showing promise [54]. MS offers a wide variety of disease-modifying therapies to reduce progression and relapses [55]. PSO is treated with topical agents, phototherapy, systemic medications, and biologics targeting specific inflammatory pathways [31] [56]. In pemphigus, broad-spectrum immunosuppression is being replaced by targeted therapies such as rituximab [32] [57]. T1D requires insulin replacement and careful blood glucose control [33] [42]. CD and UC are managed with 5-aminosalicylates, corticosteroids, biologic therapies, small-molecule inhibitors, and surgery if necessary [58] [59] [60]. A lifelong gluten-free diet is the only effective treatment for CED in most patients [61]. A complex interaction between genetics and environment causes immune dysregulation and synovial inflammation in RA [62], while SLE pathogenesis involves genetic, epigenetic, and environmental factors [63]. In MS, genetic and environmental factors (including vitamin D deficiency and Epstein-Barr virus infection) contribute to demyelination [64]. PSO pathogenesis involves complex interactions between innate and adaptive immune systems, primarily through the IL-23/Th17 pathway [65]. In pemphigus, IgG autoantibodies target desmosomal proteins [41]. Approximately 50% of genetic susceptibility to

T1D is attributed to the Human leukocyte antigen (HLA) region on chromosome 6p21 [66]. CD and UC are thought to result from multifactorial etiologies including genetic susceptibility, environmental influences, alterations in gut microbiota, and dysregulated immune responses [67] [68]. CED depends on HLA-DQ2 or HLA-DQ8, alongside gluten exposure [69].

1.2 Integrative Overview on Genetic Variants

Genomic variations are differences in the DNA sequence found at various locations within a single genome. Single nucleotide variants (SNVs) differ in a single base on a single strand, insertions or deletions of up to 50 bases, and structural variants are genomic alterations over 50 bases. Variants are common genetic variations in humans [70], occurring about once every 300 base pairs [71]. Genotyping arrays interrogate common polymorphisms, with each variant defined in relation to a standard reference genome. Variants can occur in gene regions like promoters, exons, introns, and untranslated regions.

Small DNA sequence variations can significantly influence drug responses and disease risk [72]. Variants are predicted to account for 90% of genetic variations [73]. In disease susceptibility, severity, and treatment outcomes, variants may affect gene expression and protein function [74]. They disrupt transcription factor binding or enhancer activity, disrupting normal gene expression [75], also occurs in super-enhancer regions, which regulate cell-type-specific gene expression [76]. Variants can also regulate physiological and pathological processes like cellular senescence, apoptosis, inflammation, and immune responses [77]. Many variants are linked to AIDs risk due to disease heterogeneity [78]. People develop AIDs differently based on their variants. [78]. Disease-associated variants are often clustered in the genome [76].

Deciphering the multilayered interactions among genetic variants, regulatory elements, and other biological systems is essential for uncovering the molecular basis of AIDs and identifying potential therapeutic targets. Genetic profiles and non-genetic factors can improve risk prediction for certain patient subgroups, despite the poor predictive power of individual variants [79]. Common variants typically have modest effects on disease susceptibility, prompting interest in rare variants, which often exhibit larger effect sizes and may contribute significantly to disease heritability. Emerging research highlights the potential role of rare variants, particularly those in coding regions, in autoimmune conditions. Although some studies suggest that rare coding-region variants at known loci have a limited role in common autoimmune disease susceptibility [80], others argue that rare variants with larger phenotype effects remain underexplored contributors to the “missing heritability” [81] [82]. Despite notable progress in uncov-

ering the genetic architecture of AIDs, the contribution of rare variants remains poorly understood. Bridging this research gap is essential for a more complete understanding of genetic risk and for improving the precision of genetic prediction in autoimmune diseases. Human SNVs have been well characterized. The 1000 Genomes Project (1000G) provided one of the first comprehensive global references for human genetic variation. The 1000G based genetic variation with respect to the reference genome was overall 84.7 million SNVs, 3.6 million indels and $\sim$ 60,000 structural variants; each individual carried 4.1 million to 5.0 million sites that differed from the reference genome.

A visual summary of how genetic alterations affect molecular and cellular processes is presented in Figure$\sim$ 1.1. The figure categorizes variants by their genomic locations such as regulatory regions, UTRs, coding sequences, splice sites, and intergenic regions and illustrates their functional consequences. These include synonymous, missense, frameshift, stop-gained, and splice-site variants, as well as transcript-level effects like transcript ablation and nonsense-mediated decay. Together, these insights highlight how genomic variation can disrupt gene expression, RNA processing, and protein function, contributing to disease development.

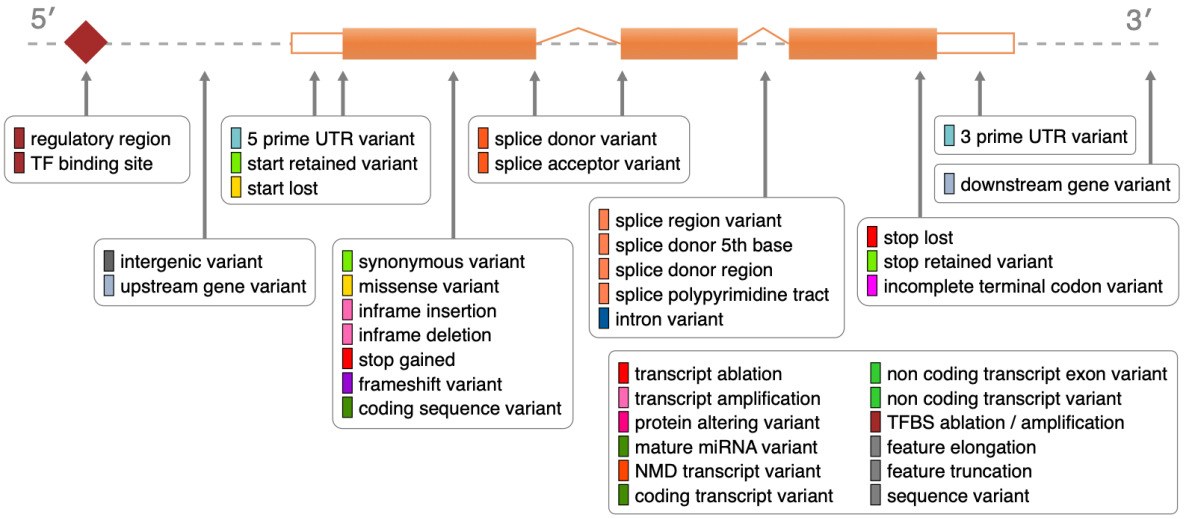


Figure 1.1: Figure provides a schematic representation of the distribution and classification of genetic variants relative to gene structure, oriented from the $5'$ to the $3'$ end. Variants are grouped by genomic context (e.g., regulatory, coding, intronic, UTR, and intergenic regions) and functional consequence (e.g., splice site changes, coding sequence alterations, transcript level effects). Color coding distinguishes each variant type, providing an overview of the potential impact of genetic variation. *Source:https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html*

1.2.1 Importance of Genetic Factors in Autoimmune Diseases

Autoimmune diseases (AIDs) exhibit a multifactorial etiology, with genetic factors playing a crucial role in their development and influencing individual susceptibility. Genetic

pathways underlie AIDs, explaining their comorbidity [83]. Different populations exhibit varying AID risk allele prevalence due to natural selection, particularly pathogen-driven selection [83] [84]. High-throughput techniques like next-generation sequencing aim to elucidate susceptibility, despite identifying only 15% of genetic factors [85]. Several AIDs are associated with major histocompatibility complex loci, with genes being the strongest genetic factor in most [86]. In addition to genetic predisposition, AIDs can be associated with specific major histocompatibility complex (MHC) class II genes. Genes from the HLA class II region (HLA-DR and HLA-DQ) are associated with autoimmune blistering diseases [87] and multiple AIDs, while specific genes like HLA-DRB1*0701 can be protective [88]. This genetic factor varies among individuals and is influenced by ethnicity [88]. PTPN22 (rs2476601), a non-HLA variants, alters T cell and B cell signalling and is shared by multiple AIDs like RA, T1D, and SLE [89]. Genetic loci linked to AIDs include 47/107 (44%) immune-mediated disease risk variants associated with multiple diseases, as identified by Cotsapas et al. [90]. Later work found 244 shared disease loci [91].

Over 90% of genetic variants linked to autoimmune diseases reside in noncoding regions, which complicates efforts to determine their functional impact [92]. These variants predominantly modulate gene regulation by altering transcription factor binding sites, histone modification patterns, and chromatin accessibility [92]. Genetic profiles are used to unravel shared disease mechanisms. Recent genetic studies have advanced our understanding of AIDs, with several loci shared across multiple disorders suggesting immune tolerance loss mediated by similar pathways [85] [93]. Both coding and noncoding variants contribute to Inflammatory Bowel Disease (IBD), and numerous genetic markers have been linked to systemic sclerosis [94]. Targeted therapies and personalised medicine approaches depend on understanding the genetic basis of AIDs. Research continues on genetic variations and integrating genetic findings into clinical practice, but challenges remain in characterising the underlying components of these diseases.

1.3 High-Throughput Genomic Data Generation in Autoimmune Disease Research

To advance the understanding of how genetic variants contribute to autoimmune diseases, it is essential to consider the foundational methodologies through which genomic data are acquired. This section outlines two principal strategies GWAS and sequencing-based approaches, that enable the systematic identification of genetic risk factors across the human genome

1.3.1 Applications of Genome Wide Association Studies in Genetic Research

Genome-wide association studies (GWAS) are a robust research methodology employed to discern genetic variants associated with distinct traits, diseases, or conditions across the entirety of the human genome. In GWAS for a binary trait, allele frequencies or genotypes are compared between cases and controls to identify genetic variants linked to the condition [95]. This process typically involves genotyping hundreds of thousands of single-nucleotide polymorphisms (SNP) in large cohorts [96], which has led to the discovery of numerous disease-associated loci and deepened our understanding of the genetic contributions to human diseases [97]. Although a range of genetic variants can be tested, GWAS have historically focused almost exclusively on SNPs [98]. Despite the success of this approach in highlighting numerous important loci, challenges remain in interpreting results and pinpointing causal variants [99]. Since the first GWAS released about twenty years ago [100], over 5,700 analyses have been conducted, yielding over 3,300 traits associated with genetic variants [96]. The process involves genotyping a large cohort, rigorous quality control to filter out low-quality samples or variants, imputation to infer missing genotypes, and regression-based association testing (e.g., logistic/linear regression) while adjusting for covariates like age, sex, and ancestry, depending on the study design and setting. In the case of a continuous phenotype or trait, a linear regression model is most commonly used, whereas logistic regression is mostly applied for dichotomous ones. Typically, the models are estimated for each single variant separately. A generalized representation of logistic regression [101] in GWAS is expressed as

$$\operatorname{logit}\big(P(Y=1)\big) = X\beta + G\gamma + \omega \tag{1.1}$$

where P(Y=1) is the probability of being a case, X is a matrix of non-genetic covariates (e.g., age, sex, ancestry). G represents the matrix of genotype values (such as SNPs), and γ is the vector of SNP/variant effect sizes. The term ω accounts for residual errors. In genetic association studies, the beta coefficient (β) represents the effect size of an allele on a trait, indicating the change in the trait value for each additional copy of the alternative allele. A positive (β) means the allele is associated with an increase in the trait, while a negative (β) implies a decrease. The standard error measures the precision of the (β) estimate; a smaller standard error denotes higher confidence in the estimated effect size. Together, (β) and standard error are used to compute p-values and construct confidence intervals, which help determine the statistical significance of the association. This model enables simultaneous estimation of genetic and non-genetic contributions to disease risk and is typically applied variant-by-variant across the genome. The typical GWAS output comprises, for each variant, a report giving the

ID of the variant according to database for single nucleotide polymorphisms (dbSNP) [102], the effect allele, the statistical effect and the corresponding p-value. Genome-wide studies typically have small p-values, so $-\log 10(p)$ gives a more intuitive number. A higher $-\log 10(p)$ indicates a stronger association.

Since GWAS test a large number of genetic variants at the same time, the statistical significance threshold has to be corrected to avoid false positive results. The widely used approach for this aim is the Bonferroni correction [96] [103], consisting of dividing the overall statistical significance threshold by the total number of independent tests, in this case, the tested independent variants. As a consequence, a threshold of 5×10^{-8} is commonly used in practice to determine genome-wide significance. This stringent cut-off accounts for the approximately one million common, independent variants [96] in the human genome, helping to reduce false positives and distinguish true genetic associations from random noise.

Interpretation of GWAS results focuses on effect sizes, linkage disequilibrium patterns (non-random association of alleles at different loci within a population), and biological context through gene annotation and pathway analysis. Post-GWAS analyses like meta-analysis, epistasis testing, and pathway analysis prioritise results. GWAS applications include insights into disease biology, heritability estimation, and clinical risk prediction [96]. Challenges include the need for large sample sizes, small effect sizes, and determining causal relationships. Multiple testing correction is essential to control for false positives when testing a large number of variants, with p-value thresholds typically ranging from 10^{-5} to 10^{-9} depending on the variants panel used [104]. Another key limitation of GWAS is that the majority of studies have been conducted in populations of European ancestry, which limits the generalization of findings to other ethnic groups [105]. Despite challenges, GWAS remain valuable for exploring complex trait and disease genetics [106].

Traditional GWAS methods analyses variants one at a time, leading to low predictive power and high significance thresholds. Using a multiple generalized linear model, Buzdugan et al., analysed all variants simultaneously while eliminating spurious correlations [107]. Shi et al., has combined false discovery rate control and LASSO regression (Used for handling numerous correlated or high-dimensional features, aiming to regularize the model and select an informative subset) into a two-stage approach that reduces false positives while maintaining power [108]. Another approach is pathway based using a truncated product statistic and a weighted Kolmogorov-Smirnov test developed by Weng et al., [109]. It considers multiple variants within genes and multiple genes within pathways.

1.3.2 Introduction and Bioinformatics Frameworks for Whole Genome Sequencing and Whole Exome Sequencing

Complementing GWAS, to investigate the genetic underpinnings of AIDs, advanced genomic techniques such as whole exome sequencing (WES) and whole genome sequencing (WGS) can be used. WES comprehensively analyses all coding regions of known genes, encompassing over 95% of exons, which harbour 85% of disease-causing mutations [110]. It presents a cost-effective alternative to whole genome sequencing, generating approximately 25,000 variants per individual [111]. WES possesses a diagnostic success rate of approximately 25% for rare diseases [112]. Through the selective capture and sequencing of exons, splice sites, and untranslated regions (UTRs) utilising hybridization-based probes (e.g., Agilent SureSelect) and high-throughput platforms such as Illumina, WES identifies coding mutations, including SNVs, small insertions/deletions (indels), and splice-site alterations. WES focuses on sequencing protein-coding regions with high mutation rates, making it cost-effective for identifying clinically relevant variants. In contrast, WGS provides a comprehensive view of the entire genome, including non-coding regions with regulatory elements and structural variations affecting gene expression.

WGS provides a comprehensive analysis of the entire genome, including non-coding regions that might regulate gene expression [113]. WGS provides an in-depth view of all genetic variations, including SNPs, insertions, deletions, and structural rearrangements, which can influence biological function. Recent genomic advancements have revolutionised our understanding of AIDs. WGS and GWAS have identified numerous genetic risk variants, many outside protein-coding regions [114]. These studies have expanded the number of known AID-associated loci from 15 to 68 [115]. While rare variants have been implicated in some cases, their role remains unclear [116]. Long non-coding RNAs co-localize with enhancers and disease-associated variants, suggesting a potential regulatory function [114]. CNVs may also contribute to disease susceptibility [117]. Despite these advances, challenges remain in identifying causal alleles and understanding their functional impacts [118]. A broader spectrum of genetic alterations can be captured by integrating findings from WGS and WES, thoroughly investigating both coding and non-coding contributors to AID risk. Initial milestone toward establishing a comprehensive global reference of human genetic variation, whole genome sequencing was conducted on DNA data from 2,504 individuals and 26 populations [119]. The Human Genome Diversity Project expanded global representation of variants through whole-genome sequencing of 929 individuals [120]. Among the numerous large-scale, often nationally coordinated genome sequencing initiatives, gnomAD (71,702 whole genomes sequenced) [121] and TOPMed (53,831 whole genomes sequenced) [122] are particularly notable for their extensive sample

sizes and contributions to human genetic diversity research.
The bioinformatics and statistical pipeline for WGS and WES begins with raw data quality control to remove low-quality reads and adapter contamination, followed by aligning the cleaned reads to a reference genome using tools like Burrows-Wheeler Aligner (BWA). Post-alignment, duplicate reads are marked to avoid amplification biases, and variant calling is performed to detect SNVs, indels, and structural alterations. Subsequent filtering steps refine the variant list based on quality metrics, and functional annotation tools (e.g., Variant Effect Predictor (VEP)[123], ANNOtate VARiation (ANNOVAR)) provide insights into the potential impact of these variants. Finally, statistical analyses such as GWAS or burden tests integrate these findings with phenotype data, while visualization and reporting tools facilitate the interpretation of results all of which are crucial for moving from raw data to clinically and biologically meaningful insights.

1.4 Computational Approaches for Exploring Genetic Structure and Risk Prediction

Building on advances in large-scale genomic data generation through methods like GWAS, WGS, and WES, this section introduces the key computational approaches employed to analyze these datasets and derive meaningful biological insights.

1.4.1 Principal Component Analysis for Detecting Genetic Structure and Population Stratification

Principal component analysis (PCA) is a foundational tool in genetic research for uncovering and correcting hidden structure in high-dimensional data. PCA is employed to reduce the dimensionality of complex genetic datasets by transforming correlated variables into a smaller set of uncorrelated principal components. This approach preserves the most critical variance within the data, thereby facilitating the identification of population structure, delineation of subgroups, and detection of disease-associated genetic variants [124]. It's widely used in exploratory data analysis and predictive modelling across various fields [125]. PCA transforms high-dimensional data into a smaller set of uncorrelated variables called principal components, retaining as much variance as possible. PCA analyses the original dataset and identifies recurring patterns, creating artificial variables (components) that maximise explained variation. The components are ranked by variance explained, with the first accounting for the most variation. By focusing on the top few components, PCA significantly reduces data di-

mensionality while retaining important information. Based on covariance (a square grid summarizing all pairwise relationships among variables), variance (captured data spread along each principal component), eigenvalues (variance captured by each principal component), and matrix decomposition (eigen decomposition of the covariance matrix into orthogonal eigenvectors and their eigenvalues), PCA analyses and summarises complex datasets. It identifies patterns by computing the covariance matrix and determining variable correlations.

1.4.2 Polygenic Risk Scores for Quantifying Inherited Disease Susceptibility

It is important to explore how these genetic findings can be quantified to assess an individual's disease risk. While GWAS identify specific genetic variants, PGS combine the effects of multiple variants to provide a comprehensive estimate of genetic risk, offering complementary insights. Polygenic Scores (PGS) is also known as Polygenic Risk Scores (PRS), a statistical tool used to estimate an individual's genetic predisposition to a particular trait or disease based on their genetic profile [98]. It is useful for estimating genetic risk for diseases [126] [127]. By aggregating the effects of multiple genetic variants, it can predict disease susceptibility and drug response [128] [129]. GWAS scores combine multiple genetic variants [130] to provide a more comprehensive risk assessment. In addition to traditional risk factors and family history [127], PGSs can predict significant disease risk after combining individual small effect genetic variants [131]. The method uses genotype data and GWAS results [98].
PGS are based on the understanding that most complex traits and diseases have a polygenic genetic architecture, meaning they are influenced by many genetic variants, each with small effects [132] [133]. A PGS is calculated by analysing genotypes with GWAS data [98]. The score is derived from a weighted count of thousands of genetic risk variants present in an individual's DNA [134]. These risk variants and their corresponding weights are identified through large-scale GWAS [134]. Standard equation to calculate a weighted PGS of an individual *j* is:

$$\text{PGS}_j = \sum_{i}^{N} \beta_i * \text{dosage}_{ij} \tag{1.2}$$

where *N* is the number of variants in the score, βi is the effect size (or beta) of variant *i* and *dosageij* is the number of alleles, *i* in the genotype of individual *j*. PGS have become routinely applied across biomedical research due to their correlation with genetic liability, which is the single largest contributor to phenotype variation [98].
PGS currently explain only a fraction of trait variance, but they show promise in improving diagnosis accuracy and risk stratification for rheumatic diseases and various

diseases [126] [135] [136]. They can identify individuals at higher risk for conditions like coronary artery disease and MS [137] [136]. They are sometimes better at predicting susceptibility and severity of AIDs than traditional biomarkers [126] [129]. PGS are useful for predicting risk, selecting treatment, and identifying early symptoms of diseases like SLE and RA [138] [139]. Genetic information is increasingly being used to combine with clinical and demographic factors in PGS research [140].
PGS are more accurate among European populations, potentially aggravating health disparities [141]. Some limitations include insufficient evidence for non-European ancestry and difficulty translating scores to lifetime risk [105]. Consequently, high-risk individuals can be identified and treated early [127]. PGS will play a crucial role in stratified medicine as GWAS sample sizes increase [98].
However, it is restricted by conceptual constraints and limitations related to biological complexity [142]. There are challenges in demonstrating clinical utility and implementing genomic technologies in healthcare systems which prevent PGS from being widely adopted by clinical practitioners [128].
A new approach for predicting drug response based on PGS is under development, called PRS-pgx [143]. Adeyemo et al., pointed out that ethical implications and regulatory frameworks are taken into consideration as PGS move towards clinical implementation [144]. A PGS can be used to improve disease prevention, diagnoses, and treatment precision, regardless of current limitations [132] [140]. By integrating PGS data with PCA, the method uncovers shared and unique genetic architectures across conditions, offering an intuitive and accessible way to understand the genetic basis of complex traits and diseases.

1.5 Structured Genomic Databases for Research and Risk Interpretation

Following the exploration of statistical methodologies for the analysis of genetic variation, the subsequent section addresses structured genomic databases that underpin large-scale data integration, interpretation, and application. Curated repositories such as the GWAS and PGS catalogs serve a pivotal role in the systematic organization and standardization of genetic association data.

1.5.1 Comprehensive Outline of Genome wide association studies Catalog

The GWAS catalog [145] is a publicly available, manually curated resource, which contains published GWAS and association results and is developed by the NHGRI and EMBL-EBI [146]. Catalog data is provided for the latest reference genome version (GRCh38.p13) and variant database version (dbSNPBuild 154). Genetic associations with complex traits and diseases can be studied using the GWAS Catalog. It was established in 2008 [147] and maintained by NHGRI-EBI as a comprehensive database of genome-wide association studies results [148]. This database contains over 1,800 phenotypes and 28,000 SNPs [149] along with SNP-trait associations, summary statistics, and supporting metadata curated from thousands of publications [150]. Access options include web interfaces, APIs, and downloadable files [151]. GWAS Integrator, pandasGWAS, and Genome-Wide Repository of Associations Between SNPs and Phenotype (GRASP) have been developed to facilitate analysis and access of this data [152], which allow to explore genetic associations, identify candidate genes, and investigate potential pleiotropic effects [153]. GWAS have expanded to include epigenome-wide association studies (EWAS), with 1.7 million associations [154]. The Catalog now covers a wide range of traits and populations, making it easy to explore genetic variants and their functional implications [150]. Manual curation of published genome-wide association studies is the primary method of data collection, including study cohort details, SNP-disease association data, and trait information. Two levels of curation ensure accuracy, including careful information extraction [155] and validation. Structured metadata ensures interoperability and usability in the catalog, which now includes a new submission portal and validation tool for faster data release and inclusion of unpublished results [148].

1.5.2 Understanding the Structure and Utility of the Polygenic Score Catalog

The rapid growth and widespread application of PGS necessitate the establishment of standardised repositories to facilitate their accurate reporting, reproducibility, and broad utilisation in research and clinical settings. One such pivotal resource is the PGS Catalog [156] which provides access to previously published PGS as well as variants, alleles, weights, and curated metadata that are necessary for reproducibility and independent applications along with their systematic evaluation [156]. The PGS Catalog aims to address the notable heterogeneity in the implementation and reporting of PGS, which has been a barrier to their clinical application [157]. Providing standardized data and methodologies, it is a centralized resource for PGS. The PGS catalog

provides researchers with a centralized repository of PGS, facilitating easier access to existing scores for various traits and diseases. In the PGS Catalog, researchers can deposit and share their PGS, allowing reproducibility and comparison of PGSs [157]. This accessibility allows for more efficient replication studies and enables researchers to compare the effectiveness of different PGS across diverse populations. Additionally, the catalog encourages transparency and standardization in PGS reporting, enhancing the reliability and reproducibility of genetic research. Together with the Clinical Genome Resource (ClinGen) complex disease working group, the PGS Catalog has developed the Polygenic Risk Score Reporting Standards (PRS-RS), a comprehensive set of standards for reporting PGSs.

1.6 Standardizing Biomedical Data with the Experimental Factor Ontology

Following the establishment of the GWAS and PGS catalogs, it is essential to highlight how these resources standardize and integrate complex biomedical data. Such standardization is made possible through the use of ontologies. An ontology is a formalised vocabulary that structures domain knowledge and its relationships. In biomedical research, ontologies cover specialised domains like genetic susceptibility factors [158], experimental design [159], and statistical analysis [160]. They simplify biological concept descriptions and experimental variable annotations, enhancing data sharing, integration, and collaboration. The need for introducing and using ontologies like the Experimental Factor Ontology (EFO) curated by the European Bioinformatics Institute (EBI) arises from the complexity and heterogeneity of biomedical data. Without a standardized vocabulary, inconsistencies in terminology and classifications can create significant challenges for data integration, comparison, and meaningful analysis. EFO directly addresses these challenges by providing a structured framework for data standardization. EFO systematically organizes biomedical concepts and experimental variables into standardized terms and relationships. It integrates diverse biomedical metadata related to diseases and phenotypes into a unified repository [161]. The GWAS and PGS catalogs both utilize EFO to classify diseases accurately and consistently, enabling effective data integration and comparison across various datasets and studies. This consistency supports in identifying genetic associations and risk factors, ultimately accelerating discoveries related to therapeutic targets. To facilitate semantic mapping and ensure consistent annotation, EFO employs a combination of custom scripting, expert curation, and specialized tools like Zooma, developed by European Molecular Biology Laboratory (EMBL)-EBI [161]. Zooma systematically links disease codes and free-text conditions to standard-

ized ontology terms, including EFO, Disease Ontology, and Human Phenotype Ontology. Through controlled vocabularies, Zooma resolves inconsistencies in data annotations and enhances interoperability across large biomedical databases such as the GWAS Catalog and ArrayExpress. Using EFO, datasets achieve approximately 80% mapping coverage from various sources, effectively integrating clinical and genetic phenotype data [161],[162]. This high level of standardization is critical because without a unified ontology like EFO, data integration becomes challenging due to inconsistencies in terminology and classification. Misaligned datasets can lead to misinterpretations, overlooked correlations, hindered collaboration, and ultimately, delayed insights into disease mechanisms and potential treatments. Indeed, the integration of electronic health records in genetic research has significantly benefited from the standardized terminologies provided by EFO [163]. In this study, analysis focused on the autoimmune-disease branch of the Experimental Factor Ontology (EFO v3.42.0; https://bioportal.bioontology.org/ontologies/EFO), extracting all “is a” relationships that link parent and child terms (e.g., rheumatoid arthritis as a child of autoimmune disease). To ensure comprehensive coverage, additional disease terms are incorporated from complementary ontologies: the MONDO Disease Ontology (EBI-curated; https://bioportal.bioontology.org/ontologies/MONDO) , Orphanet’s rare-disease repository (INSERM; http://www.orpha.net), and the Disease Ontology (http://www.disease-ontology.org).

1.7 Network Propagation

Network propagation is a computational method that diffuses information through molecular networks, enabling integration of information about mutational status of a gene and prioritization of disease-associated genes [164] [165] without requiring direct mutations. By transmitting signals across protein-protein interaction networks [166], this approach amplifies biologically coherent patterns, reveals functionally cohesive subnetworks, and identifies perturbed pathways beyond the immediate mutation sites. Its effectiveness depends on appropriate parameter optimization and normalization techniques to mitigate topological biases [164], ultimately capturing how genomic alterations create ripple effects throughout interconnected biological systems [167].
Network propagation initializes gene scores on a molecular network’s nodes and uses Random Walk with Restart (RWR) to diffuse these signals across the graph structure. In this algorithm, also called personalized PageRank or insulated diffusion, a random walker traverses the network by either following connections defined by the adjacency matrix or jumping back to seed nodes with a fixed restart probability. For any connected network with a normalized adjacency matrix W whose eigenvalues remain bounded by

one, this diffusion process mathematically converges to a unique steady-state distribution, providing a robust computational foundation for identifying network-influenced genes.

$$p_s = (1-\alpha)\,(I-\alpha W)^{-1}\,p_0$$

W is the probability of stepping from one gene to another gene in one random walk move [168] and encode gene level significance as an initial score vector p0, then at each iteration the updated score vector p is a mixture of two terms: with probability α (the restart probability), the walker follows network edges via W, and with probability 1 $-\alpha$, it returns to the original vector p0 [169]. Through diffusion, coherent signals within subnetworks are amplified: genes closely linked to high scoring seeds accumulate stronger propagated scores, whereas isolated or spurious connections are dampened. This selective enhancement makes it possible to pinpoint candidate disease genes or functional modules that would evade detection by direct statistical ranking [169].
This approach leverages molecular networks to prioritize candidate genes and improve the interpretation of GWAS results [169]. Restricting propagation to a curated network of hallmark pathways further sharpens this strategy. Because diffusion is confined to well validated interactions, biological relevance is heightened, and noise is curtailed.
These advantages notwithstanding, network propagation is limited by several methodological challenges. Topology bias can skew prioritization when highly connected regions dominate diffusion, especially under inadequate normalization [164]. Degree bias similarly favors hubs often well studied genes at the expense of less connected but potentially critical nodes; eigenvector (vector that does not change direction when a certain transformation is applied) or degree aware corrections are therefore essential [170]. Propagation parameters, such as the restart probability α in random walk implementations, are highly context sensitive: poor choices either over smooth true signal or retain excessive noise, undermining reproducibility. Finally, when the seed set is minuscule or noisy, statistical robustness diminishes, making it difficult to disentangle genuine biological associations from artifacts of network structure or parameter selection. Rigorous normalization, systematic parameter tuning, and validation against orthogonal datasets are therefore indispensable for drawing reliable biological inferences.

1.8 Environmental Influences on Autoimmune Disease

While the genetic underpinnings of AIDs offer crucial insights, they represent only part of the equation. Environmental factors ranging from infectious triggers and dietary patterns to chemical exposures can act in concert with genetic susceptibility to influ-

ence disease onset and progression. Environmental factors significantly contribute to the development of AIDs, accounting for approximately 70% of cases. However, it is important to note that these factors do not apply uniformly across all AIDs [171]. Vitamin D deficiency, parasitic infections, Epstein-Barr virus, and UV exposure are among these factors. Vitamin D deficiency has been linked to MS and SLE, though its exact role in disease progression is unclear [172]. Chemicals like silica can trigger inflammation and autoantibody production [173]. Smoking and exogenous estrogens increase the risk of SLE [174]. Pesticides, mercury, trichloroethene, and smoking also cause AIDs [175]. Diet and birth mode may affect the gut microbiota composition in T1D, contributing to disease onset [176]. Environmental factors are strongly associated with RA development, including tobacco smoke and silica dust [177]. In addition to infections and mental stress, sleep deprivation, age, perinatal factors, gender, diet, and obesity are also associated with RA risk [178]. Epigenetic changes caused by these factors, such as DNA methylation, have been linked to diseases like SLE and RA [179]. Environmental factors influence the immune system through mechanisms like Toll-like receptor signalling, B-cell activation, and T-cell regulation [180]. Pollution and nutritional factors primarily cause thyroid AID [181]. MS is linked to altered immune responses and gut microbiota due to high saturated fat and salt consumption [182]. Nutritional interventions and balanced diets can improve clinical outcomes [174]. Supplements may reduce autoimmune condition incidence and improve symptoms [172]. These environmental influences and genetic predisposition provide insights into disease pathogenesis, potentially leading to improved prevention and early intervention strategies. AIDs arise from a complex interplay between genetic predispositions and environmental triggers, both contributing significantly to disease risk. Advanced genetic analysis, such as WGS and WES, is needed to identify specific genetic variants, bridging the gap between inherited susceptibility and environmental influences.

1.9 Genetic Data and Autoimmune Diseases Covered by the UK Biobank

Since both genetic and environmental factors play a role, the selection of the database is critical for capturing this complexity. This section provides an overview of the database utilised in the current study for phenotype analysis and to explore risk scores. The UK Biobank (UKB) [183] is a major collaborative research project established in 2002 [184]. Over 500,000 United Kingdom participants aged 40–69 years are involved in the UKB between 2006 and 2010 [185] [186]. It is the largest source of genomic and phenotype data for global academic health research [187]. The project's open-access policy allows researchers worldwide to utilize this valuable resource for health

research [188], with over 26,000 researchers currently using the data to study various health conditions [186].

In the UKB, genetic, environmental, and lifestyle factors are investigated in relation to disease development [189]. The resource integrates deep genetic and phenotype information, health, and lifestyle data, enabling to examine correlations between these factors and health outcomes [187]. UKB provides genome-wide genotype data on all participants, with around 96 million testable variants after imputation [187]. Biological samples, such as blood, urine, and saliva, are collected and processed using highly automated methods to ensure long-term integrity [189]. Electronic medical records and neuroimaging data are also included in the biobank's comprehensive infrastructure, enhancing its utility for diverse research projects.

UKB collects extensive information from participants through a combination of questionnaires, clinical assessments, and biological sampling [190]. Online and in-person assessments are used to gather detailed lifestyle, medical, and genetic information. Physical measurements such as height, weight, blood pressure, and lung function are taken during visits to assessment centres [191]. These samples are processed and stored for future analysis [189]. Moreover, participants consent to the use of their electronic health records, which include hospital admissions, primary care visits, and disease registries, allowing for long-term follow-up. UKB continues to enhance its data through repeat assessments, web-based questionnaires, and multi-modal imaging of 100,000 participants [192].

The biobank includes deep phenotyping data, encompassing biological measurements, lifestyle indicators, and imaging data, alongside the genome-wide genotype data for all participants [187]. Medical diagnoses in UKB follow the international statistical classification of diseases and related health problems from the World Health Organization (WHO). In addition to International Classification of Diseases- 10th Revision (ICD-10) codes derived from medical records, the UKB provides self-reported diagnoses, referred to by dedicated internal IDs (starting with "20002_"). These self-reported data offer more detailed accounts of an individual's medical history, lifestyle choices, and environmental exposures. This integration of genetic, phenotype, and environmental information makes the UKB a powerful platform for developing PGS and advancing precision medicine [187].

Genome-wide genotyping is conducted for all UKB participants using the UKB Axiom Array, which directly measured approximately 850,000 variants [187]. The genotyping data were subsequently imputed using the Haplotype Reference Consortium and UK10K + 1000 Genomes reference panels, increasing the variant count to approximately 96 million. These imputed Axiom array genotypes (approx 825,927 markers) were included in the first UKB genetic data release in 2018 [187]. In addition to genotyping, large-scale sequencing efforts have been carried out. WES has been com-

pleted for over 470,000 participants, identifying approximately 12 million protein-coding variants. A published dataset from 2021 includes exome data for 454,787 individuals, capturing around 2 million exonic SNVs [193]. WGS was completed for 500,000 participants by 2023, following an earlier release of 200,000 genomes in 2021 [194]. Sequencing was performed using Illumina NovaSeq 6000 instruments with paired-end reads on S4 flow cells. In the initial WGS data release, 150,119 genomes were sequenced [195], yielding an extensive variant profile: approximately 585 million SNVs (representing $\sim$ 7% of all possible human SNVs), 58.7 million insertions and deletions (indels), $\sim$2.5 million microsatellites, and $\sim$ 900,000 structural variants (SVs). Notably, 46% of SNVs were classified as "singletons," present in only one individual, while just 3.4% ($\sim$ 20 million) had a frequency exceeding 0.1% [195]. Among the many national-scale genome sequencing initiatives, the UKB stands out for both the scale and depth of its genomic data, providing one of the most comprehensive variant catalogs in population genetics research.

There are currently 61 UKB-approved projects specifically relating to AIDs (search for "autoimmune diseases", June 13th, 2022). Several genetic and phenotype datasets related to AIDs are publicly available in the UKB, making it a valuable resource for investigating the genetic basis and clinical features of these conditions.

1.10 Sex Differences and Demographic Disparities in Autoimmune Diseases

Globally, millions of individuals are affected by AIDs. Industrialised nations are witnessing a surge in the incidence of these diseases, which frequently manifest during midlife, typically between the ages of 40 and 50 [196] [197]. Approximately 20 million Americans are affected by autoimmune disorders, which exhibit a higher prevalence in women [196] [197]. The overall prevalence of AIDs stands at 4.5%, with 6.4% of females and 2.7% of males affected [197]. For instance, SLE, sjogren's syndrome, and primary biliary cirrhosis exhibit a female predominance of approximately 9:1, while RA and MS have a female predominance of 3:1, and IBD and T1D have a female predominance of almost 1:1 [198]. Approximately 1% of the global population is affected by RA [199]. The majority of women with SLE are of reproductive age [63]; in the United Kingdom, one in 2000 adult women is estimated to have SLE [200]. There are over 2 million people with MS worldwide, and it typically affects young adults, predominantly females between 20 and 40 years of age [201] [202].

Approximately 80 distinct disorders have been identified [203] that target specific organs or manifest broadly. Notably, women and certain ethnic groups are disproportionately affected by AIDs [204]. While sex differences reveal striking disparities in AID

susceptibility, ethnic disparities in AID prevalence, severity, and clinical manifestations are also evident. Ethnic variations in the incidence and prevalence of SLE have been observed in black populations [205]. Recent studies have also documented the endemic nature of AIDs. Almirall et al., reported that AIDs are commonly reported in Catalonia and a higher prevalence among women at 1.5% [206]. Researchers observed a significant increase in the age-standardised prevalence rate (ASPR) of RA between 1990 and 2019, while rates for MS, IBD, and PSO decreased [207]. Borchers et al., have demonstrated that SLE is less prevalent and typically less severe in European populations compared to other ethnicities. Notably, renal involvement is more prevalent in non-European patients [208]. Shapira et al., further emphasise that the geographical distribution of AIDs exhibits significant ethnic and geographic gradients, which are shaped by both genetic and environmental factors [209]. These findings underscore the intricate interplay in the development and distribution of AIDs among different sexes, racial, and ethnic groups. Pemphigus varies based on geography and ethnicity between 0.76 and 16.1 per million people per year, and it is more common in certain populations, notably Ashkenazi Jews [210]. A chronic inflammatory skin disease, PSO affects about 2-3% of the global population [31] [65]. T1D has an estimated 34.2 million individuals worldwide affected, with prevalence increasing over the past few years [211]. CD is most prevalent in North America and Europe [212], and UC incidence is increasing in developing countries [213]. CED affects about 1% of the western population [18] [19].

Evidence suggests a tendency for AIDs to co-occur, referred to as “multiple autoimmune syndrome (MAS)” [214] or “polyautoimmunity” [215]. Several mechanisms have been proposed to explain the increased risk of MAS, including genetic predisposition, environmental triggers such as the gut microbiome and obesity, and complex mediators [216] [217] [218] [219] [220] [221]. Pemphigus has been linked to Sjögren’s syndrome, SLE, PSO, and UC and other autoimmune diseases [222] [223] [224]. T1D may coincide with other endocrine AIDs in polyglandular autoimmune syndromes [33]. CD and UC can coexist with other autoimmune disorders [67]. CED shares similarities with other AIDs and shows an increased risk of autoimmune coexistence [225]. A nationwide case-control study conducted in Taiwan [222] identified an elevated risk of developing Sjögren’s syndrome, SLE, and PSO. Smaller studies conducted in Israel and the US [226] [227] also reported associations between autoimmune thyroid diseases and RA. A review of pemphigus epidemiology [228] corroborated these findings and highlighted a novel association with PSO [223] [229] and UC [224]. Limited evidence exists for a link between myasthenia gravis and paraneoplastic pemphigus [230] [231]. However, the generalizability of these associations across different ethnic groups remains untested.

1.11 Electronic Health Data Covered by the TriNetX

A robust data platform and real-world data are essential for uncovering the complex interrelationships underlying disease comorbidities. Capturing diverse patient profiles and clinical nuances is key to achieving a comprehensive understanding of these complexities. The TrinetX (TNX) [232] database supports this effort through a real-time, electronic, federated data network that aggregates de-identified patient data from multiple healthcare organizations, enabling extensive clinical research. It has been utilized in various studies to analyse patient outcomes, treatment efficacy, and disease associations across diverse medical conditions. The global health research platform collects and aggregates real-world health data from healthcare organizations, including hospitals, clinics, and research institutions. In contrast to traditional biobanks, TNX does not collect data directly from its participants. Electronic health records of participating Health Care Organizations (HCO) are the primary source of data for TNX [233]. In addition to demographics, diagnoses, procedures, medication, laboratory results, oncology-specific data, and genetic variants obtained from molecular diagnostic genomic analyses. Data may be acquired as discrete data elements or by using natural language processing to analyse medical reports and notes [233]. The data is first harmonized syntactically by being ingested into the TNX common data model. Custom connectors and toolkits are used to infuse data from various common data models [233]. TNX uses a federated ecosystem with hardware residing in HCO's data centers [233]. Continuously updated databases are ingested and harmonized using new data. The data remains within the healthcare institution's secure environment, and TNX uses a federated data model that allows researchers to query datasets without extracting individual patient information. Compliance with privacy regulations such as Health Insurance Portability and Accountability Act (HIPAA) and General Data Protection Regulation (GDPR) is ensured by this approach. By integrating large-scale, real-time clinical data, TNX enables researchers and pharmaceutical companies to conduct observational studies, identify patient cohorts for clinical trials, and analyse treatment effectiveness across diverse populations. The data quality of TNX is ensured by several measures. Missing data rates in most fields are less than 5%, with most fields checked for completeness. Data validity is enhanced by a comprehensive data quality program [234]. The process of data de-identification has been attested to through a formal determination by a qualified expert as defined in HIPAA Privacy Rule [233]. TNX is a federated data platform that facilitates collaboration between healthcare organizations, pharmaceutical companies, and contract research organizations for clinical research [235]. A total of 220 HCOs has been established in 30 countries by 2022, compared to 55 HCOs in 7 countries in 2017 [233]. In addition, the platform allows the design of clinical trials to be data-driven, thereby reducing accrual failure

and protocol amendments [235]. Palchuk et al., 2023, report that TNX facilitated over 19,000 sponsored clinical trial opportunities. TNX is used by researchers to study various medical conditions [236] (Trocchia et al., 2023) including AIDs. A growing number of organizations are participating in the network, as is the breadth and depth of data collected. Using this extensive network of real-world data, TNX can facilitate the design of clinical trials, feasibility assessments, and research on biomedical and clinical topics.

1.12 Current Gaps in the Literature

Despite substantial progress in AID research through GWAS and PGS, several critical gaps remain. Although numerous GWAS have identified variants associated with AIDs, existing GWAS and PGS tools are typically broad in scope, with limited analyses tailored to the distinct and overlapping genetic architectures of various autoimmune conditions.

Many studies fail to adopt an integrative approach that systematically prioritizes genetic variants by considering both allele frequency (AF) and functional annotations, such as those provided by tools like VEP. This gap results in an incomplete understanding of the genetic foundations of AIDs, especially regarding the roles of both common and rare variants [118]. While thousands of associated variants have been identified, most studies focus solely on these variants without linking them to biological pathways. This lack of context limits our understanding of disease mechanisms, and efforts to identify shared pathways across different autoimmune diseases are still uncommon.

Additionally, comorbidity research has primarily focused on common autoimmune diseases such as RA, SLE, and T1D, while rare conditions like pemphigus are often underrepresented. Pemphigus patients exhibit heightened susceptibility to various secondary autoimmune diseases. [237]. This has limited insights into their comorbidity profiles and clinical implications.

A major limitation of current research is the overrepresentation of individuals of European ancestry [129], which restricts the generalization of findings to other populations and poses a barrier to equitable implementation of genomic medicine. There is a notable lack of integrative, multi-source studies that assess ancestry-specific patterns in genetic risk and clinical manifestations of AIDs. Most current efforts remain confined to single datasets, perpetuating biases and leaving significant gaps in our understanding and prediction of autoimmune disease risk across populations.

1.13 Objective of Study

This study integrates GWAS, PGS, and data from the UKB and TNX to elucidate the genetic and comorbidity profiles of AIDs. The specific aims are as follows:
1. The scope of AIDs covered in genetic databases will be assessed, along with an evaluation of the current state of genetic research in autoimmunity.
2. Investigate the risk scores of paired autoimmune diseases in the UK Biobank to visualize and find clustering patterns, associated genetic variants, and shared genes for biological pathway analysis.
3. Ancestry related risk of comorbid autoimmune diseases in patients with pemphigus employing TNX data.
4. Validating findings and enhancing the understanding of autoimmune disease comorbidities through the utilization of UKB and TNX databases.

1.14 Study Background and Framework

The initial study systematically investigated the genetic and biomedical landscape of autoimmune diseases by integrating public databases such as biomedical ontologies, and other genetic data. Here, current publicly available autoimmune disease GWAS and PGS data are reviewed using information from the GWAS Catalog and PGS Catalog, respectively, with a summary of the diseases studied, the studies conducted, and their key results. Study also examines genetic data and autoimmune disease patient records in the UK Biobank. Public resources such as these hold great promise for enabling systems-based, genetics-driven approaches to autoimmune research. This comprehensive review appeared in Frontiers in Immunology (Vol. 13, Aug 5, 2022; DOI: 10.3389/fimmu.2022.972107).
The second part of this study centers on data obtained from the UKB and an additional study by Thompson et al.,[238] incorporating phenotype data (initially reported information), PGS data, and genetic variant data. The phenotype data consist of binary values, where 1 indicates the presence of disease and 0 indicates its absence. This dataset is also complemented by PGS data, which include both the standard and enhanced datasets released by the UKB. This part in second study builds upon the recent work of Thompson et al.(2024), which systematically evaluated the UKB data. Their study providing PGS for 28 diseases and 25 quantitative traits. Thompson et al. (2024), developed a standardized evaluation pipeline to benchmark these PGSs against 76 previously published scores. Two PGS datasets were generated: the standard PGS, calculated for all individuals in the UKB and trained on external data from the 100,000 Genomes Project; and the enhanced PGS, calculated for a testing sub-

group within the UKB (White British Unrelated subgroup), using both external data and additional training data from the UKB itself. The PGS algorithms were validated using data from both the UKB and the 100,000 Genomes Project (which aimed to catalogue human genetic variation). Their findings indicated that the UKB PGS generally outperformed earlier models, particularly in individuals of European ancestry, although reduced performance is observed in non-European populations. Additionally, variant data provided by Thompson et al. (2024), through Zenodo are incorporated as the third key dataset in this part of the study. This portion of the study has not yet been published.

In the third part of study, to quantify comorbidity risks associated with AIDs particularly focusing on rare conditions such as pemphigus, this analysis examines the influence of comorbidities on disease susceptibility. The primary objective is to analyze comorbidity patterns in order to better understand variability in disease susceptibility at the population level. To conduct this analysis, data from TNX is utilized. Because the most databases including UKB are dominated by European ancestry [129]. This part of study is extended to an ancestry stratified framework in pemphigus, enabling the evaluation of race and ethnicity specific comorbidity profiles. This part of the study is published in scientific reports on December 3, 2024 (Volume 14, 2024), and is accessible via DOI: https://doi.org/10.1038/s41598-024-78031-z

Last part of study focuses on investigating comorbidity patterns using a population-based biobank (UK Biobank) and a clinical cohort (TriNetX) employing phenotype data. Integrating and harmonizing these resources extends earlier race and ethnicity centred work from a single rare autoimmune disease to a broad spectrum of rare and common AIDs. The unified framework enables cross validation of disease associations, exposes discrepancies between population and clinical settings, and quantifies comorbidity risk within the White population. This portion of the study has not yet been published.

Chapter 2

Materials and Methods

2.1 Retrieving GWAS, PGS, and UKB Data Using EFO for Autoimmune Disease Research

This study begins with a comprehensive review of autoimmune disease information using three primary resources detailed above; (i) Genome wide association studies (GWAS) catalog [146] (ii) Polygenic score (PGS) catalog [156] (iii) UK Biobank (UKB) [187]. Experimental Factor Ontology (EFO) terms for autoimmune diseases are retrieved from the "autoimmune disease" branch of EFO (version 3.42.0), and its hierarchical structure is downloaded from the EFO website (https://bioportal.bioontology.org/ontologies/EFO). GWAS catalog source files of studies and associations used in survey (files gwas-catalog-studies_ontology-annota-ted.tsv and gwas-catalog-associations_ontology-annotated.tsv) is obtained from (http://ftp.ebi.-ac.uk/pub/databases/gwas/releases/2022/05/23/), and entries with "MAPPED_TRAIT-_URI" an autoimmune EFO IDs are extracted. The PGS catalog source files are obtained from (https://www.pgscatalog.org/downloads) and extracted from the respective files via EFO IDs all information related to AIDs. AID data from UKB are extracted using the mapping file embedded in Zooma, the ontology-mapping tool of the European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI) ontology lookup service [239]. This file originates from a large-scale effort that systematically links UKB Classification of Diseases (ICD)-10 codes and self-reported conditions to EFO terms (https://github.com/EBISPOT/EFO-UKB-mappings).

2.2 Framework for Analyzing Phenotype Data, Risk Scores, and Genetic Variants Utilizing UKB Data

2.2.1 Sample Overlap Analysis and Quantitative Assessment of Disease Associations Utilizing Phenotype Data

To ensure a comprehensive understanding of the study population and to investigate AID patterns, this subsection is designed to address two key research: (1) The sample overlap between different AID datasets, and (2) The strength of comorbid relationships among autoimmune diseases. To address these two, the study utilizes phenotype information retrieved from the initially first reported data in the UKB to classify and analyze AID samples. This data is retrieved from UKB 18th of March 2024. The GWAS used array and imputed genomic data (fields 22418 and 21008, respectively). The AIDs included in this analysis are similar to those utilized by Thomson et al. in their risk-score study and comprise Crohn's disease (CD), multiple sclerosis (MS), psoriasis (PSO), rheumatoid arthritis (RA), systemic lupus erythematosus (SLE), and ulcerative colitis (UC). Celiac disease (CED) and type 1 diabetes (T1D) were excluded owing to the unavailability of their initial reported data at the time. Further, this phenotype data is also employed to assess the relationships between different AIDs through odds ratio (OR) analysis. The Fisher's exact test in R is utilised to compute the p-value, simultaneously providing an estimate of the OR along with its corresponding confidence interval (CI). This phenotype can be retrieved from UKB using its UKB ID or ICD 10 code. The OR is a measure of the strength of association between an exposure and an outcome, representing the chance of the outcome occurring in an exposed group compared to an unexposed group. The formula for the OR is:

$$OR = \frac{\left(\frac{\text{Exposed with disease}}{\text{Not exposed with disease}}\right)}{\left(\frac{\text{Exposed without disease}}{\text{Not exposed without disease}}\right)} \tag{2.1}$$

An OR $>$ 1 indicates that the comorbidity and positive association between exposure and outcome, meaning the exposure increases the odds of the outcome occurring. OR$<$1 suggests that the comorbidity is less frequent and implying a potential protective effect. OR =1 means there is no significant association between exposure and outcome. Analysis of OR values enable identification of potential shared pathogenic mechanisms. A CI for an odds ratio provides a range of values that likely contain the true OR in the population. It is used to assess the precision and statistical significance of the estimated OR. The most commonly used is the 95% CI, which means that if the study were repeated multiple times, 95% of the computed intervals would contain the true OR. The calculation of CI includes exp refers to the exponential function, which

is the inverse of the natural logarithm. OR = odds ratio, ln(OR) = Natural logarithm of the odds ratio, Z = Z-score (for a 95% CI), SE = Standard error of ln(OR). The natural logarithm (ln) is used to make it more symmetric.

$$\text{CI} = \exp\left(\ln(\text{OR}) \pm Z * \text{SE}\right) \tag{2.2}$$

$$SE = \sqrt{\frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}} \tag{2.3}$$

The SE provides a measure of the variability or uncertainty in the estimate. A and B correspond to number of exposed individuals with and without disease and C and D correspond number of unexposed individuals with and without disease. The CI excludes 1, the OR is statistically significant. If the CI includes 1, the association may not be statistically significant. A narrow CI suggests more precise results, while a wide CI suggests greater uncertainty in the estimate.
The odds ratio is transformed into a log10 scale simplifies the representation and makes the resulting distribution more symmetric on the log scale. The number of male and female are calculated using initial first reported data.

2.2.2 Process for Analyzing Cohort Overlap, PGS Correlation, and Genetic Risk Score Distribution in Autoimmune Diseases Using PGS Datasets

Approach to Assessing Cohort Overlap and Polygenic Score Correlations Across Standard and Enhanced Datasets

Following phenotype classification, two pivotal questions are addressed to better understand shared genetic architecture and pleiotropy across AIDs: (1) sample overlap within and between the enhanced and standard PGS datasets, and (2) correlation of PGS for diseases and traits from both datasets. To investigate these two questions, the enhanced and standard datasets are analyzed using the first reported AID information from the UKB to classify individuals by disease. The sample overlap is examined to assess how many individuals are shared across diseases and datasets. This step establishes the degree of intersection between cohorts and forms the foundation for further stratification analyses.
Polygenic scores from the standard and enhanced datasets are correlated to explore pleiotropy, scores highlights groups of diseases that share a common genetic basis. The correlation analysis of risk scores originally conducted by Thompson et al. (2024),

is replicated, and the PGS values are standardized according to their methodology. Consequently, this study employs Pearson's correlation method, given the standardized nature of the PGS data. The resulting correlation matrices provide a compact, reproducible map of shared genetic architecture for downstream clustering. The Pearson's correlation is calculated using Hmisc package's rcorr function.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

x_i and y_i are the individual data points, $\bar{x}$ and $\bar{y}$ are the means of the x and y variables, n is the number of paired observations. The matrix of p-values corresponding to tests of the hypothesis that each correlation is zero. The p-value is typically calculated using the t-distribution, where the test statistic is computed as

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

and the p-value is determined from the t-distribution with $n - 2$ degrees of freedom. The n is the matrix of the number of observations used for each pairwise correlation (this accounts for any missing values). The r is the correlation matrix, with each entry computed using the Pearson formula shown above. The correlation study is done using PGS from both the enhanced and standard datasets.

Strategy to Genetic Risk Profiling using Enhanced Dataset

Comparative evaluations are also conducted, including two-dimensional PGS comparisons assessing relationships between two quantitative PGS variables and visualizing group separation based on PGS encourages these comparative evaluations. The analysis is focused on four AIDs such as MS, PSO, RA, and SLE using data from the enhanced dataset. T1D and CED are excluded as first reported data for these conditions has not been retrieved from UKB. To avoid the masking of underlying patterns due to an excessive number of data points, the enhanced dataset is selected for this study because of its relatively smaller sample size. The scatter plot is generated using the ggplot2 package in R (version 4.4.1) to facilitate clear visual representation of the relationship between PGS values and disease classification. All participants from the enhanced dataset are included in the analysis, except for individuals diagnosed with two AIDs where the overlapping sample size is fewer than 30; these cases are excluded to maintain analytical robustness. Information related to PGS has been retrieved on 6th of February 2024.

2.2.3 Process for Identifying Variants in the Unrelated White British UK Biobank Cohort and Mapping Associated Genes to Pathways

Approach for Identifying Significant AID Variants and Their Functional Annotation

After analyzing the PGS datasets, the study transitions to the genetic data processing phase. For this investigation, variants related data and associated information were downloaded from Zenodo on May 30, 2024. Zenodo is a free, open-access digital repository designed to store, share, and preserve research outputs across multiple disciplines. The variant dataset, provided by Thompson et al. (2024), via Zenodo, is generated using a cohort of unrelated White British individuals from the UKB. To identify significant variants, a genome-wide association significance threshold of $P < -\log_{10}(5 \times 10^{-8})$ (minus_log10_p $\geq$ 7.30103) is applied. The Human leukocyte antigen (HLA) region, located on chromosome 6, is a gene-dense and highly complex area. It is often challenging to pinpoint the specific variants responsible for traits and diseases within this region. Therefore, the HLA region has been excluded from the study. Many studies have shown that variants near the HLA region are likely to be correlated with each other due to their physical proximity, which can impact the interpretation of genetic associations and may obscure the relationship between variants and AIDs. Removing this region and its surroundings can thus improve the clarity of genetic analyses. The assembly used in the Thompson et al. (2024), study is GRCh37, and accordingly, the genomic coordinates for the HLA region in the hg37 reference genome are used (28,477,797–33,448,354). A buffer of 100,000 base pairs is applied upstream and downstream, resulting in the exclusion of the region spanning from 27,400,000 to 34,400,000. After filtering out the HLA region, the number of variants decreased for most AIDs. To improve mapping accuracy and the relevance of genomic studies for disease-associated variants, the coordinates from GRCh37 are converted to the latest human reference genome assembly, GRCh38. This conversion is carried out using the Lift Genome Annotations tool from the University of California Santa Cruz (UCSC) genome browser, which facilitates the translation of genomic coordinates between different genome assemblies. It identifies genes with changed annotations and uses pre-generated chains to convert genomic features between assemblies [240], [241] [242]. The analysis is performed between 22nd October 2024 and 25th February 2025, using the “Dec.2013 (GRCh38/hg38)” assembly.

The newly converted genomic coordinates of Single Nucleotide Polymorphisms (SNPs) (from GRCh37 to GRCh38) are used as input for the Variant Effect Predictor (VEP), accessed through the Ensembl Genome Browser. VEP is widely used for analyzing,

annotating, and prioritizing genomic variants in both coding and non-coding regions. For this analysis, the reference genome assembly used is GRCh38.p14, with VEP version v113.0 and the database homo_sapiens_core_113_38. Additionally, the Genome Aggregation Database (gnomAD) version 4 is integrated into the analysis. The study workflow is illustrated in the flowchart (Figure$\sim$ 2.1). The VEP output file holds information on chromosomes, variant's position, alternative and reference allele, beta value, standard error, -log 10 p value and number of case and control samples. The alternative allele is the variant differing from the reference at that position, while the reference allele is not necessarily the most common in a population. Additionally, it is essential to ensure that allele coding (reference/alternative) is consistent across datasets to avoid errors in interpreting the direction of the effect. The $-\log 10(p)$ transformation is usually applied to variant associated p-values to simplify their interpretation.

Procedure for Analyzing Significant Variants Based on Allele Frequency and Functional Impact

To identify novel gene and variant associations relevant to AIDs, genetic variants are prioritized based on allele frequency (AF) and predicted impact using output from the VEP tool, which annotates and characterizes genomic variants. Two primary filtration criteria are applied to the VEP output: AF and impact. AF measures the proportion of a specific allele relative to all alleles of that gene within a population, offering insights into the genetic diversity and distinguishing between common and rare variants. Variants filtered based on AF are divided into two groups: those with an AF greater than or equal to 0.05; common variant, and those with an AF less than 0.05; rare variant. To address the clustering of GWAS variants within regions of linkage disequilibrium, variants located within 500,000 base pairs of each other are aggregated into loci. This aggregation is performed separately for each AF group, resulting in two distinct sets of genes for each AID. The impact scores assigned by VEP categorized as "high", "moderate", "low" or "modifier" further characterize the functional consequences of these variants. Most selected variants based on AF filtering are of "modifier" or "low" impact, typically affecting non-coding regions or causing minor amino acid changes unlikely to alter protein function significantly. A second analytical approach focuses specifically on the biological impact of variants. Variants predicted to have high or moderate impact (such as missense variants, splice acceptor or donor variants, frameshift variants, and inframe deletions) are selected, as they are more likely to result in loss of function or significant alterations to protein structure. This filtering yields a separate set of genes associated with each AID.

Finally, the rsIDs and their corresponding genes are examined in the GWAS Catalog to investigate known associations, related traits, and reported studies. A comprehensive

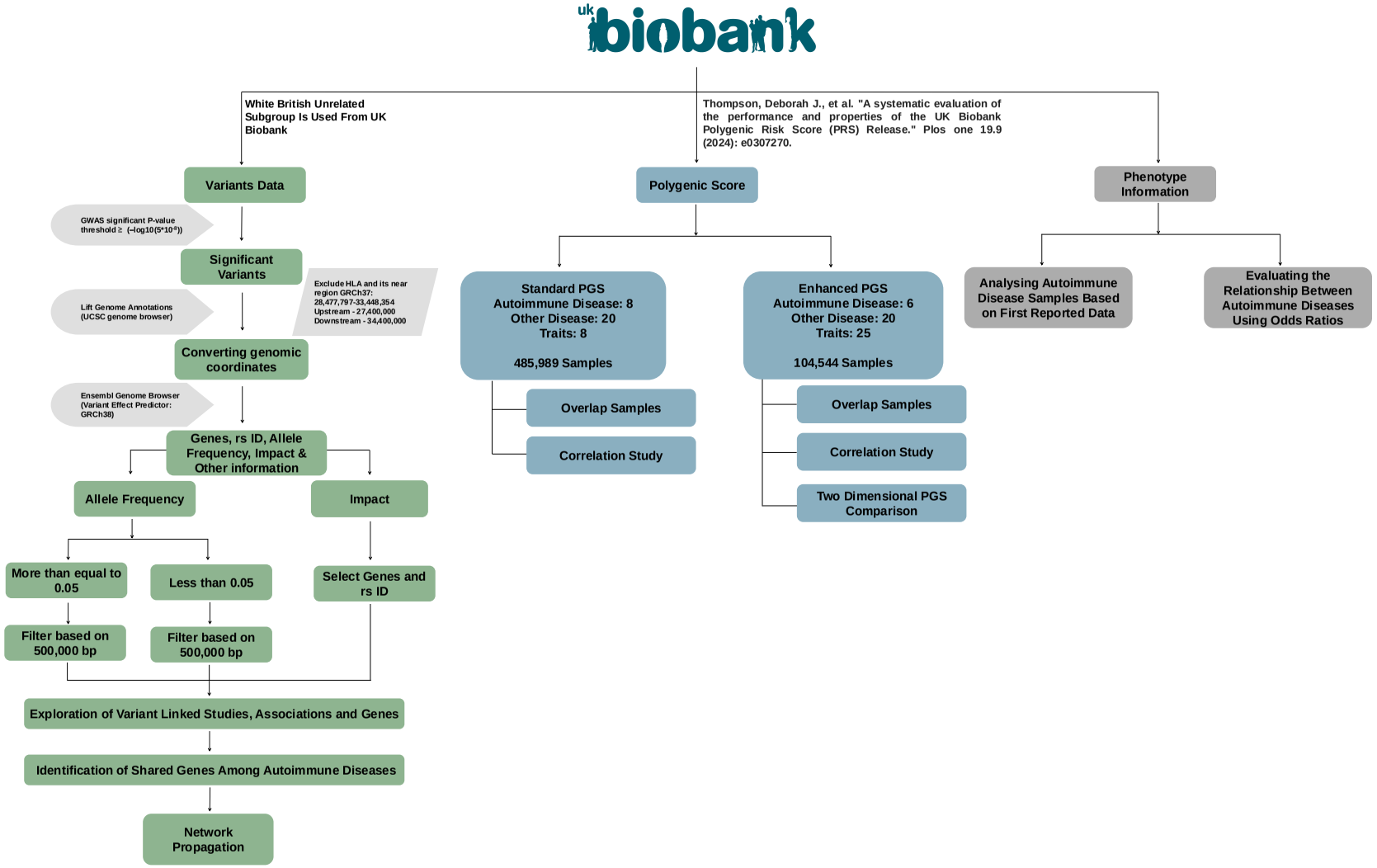


Figure 2.1: The flowchart presents the study design for analyzing data from the UKB. Color scheme representing three types of UKB data and corresponding analytical steps. It includes three analysis: the right-hand panel (dark gray) shows phenotypic data are used to assess AID co-occurrence and explore sample characteristics. The central panel (blue) illustrates analyses of PGS data, including correlation assessments and risk-score distribution studies. The left panel (green) depicts steps of variant-level analyses in the White British unrelated subgroup, highlighting the identification of genes shared across multiple AIDs.The light gray sections illustrate the tools used and any data extraction steps performed during the analysis. See the “Study Background and Framework” section for detailed descriptions of the standard and enhanced datasets.

interpretation is achieved by incorporating multiple layers of annotation, filtering based on unique rsIDs, and retrieving both gene and transcript-level information. Special attention is given to genes commonly identified between AIDs through any of approaches (rare variant analysis, common variant analysis, and predicted impact analysis).

Methodology for Pathway Mapping of Common Genes Associated with Significant Variants via Network Propagation

Network propagation is performed on genes common to several AIDs to leverage existing interaction data and sharpen the shared genetic signal. This expands the initial overlap into a broader disease module, adding neighbouring genes that likely act in the same pathways and exposing shared pathways or key hub proteins. In this study, hallmark gene sets from the Molecular Signatures Database (MSigDB) are employed to explore biological systems and pathways associated with the identified genes. Hallmark gene sets represent distinct, well-defined biological states or processes, characterized by coherent patterns of gene expression. These sets are constructed through a computational methodology that identifies overlaps among gene sets and retains genes exhibiting coordinated expression. By reducing noise and redundancy, the hallmark gene sets provide a clearer and more biologically relevant framework for analysis. The microarray data used for the refinement and validation of these signatures are available on the hallmark gene set pages, and the hallmark collection is freely accessible as the H Collection within MSigDB [243]. Network propagation begins with a defined set of seed genes and is executed on the STRING v12 interaction network [244]. Only high-confidence edges (score $\geq$ 700) are retained, ultra-connected hubs (degree $\geq$ 900) and self-loops are removed, and the analysis is limited to the largest connected component. Signals are then diffused with a regularized Laplacian kernel, which smooths local neighbourhoods while damping extreme values. The raw diffusion scores are converted to permutation-based z-scores (1 000 random seed shuffles) and subsequently standardised, producing a robust map of network-level effects. All propagation steps are carried out in R (version 4.4.1) using the diffuStats and GSVA libraries. This network-based approach identifies biological pathways that are functionally linked to genes harbouring disease-related variants. After the high-confidence edge filter, hub removal, Laplacian diffusion, permutation-based z-scoring, and cross-validation across 1000 random seed shuffles, the propagated signals remain significant, greatly reducing spurious noise and increasing confidence that highlighted nodes reflect genuine network-level associations. The final output is a table of propagated scores in which each row represents a pathway and each column represents a disease. Scores range from –1 to +1: values near +1 indicate strong signal accumulation and likely disease relevance, values near –1 suggest pathway under-activation or suppression,

and scores around zero denote minimal propagation.

2.3 Retrospective Analysis of Comorbidities in Pemphigus Patients Across Diverse Ancestral Backgrounds utilizing TNX Database

2.3.1 Study Design

The analysis is divided into two parts to systematically explore AIDs with a high comorbidity with pemphigus (Figure$\sim$ 2.2). The primary analysis, termed the 'Competing Risk' analysis, aims to identify AIDs with a high prevalence in the pemphigus group. Diseases with a patient count of at least 30 are considered for further statistical analysis in the follow up "Compare Outcomes" study. The secondary analysis compares pemphigus patients to a control group, with additional consideration for ethnicity-specific risk assessment. The control group, propensity-matched (establishes a comprehensive and balanced comparison between treatment and control groups) to the pemphigus group, is designed to control for potential confounders.

2.3.2 Dataset Definition and Propensity Matching

For the initial 'Competing Risk' analysis, patients diagnosed with pemphigus (ICD10CM:L10) are selected from the database without consideration for sex, age, or ethnicity. Out of 108 healthcare companies worldwide, 103 provided accessible information on 129,786,809 patients, of which 17,474 are diagnosed with pemphigus. Patients with data recorded more than 20 years ago were automatically excluded by TNX, resulting in a final cohort of 14,801 individuals. Subsequently, this cohort underwent analysis, encompassing 92 AIDs categorized by their ICD-10-CM codes, as derived from Samuels et al. [245]. The analysis spanned from one day after the index event (diagnosis of pemphigus, ICD10CM:L10) to any time thereafter. Eighteen AIDs are excluded from the list due to various reasons: four lacked data on TNX, eight had uncleared ICD-10 codes, three had duplicate names, and one is restricted on TNX. After filtering, 26 diseases with at least 30 patients progressed to the follow-up study. Note that some of the diseases have slightly different names on TNX than in the literature.
The second "Compare Outcomes" study aimed to measure actual comorbidity by comparing pemphigus patients to a propensity-matched control group without pemphigus. Propensity matching is based on sex, age, age at index (time of pemphigus diagnosis, ICD10CM:L10), and ethnicity. It is noted that TriNetX distinguishes between "ethnicity" and "race" on a technical basis. However, both classifications rely on self-identification

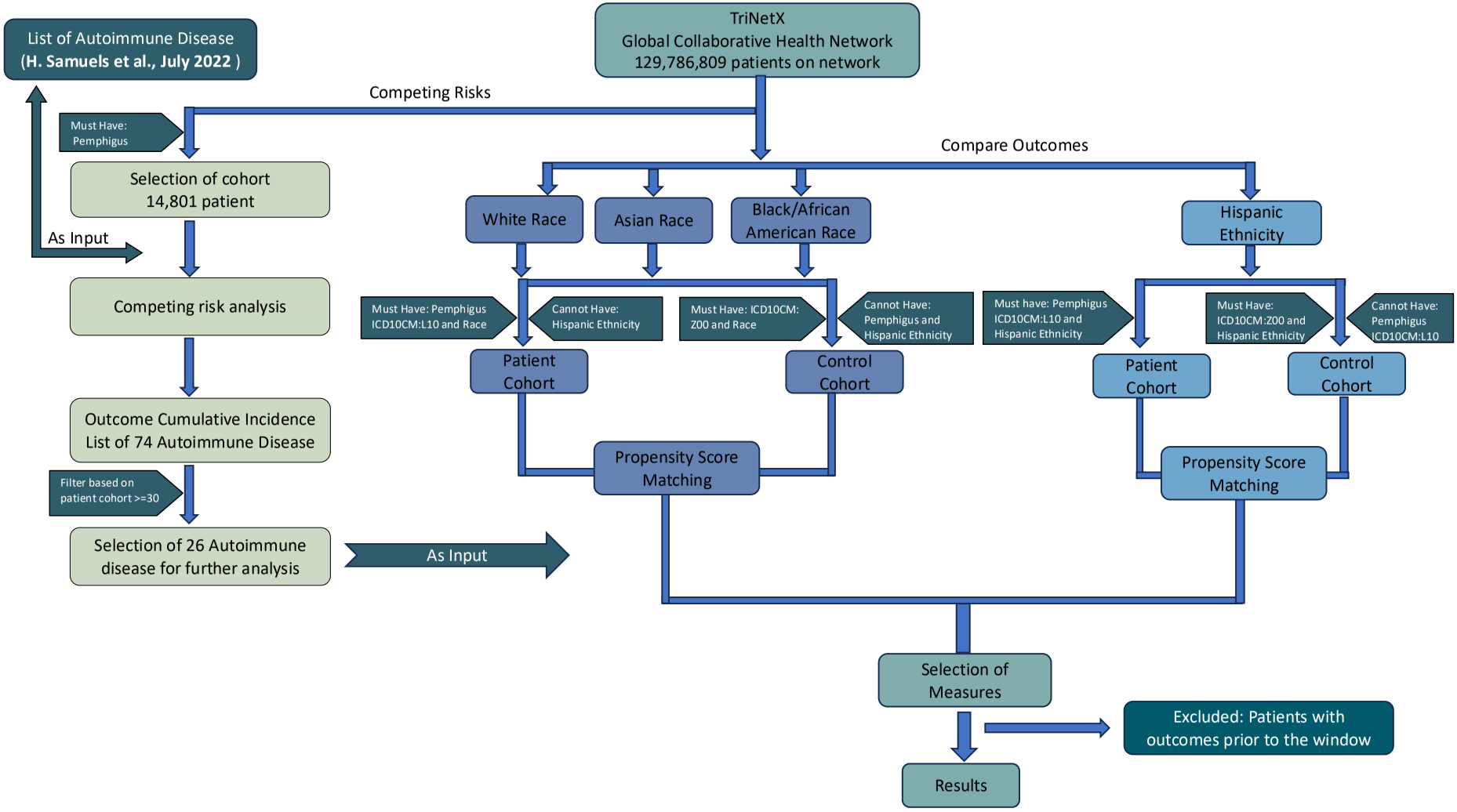


Figure 2.2: Study design outline for investigating ancestry and comorbid AIDs risk in patients with pemphigus using the TriNetX database. It includes two analysis: on the left hand the preliminary competing risks study, in which systemically identify a comprehensive overview of AIDs with high prevalence in the studied pemphigus group. On the right-hand side is the comparative outcomes study, where prevalences are compared to a control group to assess actual comorbidity among White, Black/African American, Asian race and Hispanic ethnicity.

and are treated similarly in the dataset; therefore, the terms are used interchangeably in this study. The database is queried to select patients of any age and both sexes recorded on the TNX platform in the last 20 years, meeting the criteria for pemphigus diagnosis (ICD10CM:L10). As TNX limits the number of potential confounders that can be controlled simultaneously, the analysis is divided into ethnicity groups, namely White, Black/African American, and Asian (Groups 1–3). Since people with a Hispanic ethnicity can belong to any or multiple of these groups, they are excluded and a Hispanic group (Group 4) is established, exclusively comprising patients of Hispanic ethnicity. Propensity score matching aligned the ethnicity-based cohorts, focusing solely on demographic criteria such as current age, age at the index (time of pemphigus diagnosis, ICD10CM:L10), and sex.

Following propensity score matching, the high-prevalence diseases from the initial study are used as input. Patients experiencing the diseases before pemphigus itself are excluded. This analysis employed several measures to assess comorbidity, including risk ratio and odds ratio obtained through the "Measures of Association" tool in TNX, providing direct comparisons of prevalence between groups. Additionally, hazard ratio (HR) is derived using Kaplan-Meier survival analysis to account for risk over time. A HR is a measure used in survival analysis to compare the risk of an event (such as death, disease recurrence, or failure) occurring at any given time between two groups cohorts under study. Derived from the Cox proportional hazards model, HR is expressed as

$$\text{HR} = \exp(\beta) \tag{2.4}$$

where (β) represents the regression coefficient for a predictor variable, adjusting for covariates. This semi-parametric model assumes proportional hazards over time, implying the HR between any two groups remains constant over time. HR calculation involves fitting time-to-event data to the Cox model, estimating (β), and exponentiating it to derive the ratio, accompanied by 95% confidence intervals and p-values to assess statistical significance. An HR $>$ 1 indicates elevated risk in the exposed group, while HR $<$ 1 denotes protective effects. Unfortunately, the OR could not be properly calculated for several subcategories due to limited data combined with the obfuscation approach used by TNX for patient anonymization. The analysis is primarily based on HR, as this obfuscation does not apply to the HR. Furthermore, significance can only be shown up to $p < 0.001$ due to platform limitations.

Valuable feedback from Professor Ralf Ludwig highlighted a significant methodological concern regarding disease classification. Specifically, he noted that a major confounder in the study stems from the definition of pemphigus based on the ICD-10 code L10. This code encompasses not only the classical forms pemphigus vulgaris (PV)

and pemphigus foliaceus (PF) but also includes endemic PF, paraneoplastic pemphigus, and other subtypes. Additionally, he emphasized the importance of excluding other autoimmune blistering diseases to ensure a more accurate and specific definition of pemphigus for analysis. An additional analysis was undertaken based on his suggestion, selecting ICD-10 codes L10.1, L10.0, L10.2, and L10.4 on TNX to define the pemphigus cohort. The refined analysis is performed specifically for individuals of the white race, using a narrower case definition to exclude non-classical subtypes and other autoimmune blistering diseases.

2.4 Divergence and Convergence in Autoimmune Comorbidity Patterns

This comorbidity analysis encompasses both common and rare AIDs spanning multiple organ systems. Neurological conditions include MS, myasthenia gravis, inflammatory polyneuropathy, and other demyelinating diseases. Dermatological and cutaneous disorders comprise vitiligo, pemphigus, and PSO. Gastrointestinal AIDs cover CD and UC, while multi organ or systemic conditions include RA, ankylosing spondylitis, psoriatic and enteropathic arthropathies, SLE, and reactive arthropathies. Juvenile arthritis represents internal organ involvement. All analyses focus on individuals of White ancestry. Because the onset age varies by disease and AIDs can emerge at any stage of life [246] age at onset (defined as the first symptomatic manifestation) is incorporated for each AID. In the TNX cohort, this variable corresponds to "Age at Index," whereas the UKB supplies analogous disease specific age metrics. ICD-10 codes are employed to identify the same AIDs in both the databases. Harmonised ICD-10 mapping ensures consistency between the TNX and UKB cohorts.
Comorbidity in both TNX and UKB is assessed via odds ratios calculated for AIDs. Data retrieval and analysis for TNX were completed on 30 May 2024, using its global collaborative network of 120 health care organisations. Of the 163,606,965 patients available, 71,069,654 are recorded as White; cohorts are defined by query criteria in which each AID case cohort contains a disease diagnosis and White demographic, with no control samples (ICD10CM:Z00), whereas each control cohort contains only ICD10CM:Z00 diagnoses, White demographic, no specific AIDs cases which are being used in the analysis. Propensity matching on sex, age, age at index, and ethnicity is applied. The TNX platform then yields odds ratios for comorbid conditions in White AID samples. UKB phenotype data, extracted on 13 July 2023, span 15 AIDs (diagnoses dated 1902 – 2022). Disease specific age at index cut offs are applied, and OR are computed in RStudio using coded Fisher's exact tests; results from the two datasets are visualised with the "ggplot" library.

Chapter 3

Results

3.1 Results from GWAS, PGS, and UKB Data Retrieval Using EFO for Autoimmune Disease Analysis

3.1.1 Characterization of Autoimmune Disease Samples and Studies Across Ontologies, UKB, GWAS, and PGS Catalogs

The ontology branch of child terms for autoimmune diseases (AIDs) contains 120 terms organized at up to five levels (Suppl. Table S1). A total of 13 terms have been taken from the Mondo disease ontology, 3 terms have been derived from Orphanet and 1 term from the disease ontology (Suppl. Table S2). Overall, the Genome-Wide Association Study (GWAS) catalog studies contain 442 autoimmune disease GWAS ("STUDY ACCESSION") published between 2006 and 2022 in 58 different journals with 221 unique PubMed IDs (Suppl. Table S3); these studies have been conducted on 377 different datasets (according to column "INITIAL SAMPLE SIZE" in Suppl. Table S3). A subset of studies (n=179 (47%)) reported no genome-wide significant variants. The GWAS catalog contains 5,023 associations that cover 41 autoimmune diseases (according to EFO ID) based on 253 datasets (according to column "INITIAL SAMPLE SIZE") relating to 200 unique PubMed IDs (Suppl. Table S4). These associations correspond to 3,212 unique variants (according to column "SNPS") and 1,760 unique genes or gene combinations reported in the literature (column "REPORTED GENE(S)"; Suppl. Table S4). The polygenic scores (PGS) catalog contains 18 AIDs for which 47 PGS are published in 14 papers between 2018 and 2022 (cf. Suppl. Table S5). These have been developed with 15 different computational methods, mostly with the tools snpnet [n=18 scores; [247]], using genome-wide significant GWAS variants (n=7 scores), LDPred [n=6; version 1 and 2; [248] [249]] or by applying pruning and threshold (n=4). Corresponding to the method or tool applied for PGS constructions, the number of variants considered in the scores ranges from 3 to 6,907,112. Scores

have been developed using several cohorts of mainly European and/or East Asian (primarily Chinese) ancestry, primarily as source GWAS, but also to train parameters (SNP weights, Effect sizes across different populations, linkage disequilibrium) (Suppl. Table S6). Further, many PGS have been developed using the UK Biobank (UKB). The PGS catalog source files obtained via experimental factor ontology (EFO) IDs for all information related to AIDs can be found: PGS (Suppl. Table S3), score development samples (Suppl. Table S4), performance evaluation metrics (Suppl. Table S5) and evaluation samples (Suppl. Table S6). PGS that are associated with autoimmunity have been assessed in 124 data sets (Suppl. Table S7), leading to 225 performance assessments published in 16 papers (Suppl. Table S8). One of the most common performance measures is the Area Under the Receiver-Operating Characteristic Curve (AUROC), which compares the number of individuals who are incorrectly identified as having a disease (false positive rate) to the number who are correctly diagnosed as having a disease (true positive rate). An AUROC value of 0.5 suggests that the PGS has no discriminatory power, performing no better than random chance. Values closer to 1.0 indicate excellent discrimination, meaning the score effectively distinguishes between individuals with and without the disease. Thus, higher AUROC values reflect greater accuracy and reliability of the PGS in predicting autoimmune conditions. It is typically found that AUROC classification performance of autoimmune PGS ranges from 0.56 to 0.99 (provided for n=124; Suppl. Table S8). As a whole,16 publications constructed and/or evaluated an AID PGS, according to the PGS catalog (columns ”PGS Publication (PGP ID)” of Supplemental Table S5 and Suppl. Table S8). Of the 120 autoimmune EFO terms examined (cf. Suppl. Table S2), 20 are found to correspond to patients and associated genotypes within the UKB dataset (Suppl. Table S9). Among these, 9 had both self-reported and ICD-10 diagnoses, 6 had only self-reported data, and 5 had diagnosis information based solely on ICD-10 codes. The number of respective patients ranges from 13 (reactive arthritis with ICD-10 code M03) to 12,556 (rheumatoid arthritis (RA) with ICD-10 code M06) (Suppl. Table S10).

3.1.2 Overlap of Autoimmune Disease Across UKB, GWAS and PGS catalog

The investigation is conducted to determine which AIDs are represented in the GWAS catalog, the PGS catalog, and the UKB. As it is achieved by comparing the unique autoimmune EFO IDs that are present in each of these three resources. 120 EFO IDs associated with AIDs (Suppl. Table S2) cover a range of different levels of diagnosis (Suppl. Table S1). According to the GWAS catalog, 41 EFO IDs are covered by the catalog (Suppl. Table S5), while the PGS catalog covers 18 EFO IDs (Suppl. Table S5). Considering that the UKB does not use EFO IDs, A published mapping of EFO

IDs to UKB data fields has been utilised and is provided in Supplement Table S10. This assigned 20 EFO IDs to traits in UKB (Suppl. Table S9). A comparison of the overlap between the three areas of AIDs covered by each of the three resources is shown in Figure$\sim$ 3.1. In all three databases, out of 120, 48 EFO IDs correspond to AIDs, the majority appearing in GWAS catalog. The GWAS catalog is sharing 15 AID EFO IDs with the PGS catalog and 15 can be mapped to UKB. Further, 12 disease EFO IDs are shared between PGS catalog and UKB. There are 11 AID EFO IDs common to all three databases. They relate to ankylosing spondylitis, appendicitis, CD, IBD, juvenile idiopathic arthritis, psoriatic arthritis, RA, Sjögren syndrome, SLE, UC and Vitiligo. Several of the AIDs related to EFO IDs that are not shared by all three resources are cases that highlight limitations with respect to the definition of terms and relationships within the EFO and in the mapping of EFO terms to external codes and identifiers, which may not be one-on-one and needs disease-specific knowledge (for details see caption of Figure$\sim$ 3.1. There are 10 AIDs with most GWAS studies according to GWAS catalog is investigated more closely (Table$\sim$ 3.1). They are SLE [250], RA [251], MS [252], IBD [253] with its two subtypes CD and UC, vitiligo [254], Sjögren syndrome [255], Grave's disease [256], and Behcet's syndrome [257]. The GWAS catalog association data of the top 10 AIDs (underlying Table$\sim$ 3.1) are provided in Suppl. Tables S10-S19. The most GWAS studied AID is SLE, for which 37 different GWAS have been performed, the largest one using 13,377 cases and 194,993 controls. These studies have reported 788 unique single nucleotide polymorphisms (SNPs) and 439 unique genes or gene combinations. In the PGS catalog, six studies are noted on SLE, which report six different risk scores. These PGS have been evaluated in 32 settings. The largest number of cases has been analyzed for IBD (n=25,042), the lowest for Sjögren syndrome (n=1,599). Overall, the number of independent genomic loci associated with disease increases with the number of studies and cases (Table$\sim$ 3.1).

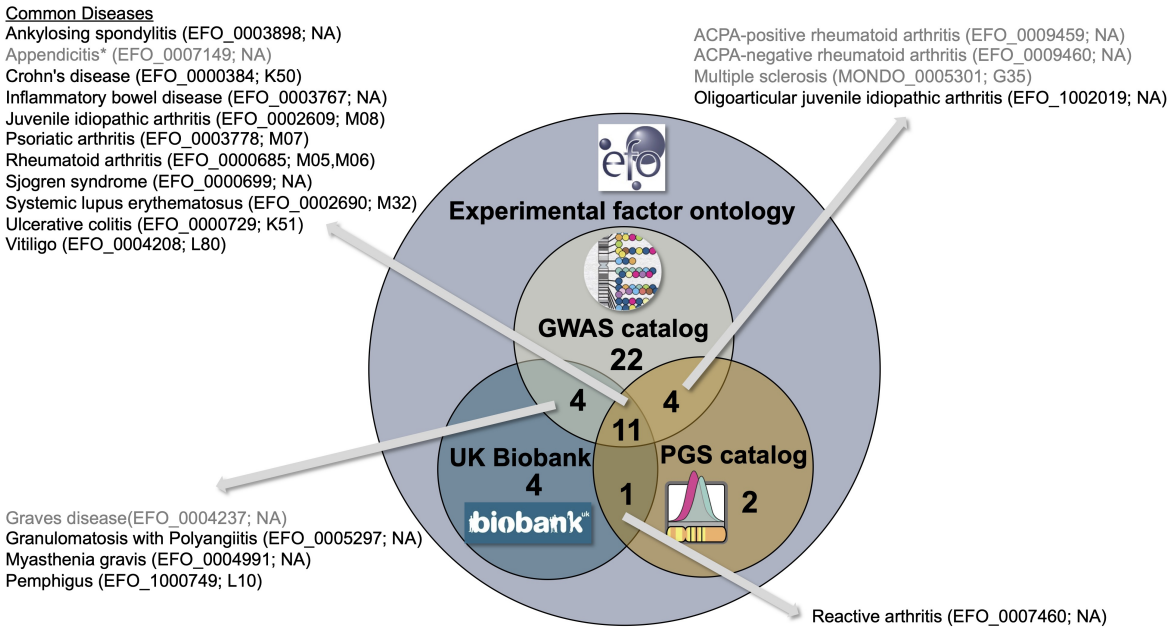


Figure 3.1: A Venn diagram representing the number of AIDs in the EFO overlapping with GWAS catalog and PGS catalog covered EFO IDs as well as UKB data field matched according to Suppl. Table S10. EFO IDs of diseases in more than two resources are listed together with their name according to EFO and their UKB ICD-10 code according to Suppl. Table S10. Diseases shaded gray are affected by issues with disease definition and classification compromising the mapping. These issues are: (i) Appendicitis is classified as an AID in EFO because it is a child term of IBD, however, it is not considered an AID. (ii) Grave's disease EFO child terms are in PGS, not the EFO ID of Grave's disease itself though. (iii) ACPA-positive and ACPA-negative RA is not mappable to UKB. UKB, however, contains seropositive and other RA, a distinction not covered by EFO. (iv) UKB has information on MS, yet since the recent EFO version is updated to using the MONDO ID for MS, the mapping to UKB data fields failed. "NA" denotes that mapping to UKB is not available for the respective EFO ID.

Table 3.1: The ten AIDs (defined by EFO term) which have the highest number of GWAS studies registered at the GWAS catalog. Displayed is the summary of information obtained from GWAS catalog, PGS catalog and UKB. With respect to GWAS catalog, this is the number of unique studies (according to column "STUDY ACCESSION" of Suppl. Table S3), the highest number of cases with corresponding number of controls, the number of unique variants reported (according to column "SNP_ID_CURRENT" of Suppl. Table S4), the number of independent, associated genomic loci reported in the literature, the number of unique genes or gene combinations reported in the respective publications (according to column "REPORTED GENE(S)" of Suppl. Table S5). With respect to PGS catalog, reported are the number of unique studies (according to column "PGS Publication (PGP) ID" of Suppl. Table S5 and Suppl. Table S8), unique scores developed (according to column "Polygenic Score (PGS) ID" of Suppl. Table S6), the range of variants utilized in the scores for the respective disease and the number of performance evaluations in independent samples (according to column "PGS Performance Metric (PPM) ID" of Suppl. Table S8). Finally, for the UKB, the UKB data field, ICD-10 code (if available in UKB) and patient number according to Suppl. Table S9 is provided.

Trait	EFO IDs	GWAS Catalog						PGS Catalog				UKB		
		Studies	Cases	Controls	SNPs	Loci	Genes	Studies	PGS ID	Variants	Eval.	Field	ICD-10	Indiv.
SLE	EFO_0002690	37	13,377	194,993	788	132[1]	439	6	6	41–293,684	32	131894	M32	1,053
RA	EFO_0000685	37	22,628	288,664	421	>150[2]	249	3	6	3–95,083	33	131850	M06	12,556
MS	MONDO_0005301	27	14,802	26,703	603	233[3]	479	3	5	36–129,077	25	131042	G35	2,518
CD	EFO_0000384	27	12,924	21,442	411	>200[4]	265	1	2	220–257	9	131626	K50	3,355
UC	EFO_0000729	25	12,366	33,609	295	>200[4]	184	2	4	179–566,637	26	131628	K51	6,451
IBD	EFO_0003767	12	25,042	34,915	387	>200[4]	238	3	2	195–690,711	7	–	–	–
Vitiligo	EFO_0004208	10	2,853	37,405	91	49[5]	80	3	3	42–77	10	131802	L80	1,201
Sjögren syndrome	EFO_0000699	10	1,599	658,316	48	25[6]	42	1	1	7	5	20002_1382	–	572***
Grave's disease	EFO_0004237	8	4,487	629,598	74	12[7]	27	–	–	–	–	20002_1522	–	183***
Behçet's syndrome	EFO_0003780	8	3,197	5,785	40	21[8]	35	–	–	–	–	41202	–	18****

[1] [250]; [2] [251]; [3] [252]; [4] [253]; [5] [254]; [6] [255]; [7] [256]; [8] [257]; * Excludes seropositive RA (M05) with 1,401 patients; ** K50+K51 combined; *** Self-reported; **** Based on hospital medical history.

3.2 Outcomes from the Analysis of Phenotype Data, Risk Scores, and Genetic Variants Using UKB Data

This section comprehensively presents the findings from analyses of phenotype data, risk-score correlations and score distributions, utilizing PGS datasets and variant information derived from the unrelated White British subgroup. Each subsection outlines the results obtained through a specific analytical approach.

3.2.1 Phenotype Based Evaluation of Sample Overlap and Comorbid Relationships in Autoimmune Diseases

Identifying frequently co-occurring autoimmune conditions highlights shared etiological pathways and guides mechanistic investigations. Accordingly, this section quantifies the overlap among AIDs case samples in the UKB and evaluates their comorbid relationships. First-reported phenotype data from UKB are used to map shared diagnoses across cohorts and measure the strength of disease co-occurrence.

The phenotype data is available for all 502,371 patients in the UKB. Among AIDs, PSO has the largest number of samples (15,978), followed by RA with 13,344 samples, and UC with 6,702 samples. The fewest samples are reported for SLE with 1,101 cases. Multiple sclerosis (MS) and Crohn's disease (CD) have 2,593 and 3,503 patients, respectively (Table$\sim$ 3.2). The distribution of shared and unique samples among AIDs based on phenotype data is illustrated in Figure$\sim$ 3.2. Each column represents a distinct intersection of disease sets, with dots indicating the sets involved and set sizes corresponding to the number of associated samples. MS exhibits the highest proportion of unshared samples (2,367; 91.2%), followed by PSO (14,379; 89.9%) and RA (11,528; 86.3%). In contrast, CD and UC show the lowest proportions of unshared samples, at (2,130; 60.8%) and (5,074; 75.7%), respectively. The most substantial sample overlap is observed between PSO and RA (993), followed by CD and UC (922), and RA and UC (247).

Positive odds ratios, many of which are statistically significant ($p < 0.05$), are identified between several AID pairs. The calculation of odds ratio is explained in method (2.2.1) section. The strongest positive associations are found between RA and SLE (1.03), MS and SLE (0.58), and PSO and RA (0.45). One non-significant positive odds ratio is observed between MS and PSO (0.06). Additionally, the odds ratio for MS and RA is slightly negative (–0.01), though this is not statistically significant. Figure$\sim$ 3.3 presents the OR transformed to a log10 scale, with corresponding exact values available in Suppl. Table S20.

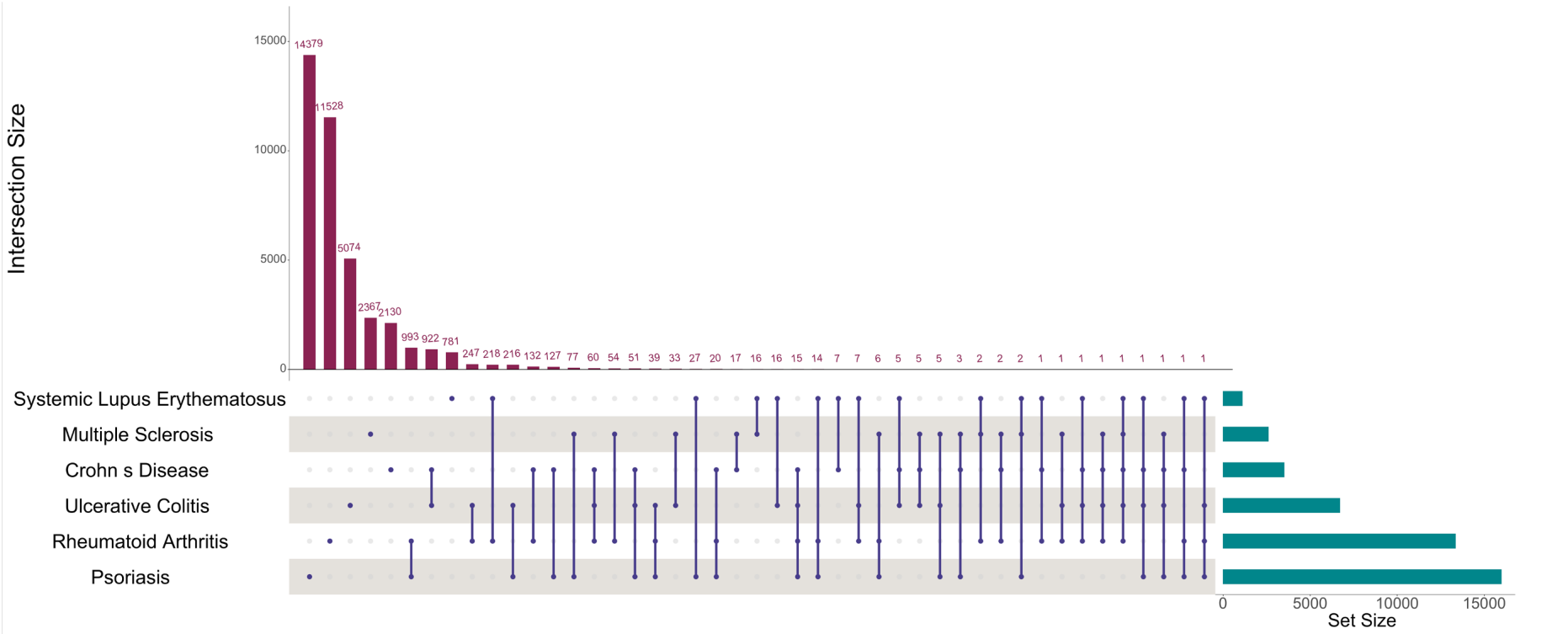


Figure 3.2: Upset plot illustrates the overlap of samples among AIDs using first reported (phenotype) data from the UKB. The dark red bar chart quantifies each intersection, with the numbers top each bar indicating the sample count shared between the sets. Each row represents an original set, while each column corresponds to a specific intersection. Within the matrix, blue filled dots denote the presence of a set in a given intersection, and empty spaces indicate its absence. Additionally, a separate dark green section displays the total size of each individual set for reference.

Table 3.2: First reported sample information for selected AIDs obtained from the UKB. The first three columns list the diseases along with their corresponding ICD-10 codes and UKB Field IDs. The (No. of Patients) column represents the phenotype data extracted from the UKB, detailing the total number of cases in each disease, while the (Male) and (Female) columns specify the number of male and female patients within phenotype dataset.

Diseases	ICD10 Code	Field ID (UKB)	No. of Patients (First Reported)	Male	Female
CD	K50	131626	3,503	1,568	1,935
MS	G35	131042	2,593	721	1,872
PSO	L40	131742	15,978	7,982	7,996
RA	M05	131850	13,344	4,364	8,980
SLE	M32	131894	1,101	166	935
UC	K51	131628	6,702	3,358	3,344

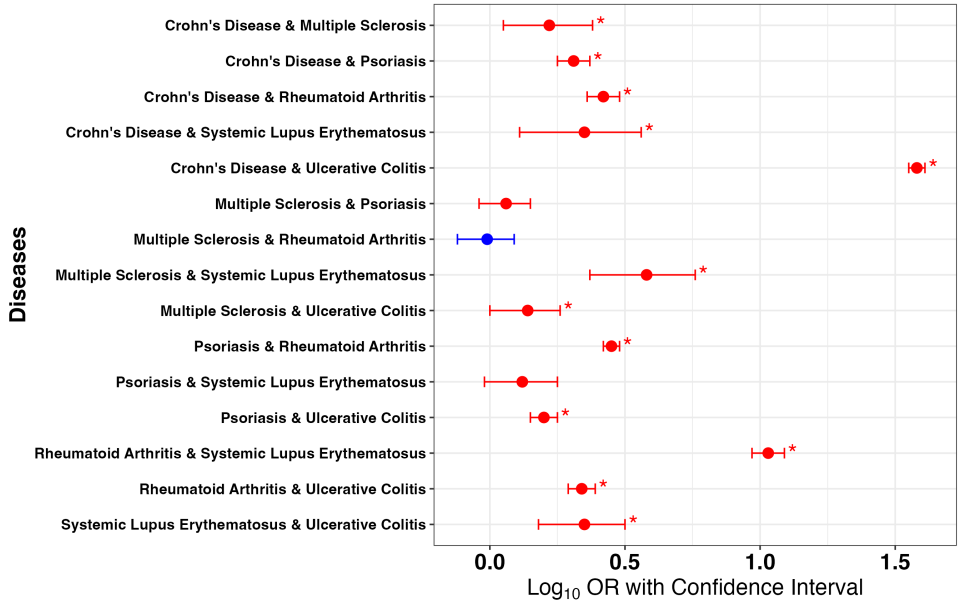


Figure 3.3: The forest plot illustrates the odds ratios for comorbid conditions between AIDs. Each line represents a unique disease combination, with the central section displaying the log10 OR and its corresponding CI. Additionally, The plot displays p-values denoting statistical significance ($p < 0.05$) for each association, represented by a red star. The position relative to the reference line (zero) determines whether the association is positive or negative. Negative odds ratios reflect a potential protective effect against the condition, whereas positive odds ratios indicate comorbidity.

3.2.2 Findings on Cohort Overlap, PGS Correlations, and Distribution of Genetic Risk Score Across Autoimmune Diseases

Analysis of Cohort Overlap and Polygenic Score Correlations Across Standard and Enhanced Datasets

Ensuring the validity of comparative analyses and guarding against sampling biases motivates an examination of cohort overlap within and between the standard and enhanced PGS datasets. The analysis draws on participant data from both PGS datasets and a broad spectrum of diseases and traits. By quantifying the degree of overlap and calculating correlation metrics across phenotypes, it illuminates shared genetic architectures and potential pleiotropic effects, thereby enriching the interpretation of genetic risk patterns

The sample overlap analysis between the standard and enhanced PGS datasets reveals that the enhanced dataset is a subset of the standard dataset (Figure$\sim$ A.1). A total of 104,544 individuals are shared between both datasets, while 381,445 samples are unique to the standard dataset. The enhanced dataset comprises 6 AIDs, 20 other diseases, and 25 quantitative traits, whereas the standard dataset includes 8 AIDs, 20 other diseases, and 8 quantitative traits (Figure$\sim$ 3.4. In terms of specific disease representation across datasets, psoriasis (PSO) again leads with 3,142 patients in the enhanced dataset and 15,527 in the standard dataset. RA follows with (2,747; Enhanced, 12,792; Standard) patients. SLE has the fewest samples (312; Enhanced, 1,054; Standard) dataset followed by MS with (478; Enhanced, 2,488; Standard) cases. Notably, SLE shows a marked female predominance. In the enhanced dataset, there are 37 males and 275 females, while the standard dataset includes 156 males and 898 females (Table$\sim$ 3.3).

Table 3.3: Patient counts in the enhanced and standard datasets across autoimmune diseases. The first three columns list the disease names, their corresponding ICD-10 codes, and UKB field IDs. The remaining columns provide patient counts for each disease. PGS data is available for all patients.

Diseases	ICD10 Code	UKB ID	Enhanced Dataset			Standard Dataset		
			Total	Male	Female	Total	Male	Female
CD	K50	131626	720	324	396	3,376	1,529	1,847
MS	G35	131042	478	123	355	2,488	690	1,798
PSO	L40	131742	3,142	1,580	1,562	15,527	7,805	7,722
RA	M05	131850	2,747	842	1,905	12,792	4,206	8,586
SLE	M32	131894	312	37	275	1,054	156	898
UC	K51	131628	1,419	722	697	6,497	3,277	3,220

Within enhanced and standard dataset, the pattern of sample overlap across diseases

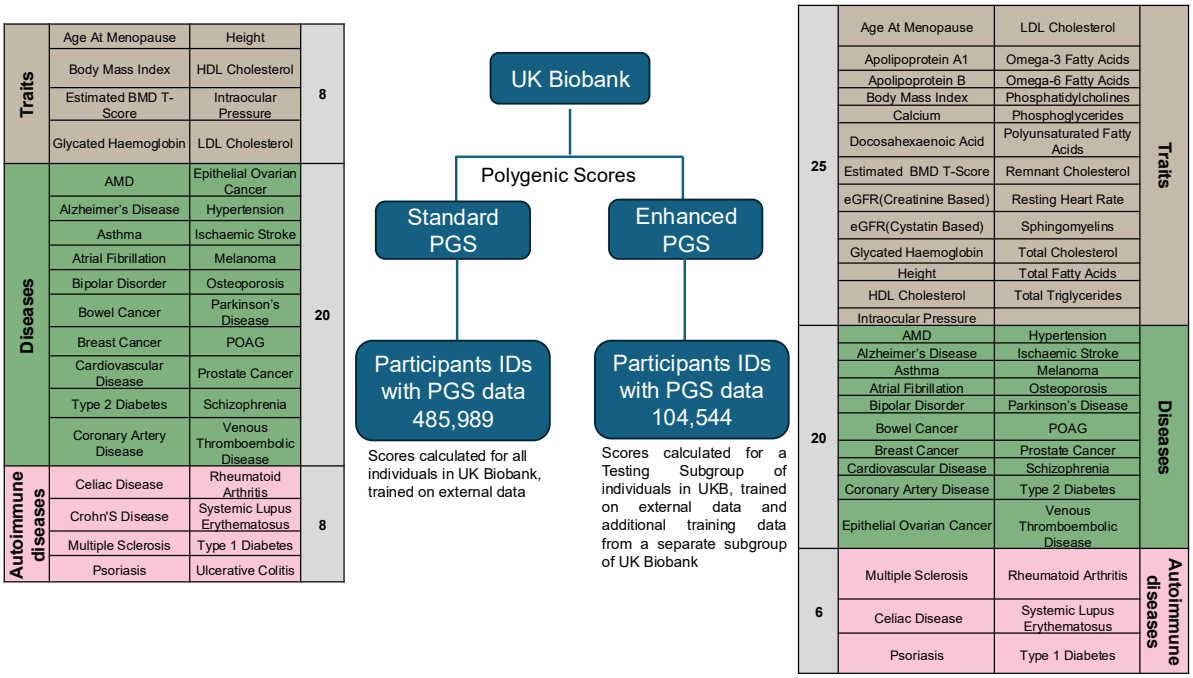


Figure 3.4: The chart displays the PGS datasets overview retrieved from UKB and the number of participants in each set. The table on the left outlines the fields included in the standard PGS dataset (participants: 485,989), which consists of 8 AIDs, 8 traits, and 20 additional diseases. The table on the right presents the fields from the enhanced PGS dataset (Participants: 104,544), which includes 6 AIDs, 20 other diseases, and 25 traits. Color coding is used, with brown representing traits, green indicating diseases, and pink denoting AIDs.

mirrors that observed in the phenotype data. MS shows the highest proportion of unshared samples (93.5% Enhanced, 91.3% Standard), followed by RA with 85.9% in the enhanced and 86.4% in the standard dataset. The lowest representation is observed for CD with 59.3% in the enhanced and 60.9% in the standard dataset, followed closely by UC) with 76% and 75.9% respectively (Figure$\sim$ A.2 and Figure$\sim$ A.3). The highest number of shared samples across diseases occurs between PSO and RA (200; Enhanced, 952; Standard), followed by UC and CD (199; Enhanced, 884; Standard).

Pearson's correlation is used across all traits, diseases, and AIDs. The correlation between diseases, traits, and AIDs is predominantly neutral and positive, with a few negative correlations. The correlations that are not statistically significant (p-value < 0.05) are represented by a cross (x) symbol (Figure$\sim$ 3.5; Enhanced dataset and Figure$\sim$ 3.6; Standard dataset). The strong positive correlations that are statistically significant are predominantly between traits. The correlation more than 0.8 in enhanced dataset is between apolipoprotein B and low density lipoprotein cholesterol (0.84), omega-3 fatty acids and docosahexaenoic acid (0.89), omega-6 fatty acids and phosphatidylcholines (0.80), phosphoglycerides (0.83), polyunsaturated fatty acids (0.87), total fatty acids (0.80), total cholesterol and remnant cholesterol (0.85), sphingomyelins(0.88), polyunsaturated fatty acids and total fatty acids (0.86). In the standard dataset, several autoimmune disease pairs show moderate positive correlations (greater than 0.4),

including T1D and RA (0.40), CED and SLE (0.42), as well as CD and UC (0.54). Notable negative correlations (less than -0.2) are observed between RA and MS (-0.24), and between MS and T1D (-0.22). In the enhanced dataset, positive correlations above 0.2 are seen between T1D and RA (0.43), CED and SLE (0.39), and between MS and SLE (0.21). Slightly, negative correlations are found between MS and RA (-0.28), as well as between CED and RA (-0.19). Diseases showing a correlation greater than 0.5 in both the enhanced and standard datasets include ischaemic stroke and hypertension (0.75; Enhanced set, 0.64; Standard set), coronary artery disease and cardiovascular disease (0.78; Enhanced set, 0.72; Standard set). The correlation more than 0.5 between traits and disease are intraocular pressure and primary open angle glaucoma (0.61, Enhanced set). The disease more than 0.5 is between CD and UC (0.53; Standard set). The disease more than 0.4 is between cardiovascular disease and Ischaemic stroke (0.47; Enhanced set, 0.48; Standard set), T1D and RA (0.43; Enhanced set, 0.40; Standard set), SLE and CED (0.41; Standard set) (Suppl. Table S21 and Suppl. Table S22).

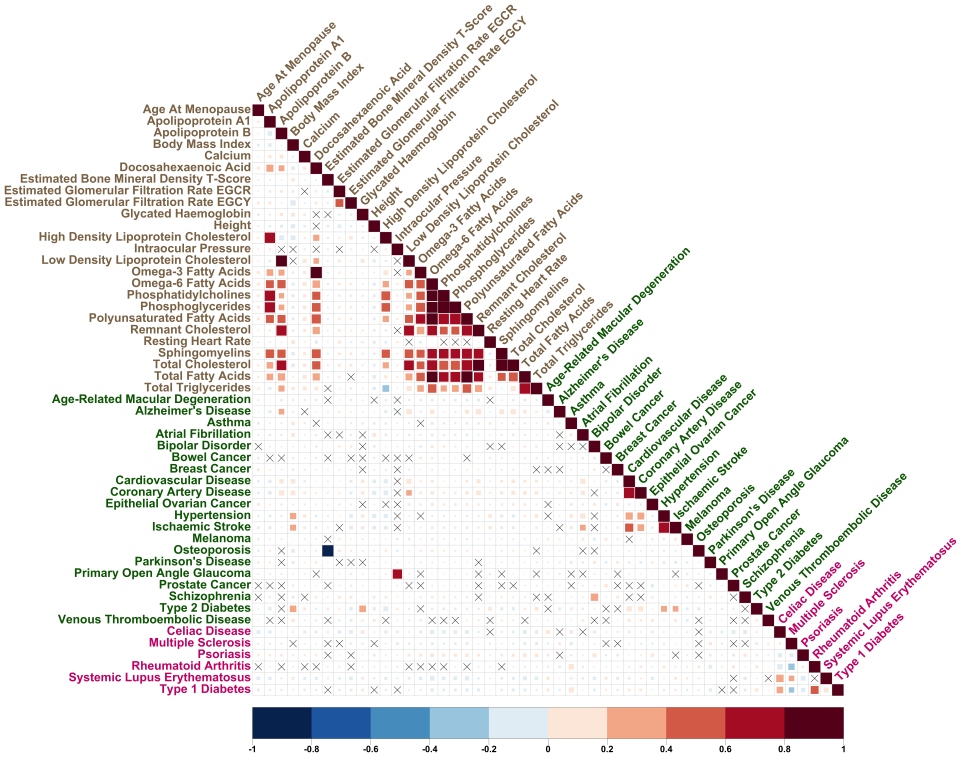


Figure 3.5: Correlation plot to visualize the strength and direction of relationships between various traits, diseases, and AIDs. The analysis utilises PGS derived from the Enhanced dataset. Brown denotes traits, green presents diseases, and pink denotes autoimmune conditions. The plot employs a colour gradient to indicate correlation strength, with red representing positive correlations (more than 0) and blue indicating negative correlations (less than 0). No association is indicated by lighter (near 0 and white) colours. The correlation coefficient, ranging from -1 to 1, quantifies the relationship, with values closer to 1 suggesting a strong positive correlation and values near -1 indicating a strong negative correlation. Each cell in the matrix corresponds to a pairwise comparison, facilitating the identification of highly correlated clusters of variables. The cross mark denotes correlations that are not statistically significant.

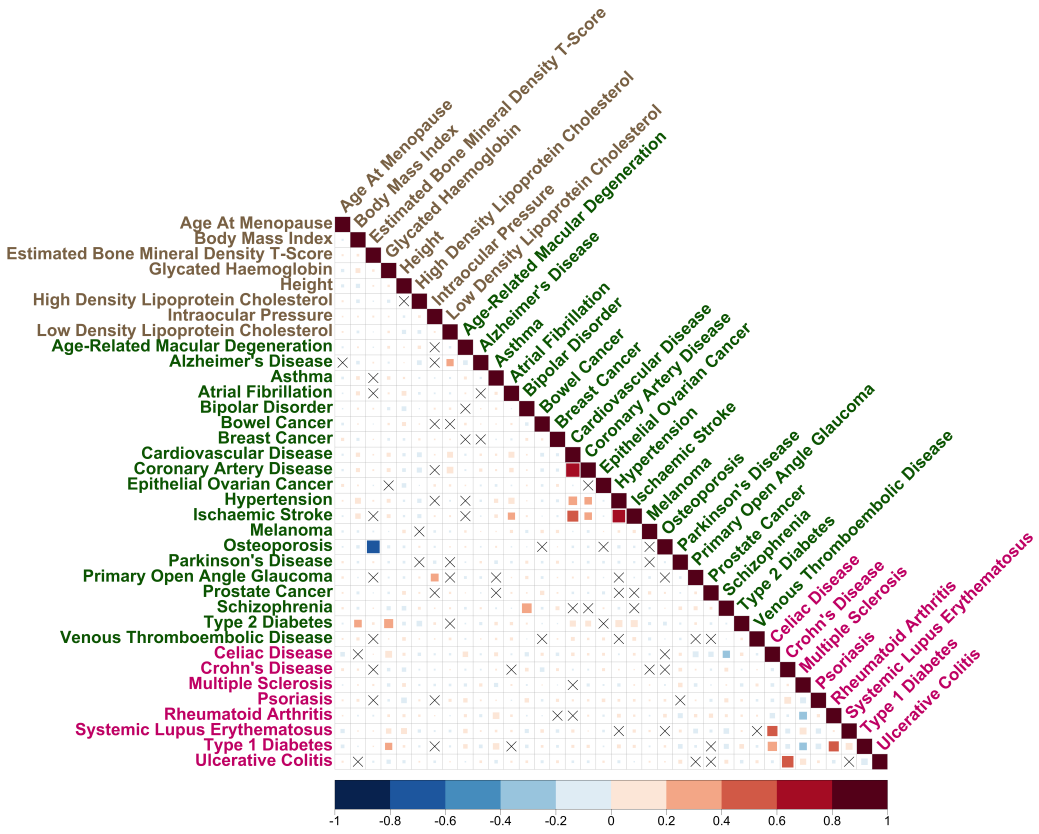


Figure 3.6: Correlation plot visualizes the strength and direction of associations between traits, diseases, and autoimmune conditions using PGS calculated from the Standard dataset. Variables are color-coded by category: traits (brown), diseases (green), and AIDs (pink). Correlation strength is depicted through a gradient, with red signalling positive relationships (more than 0), blue indicating negative relationships (less than 0), and lighter shades (near 0 and white) representing negligible associations. Numerical correlation coefficients (ranging from -1 to 1) quantify these relationships, where values approaching 1 or -1 denote robust positive or negative links, respectively. Each cell in the matrix corresponds to a pairwise comparison, highlighting clusters of interconnected variables. Statistically non-significant correlations are marked with cross symbols to distinguish them from meaningful associations.

Assessment of Genetic Risk Score Distribution Using the Enhanced Dataset

Presenting results on the underlying heterogeneity in genetic risk profiles motivates this section, drawing on PGS from the enhanced dataset among individuals with AIDs, and examining the distribution and potential clustering of PGS to identify distinct risk-score patterns, reveal shared etiological pathways, and inform stratified approaches to disease prediction and intervention.

I compared the mean MS values for PSO and RA (Figure$\sim$ 3.7A and Figure$\sim$ 3.7B). Each dot denotes a risk score for patient sample. A notable disparity exists in the mean values of MS compared to both PSO and RA. A positive trend is observed on the y-axis for the MS group mean, while the mean of PSO and RA exhibits positive trends on the x-axis. The sharp peak for MS along the x-axis and for both PSO and RA along the y-axis, most individuals cluster within the central range. The x-axis density plot for MS peaks slightly to the left, whereas the y-axis density plot for PSO and RA peaks slightly to the right. The PSO and RA distributions on the x-axis span a wider range and extend further to the right. Similarly, the MS distribution on the y-axis exhibits a broader spread extending to the left. A flatter distribution is observed along both axes, with individuals showing higher PGS values clustered in the top-right quadrant in both disease comparisons. For MS and PSO, 47% of the samples are male and 53% female. In the top-left quadrant, MS constitutes the majority (46%), while in the bottom-right and bottom-left quadrants, it accounts for 28%. The bottom-right quadrant is dominated by PSO (38%), and the bottom and top-left quadrants together contain 38% of PSO samples. In the top-right quadrant, MS and PSO are comparably represented, with 26% and 24% of samples, respectively. Female representation is higher in MS-dominated quadrants (75%), while the PSO quadrants exhibit a nearly equal gender ratio (50.4% female). In both diseases, over 50% of the samples are distributed across the four quadrants.

Similarly, among MS and RA samples, 29.8% are male and 70.2% female. In the top-left quadrant, MS again represents the majority with 47% of the samples, while the bottom-right and bottom-left quadrants contain 28%. The bottom-right quadrant is primarily composed of RA samples (37%), and the bottom and top-left quadrants together include 39% of RA samples. The top-right quadrant is shared by both diseases, with 25% of the samples attributed to MS and 23% to RA. High-percentage quadrants for both diseases are predominantly female, with 80% of samples in the MS dominated top-left quadrant and 70% in the RA dominated bottom-right quadrant belonging to females. Across all quadrants, female representation exceeds male in both MS and RA, and more than half of the total samples are distributed throughout different quadrants in each disease comparison.

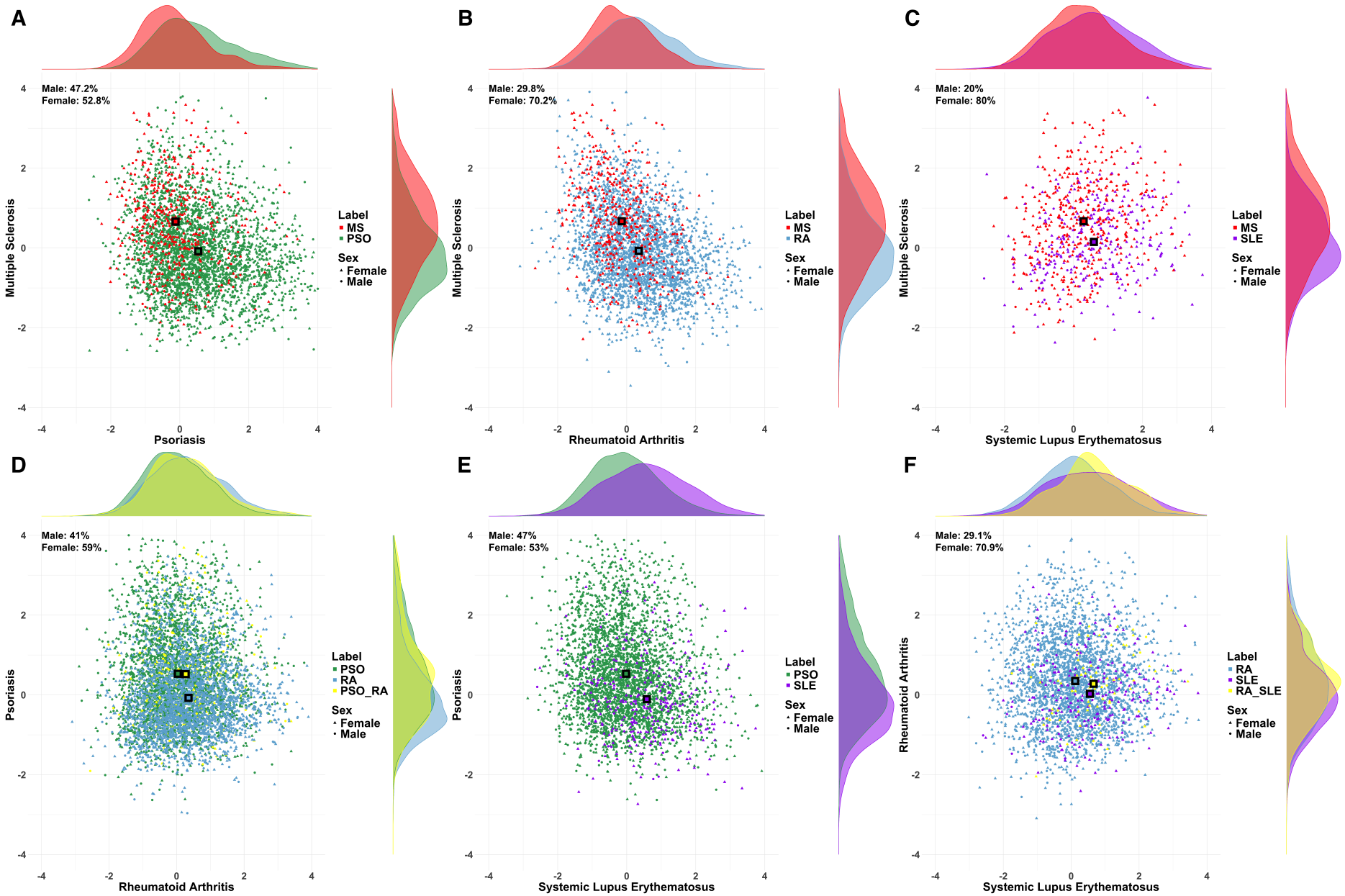


Figure 3.7: The scatter plot depicts the relationship between two diseases through their risk scores, helping to reveal underlying patterns and trends. Each point represents an individual's PGS values from Enhanced dataset, with different colours distinguishing diseases: psoriasis (green; PSO), multiple sclerosis (red; MS), Rheumatoid arthritis (blue; RA), and Systemic lupus erythematosus (purple; SLE). Patients having both diseases are shown in yellow. The range of x and y axis is 4 to -4. The shape of the data points represents sex, with circles for males and triangles for females. For each comparison, the percentages of male and female samples are displayed at the top of each plot. Additionally, density plots along the x and y axes depict the distribution of data points, illustrating how values are spread across each axis. (A) The x-axis represents PSO, while the y-axis represents MS. (B) The x-axis represents RA, while the y-axis represents MS. (C) The x-axis represents SLE, while the y-axis represents MS. (D) The x-axis represents RA, while the y-axis represents PSO, and samples with both diseases are labelled as PSO_RA. (E) The x-axis represents SLE, while the y-axis represents PSO. (F) The x-axis represents SLE, while the y-axis represents RA and samples with both diseases are labelled as RA_SLE.

Further, I investigate SLE and MS, SLE exhibits a more evenly distributed pattern with a slight right-skewed distribution, in contrast to MS, which has a higher density peak on the left side. A relatively small number of individuals with MS exhibit lower values along the x-axis compared to those with SLE. SLE has a peak density in the lower Y-axis range, while MS has a broader spread extending towards higher Y-values. Notably, there is a significant overlap between the two sets of diseases with similar risk values in most individuals. Based on the observed patterns, MS and SLE are positively correlated in a linear fashion. Both on the X-axis and Y-axis, the mean values of MS and SLE exhibit positive trends (Figure$\sim$ 3.7C). In the sample, 20% of the patients are males and 80% are females. There are the highest number of samples for both diseases in the top right quadrant, which contains 46% of samples for MS and 41% for SLE.

In contrast to the PGS indistinguishable diseases mentioned previously, the subsequent pair exhibits divergent PGS profiles when examined individually. However, when these diseases coexist, the risk associated with each disease is amplified (Figure$\sim$ 3.7D). The x-axis density plots display overlapping peaks for individual diseases, while the y-axis plot reveals a broader distribution for the comorbid group. Comorbidity samples predominantly align with the distributions of both individual disease groups, with PGS values falling within the range observed for each condition. Some individuals have relatively lower PGS values for PSO showed by RA peak on the y-axis. The samples of PSO and RA are mostly overlapping; however, the means of the two groups differ. In contrast, samples with PSO exhibit a positive mean for PSO but a negative mean for RA, while the mean of RA is positive for RA but negative for PSO. There is a higher mean in samples with both diseases compared to samples with a single disease. Male and female participants constitute 41% and 59%, respectively; RA constitutes the highest proportion of female samples. Approximately 62% of the PSO samples are found in the top left and right quadrants. Among the 62% of samples, half have high PGS for both diseases; however, half have a high PGS for PSO but a low-risk score for RA. A high percentage of samples are present in the bottom-right quadrant for RA, which is 35%, followed by 26% in the top-right quadrant. The majority of samples with both diseases belong to the top-right quadrant (33%), followed by the top-left quadrant (27%) and the bottom-right quadrant (24%).

The comparison of PSO and SLE samples presents orthogonal mean positions, with each showing elevated scores on a distinct axis while remaining baseline on the other. The PSO, shows a peak towards the lower end of the x-axis, most individuals with this disease have relatively lower PGS values for SLE. Its distribution appears more concentrated, with a narrower spread of PGS values among these samples (Figure$\sim$ 3.7E). In contrast, the SLE distribution is broader and shifted toward higher PGS values, with a peak slightly to the right. The PSO distribution on the y-axis displays a

wider spread, with a flatter shape toward the upper end and the presence of individuals with higher PGS values. The SLE peak shows that some individuals have relatively lower PGS values for PSO. In this pair of AIDs analysis, the samples consisted of 47% male and 53% female participants. For PSO, the distribution is relatively balanced between males and females, with both sexes showing an even spread across all quadrants. The top-left quadrant shows the highest overall representation (35%), with PSO samples primarily aligning with higher Y-axis values. In contrast, SLE exhibits a strong female bias, with 96% of the top-right quadrant samples. SLE samples are predominantly concentrated in the bottom-right quadrant (44%), with a tendency toward higher X-axis values. While PSO shows a more even distribution across quadrants, SLE displays a more polarized pattern, particularly influenced by female samples.
The blue and purple curves correspond to samples with RA and SLE, respectively, while the yellow curve represents individuals diagnosed with both diseases (samples with comorbidity) (Figure$\sim$ 3.7F). In the horizontal density plot (x-axis), the blue curve shows a moderate peak towards the lower end, most samples with RA cluster at lower x-axis values. The SLE curve has a wider distribution and greater variability in PGS along this axis. In the vertical density plot, RA peaks lower on the y-axis; samples typically have lower y-axis values. On the other hand, the SLE shows a slight skew towards higher values. The RA_SLE curve, representing samples having both the diseases, peaks around the center on both axes, these individuals with both diseases tend to have intermediate PGS values. In this pair of AID analysis, the samples consisted of 29.1% male and 70% female participants. For RA, and SLE exhibits a strong female bias, with 70%; top-left and 92%; bottom-right quadrants. Notably, the top-right quadrant has the highest overall representation 32% for RA, 33% for SLE. Each disease shows a distribution across different quadrants and displays a more polarized pattern, particularly influenced by female samples. In comorbid samples for both sexes together the majority of cases concentrated in the top-right quadrant at 57%. Despite the fact that the mean of the comorbidity group is higher than the mean of SLE samples and less than mean of RA samples.

3.2.3 Identifying Variants in the Unrelated White British UK Biobank Cohort and Mapping Associated Genes to Pathways

Identification of Significant Autoimmune Disease Variants and Their Functional Annotations

Elucidating the genetic variants that contribute to autoimmune disease encourages this analysis, drawing on the White British unrelated subgroup dataset, and employing VEP-based annotations to prioritize variants and their associated genes, revealing po-

tential mechanistic insights into autoimmune pathology.
The variant information retrieved from Zenodo contains high number of variants for each AID Table$\sim$ 3.4. Regarding genetic data, the total number of SNPs per disease ranges from approximately 13,000,000 to 13,700,000. RA has the highest number of variants (13,628,693), followed closely by PSO (13,628,692) and UC (13,628,549). On the lower end, SLE has the fewest variants (13,181,991), followed by T1D (13,544,480) and MS (13,600,935). CD and CED fall in the mid-range with 13,618,476 and 13,626,2-45 variants, respectively. The number of significant variants decreases after applying the GWAS significance threshold $P < -\log_{10}(5 \times 10^{-8})$ and excluding the HLA region and its surrounding areas. The filtration and analysis steps are mentioned in method (2.2.3) section. Variants meeting conventional significance criteria are extracted and their effect sizes (β) and standard errors detailed in Suppl. Table S23. The CED category exhibits the highest number of significant variants (4,228), followed by PSO (813), T1D (549) and SLE (411). Conversely, MS (10), RA (28), UC (174), and CD (187) have the lowest number of significant variants (Suppl. Table S24). The VEP output categorizes significant variants into novel and previously reported groups, identifying 39 novel variants in CED, 14 in T1D, 5 in PSO, 3 in SLE, and 1 in RA, and assigns each variant to its corresponding genes (Suppl. Table S24).

Following filtration on VEP output by common variants and a 500 kb, the greatest numbers of genes are identified for PSO (15) and CED (12), and T1D (7), while CD and RA yielded the fewest (3 each), with SLE and UC intermediate at 4 and 6 genes, respectively. After 500 kb filtration of rare variants, CED yielded the greatest number of associated genes (11), followed by SLE and T1D (9). PSO and MS each identified 6 genes, while CD identified 3, and UC and RA each identified 4. The greatest number of genes with high and moderate impact variants is observed in CED (23), followed by T1D (9), SLE (7), and PSO (6). The fewest are found in RA (1), CD (2), and UC (3). The list of genes identified by each filtering criterion across all AIDs is presented in Table 3.5, with detailed information provided in Suppl. Table S25.

Identification of Shared Genes Across Autoimmune Diseases

Identifying genes shared across multiple AIDs inspires this comprehensive comparison of gene sets related to variants selected by common, rare, and impact-based annotations from VEP output, and comparing these gene sets to pinpoint those consistently emerging across conditions, thereby offering insight into core genetic drivers of autoimmunity. There are 14 genes identified as common to the AIDs. Of these, 11 genes are related to impactful variants, 2 carry common variants, and 1 possesses a rare variant (Figure$\sim$ 3.8).

Table 3.4: The table summarizes the number of variants used in UKB PGS analyses. The total number of variants/SNPs is remarkably consistent, ranging from approximately 13.18 million to about 13.63 million across AIDs. At a stringent p-value cut-off of 0.1, each disease has roughly 2.7 million variants, whereas at the more permissive 0.5 threshold, that count rises to around 9.2 million, representing a roughly three to four-fold increase in variant numbers. Differences between diseases at each threshold are minimal, underscoring a uniform distribution of association signal strength across AID.

Autoimmune Diseases	**Abbreviation**	**<0.1**	**<0.5**	**Total**
Celiac disease	CED	2,754,875	9,199,456	13,626,245
Crohn's disease	CD	2,753,381	9,238,161	13,618,476
Multiple sclerosis	MS	2,760,611	9,294,278	13,600,935
Psoriasis	PSO	2,741,332	9,181,598	13,628,692
Rheumatoid arthritis	RA	2,749,319	9,191,652	13,628,693
Systemic lupus erythematosus	SLE	2,718,449	9,205,720	13,181,991
Type 1 diabetes	T1D	2,726,237	9,280,203	13,544,480
Ulcerative colitis	UC	2,749,703	9,188,061	13,628,549

One gene carrying rare variants is commonly identified between T1D and CED. ZNF391 is linked to variant rs67859638 in CED, which shows 2/2 association and trait in 2 studies, no autoimmune disease is associated with this variant. In T1D ZNF391 is associated with rs112328672, this variant has no recorded associations or studies along with no other AID association. Both variant is located on chromosome 6, resides in an intron and annotated with MODIFIER impact. On GWAS catalog ZNF391 gene is credited with 19 association and 13 traits in 17 studies with no recorded associations with AIDs (Suppl. Table S26).

Two genes carrying three common variants are identified between three autoimmune diseases (RA, T1D, and CED). Variant rs10774624 appears twice, in CED and in T1D, each time with 34 association records over 26 traits in 33 studies and additional links to RA and vitiligo. Both variants are MODIFIER map to an intronic/non-coding-transcript region of LINC02356 on chromosome 12, and LINC02356 has 75/53 associations/traits from 64 studies with reported links to RA and vitiligo. Variant rs1230666 is listed for RA with 9/8 associations/traits in 9 studies and no other autoimmune links; it lies in an intron of MAGI3 with MODIFIRE impact on chromosome 1. MAGI3 has no recorded associations or studies on GWAS catalog for rs1230666 variant. rs72687973 for T1D with no association or study counts and no additional AID links. It is an intronic MODIFIER variant in MAGI3 (chromosome 1), for which MAGI3 is reported with 80/41 associations/traits across 49 studies and links to CD, autoimmune thyroid disease, RA, T1D and SLE.

Table 3.5: List of genes identified using filtration based on Impact, Common variants, and Rare variants

Disease	Common Variants	Rare Variants	Impact
Celiac disease	ANKS1A, CMAHP, DCDC2, FYCO1, H3C7, MAPK14, PTPRK, PRSS16, LPP, LINC01934, LINC02356, LINC02828	ANKS1A, CCR3, C6orf62, DCDC2, FANCE, IL12A-AS1, NCF2, SLC17A1, SLC26A8, SOD1-DT, ZNF391	BLTP3A, BTN1A1, BTN2A1, BTN2A2, BTN3A1, BTN3A2, CARMIL1, FANCE, H1-1, H1-6, HFE, KIAA0319, MMEL1, NCF2, POM121L2, PRSS16, SH2B3, SLC17A1, SLC17A2, SLC17A3, SLC17A4, TDP2, ZNF322
Crohn's disease	C1orf141, IL10, NOD2	CDKN2B, IP6K3, TPBGL-AS1	IL23R, NOD2
Multiple sclerosis	-	DPYD, PCSK9, SIM1, SLC7A3, SPOPL, TCP10L	-
Psoriasis	BTN2A2, DENND1B, ETS1, FAP, GPX4, LCE3D, LINC02202, POM121L6P, REL-DT, RNU6-147P, RUNX3, SP140L, TNFAIP3, TNIP1, TYK2	C1orf141, DIP2A, LINC00240, OLFM2, RNF145, TYK2	AP1M2, LCE3D, POM121L2, RUNX3, TRAF3IP2, TYK2
Rheumatoid arthritis	ANKRD55, IKZF1, MAGI3	RSBN1, H2AC10P, LINC00649, TACC2	PTPN22
Systemic lupus erythematosus	PIK3C2B, RUNX1, SCGN, ZNF391	BTN1A1, CARMIL1, FMNL2, FNDC3B, LINC00578, SEMA4B, TRIM38, VCPIP1, VN1R11P	BTN1A1, BTN2A1, BTN3A1, BTN3A2, H1-1, SLC17A2, ZNF322
Type 1 diabetes	GRIP1, H2AC5P, IGF2, LINC02356, MAGI3, NAA25, VN1R13P	CACNA2D3, DNAAF11, H2BC6, JUN-DT, RNF169, SCGN, UBAC2, UNC79, ZNF391	BTN1A1, BTN2A1, BTN3A1, DCLRE1B, INS, PTPN22, SH2B3, TH, ZNF322
Ulcerative colitis	GPR35, IL23R, LINC03112, NR5A2, RORC, SLC26A3	E2F5, IL23R, LINC01435, LINC01832	FCGR2A, IL23R, LINC01475

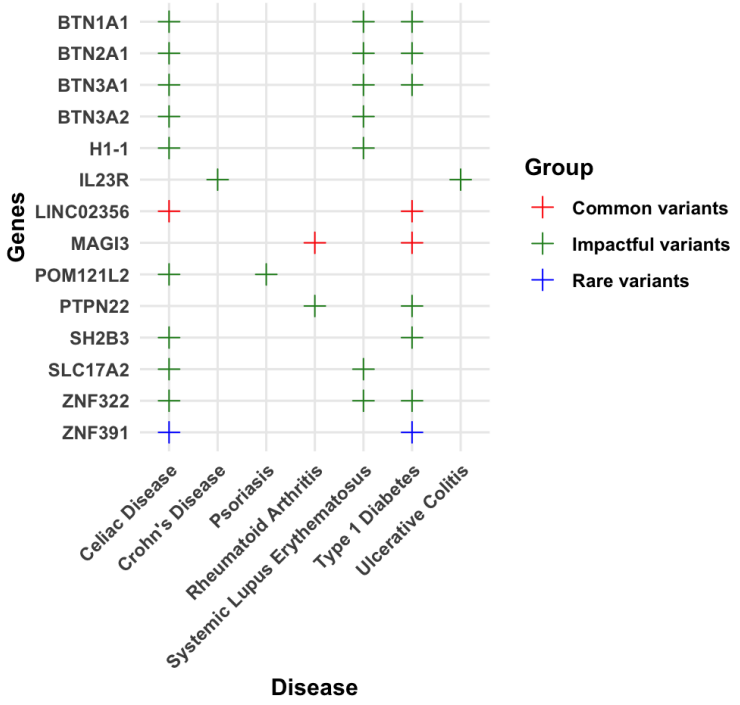


Figure 3.8: The plot depicts genes shared among autoimmune diseases, with each gene color-coded by variant category: red for common variants, green for high-impact variants, and blue for rare variants. The x-axis lists the AIDs, and the y-axis displays the common genes identified across the different AIDs. There are 14 genes identified as common to the AIDs. Of these, 11 genes are related to impactful variants, 2 carry common variants, and 1 possesses a rare variant.

The Impact-filtered output comprises 21 unique variants distributed across 11 genes on chromosomes 1, 6, 11 and 12 (Suppl. Table S26). The gene BTN1A1 (chromosome 6) and its variants show limited but notable overlap across CED, SLE, and T1D. The rs35555795 variant, a moderate-impact missense substitution, is the only one identified in all three diseases, supported by a single association from one study, with no additional AID links. BTN1A1, at the gene level, 37 association records covering 25 distinct traits across 34 studies in SLE and T1D. However, in CED there are 40 association with 27 traits across 37 studies, yet it shows no broader connections to autoimmune conditions. Two other moderate-impact missense variants in BTN1A1, rs3736781 and rs9393728, are each associated with CED in one study and share the same gene-level metrics (40 associations across 27 traits from 37 studies), also without further AID links.

BTN2A1 identified commonly in three autoimmune diseases: CED, SLE, and T1D on chromosome 6. Variant rs13195401 appears consistently across all three diseases with 8 associations across 6 traits reported in 8 studies and shows no additional AID associations, although gene-level data suggests potential links to RA and T1D. Variant rs13195402, also common to all three conditions, displays slightly higher associations for CED (9 associations across 9 traits) and shows reported links to CD and IBD at the variant and gene levels. rs13195402 associated with SLE and T1D has 6 associations

across 6 traits in 6 studies with no AID association mentioned. Variants rs13195509 and rs3734542 follow similar patterns, appearing across all diseases with a declining number of study associations (3 and 1, respectively), and no direct additional AID associations, although gene-level links to RA and T1D persist. Variant rs3734543 has no reported association data at the variant level but includes links to external databases (e.g., NCBI), and maintains gene-level autoimmune associations. Across all entries, BTN2A1consistently shows a high number of trait associations ranging from 43/32 to 49/38 (associations/traits) and 40 to 46 studies along with recurring AID links at the gene level, primarily involving RA and T1D.

BTN3A1 gene is common between three CED, SLE, and T1D. For all three conditions, rs41266839 is a moderate-impact missense variant located in BTN3A1, which maps to chromosome 6 for CED and SLE, and chromosome 11 for T1D. The variant is consistently reported across 7 studies, with 7 associations, none of these have AID associations. BTN3A1 is reported for 30 associations across 22 traits for Celiac Disease (CED), 27/17 for SLE, and 28/17 for T1D with further association with RA and T1D in different studies.

rs13216828, rs71557335, and rs9358936 variants are associated with BTN3A2 and commonly identified in CED and SLE, all located on chromosome 6. rs13216828 and rs9358936 identified as moderate-impact missense variants, and rs71557335 classified as a high-impact splice donor variant. While no variant-level association counts, study counts, or other AID links are specified for the individual variant. For CED, BTN3A2 is associated with 136 associations across 76 traits from 114 studies for rs13216828 and rs9358936, and 138 associations across 76 traits from 114 studies for rs9358936; similarly, for SLE, BTN3A2 records 127 associations across 70 traits from 105 studies for all three SNPs, with no additional AID links reported at the gene level. Notably, BTN3A2 is linked to RA in GWAS catalog.

POM121L2 is common between CED and PSO. For CED, two variants are listed: rs16897515 and rs2235233, both classified as moderate-impact missense variants associated with POM121L2. Variant rs16897515 is supported by 4 association records across 4 traits and studies, and is also linked to other autoimmune diseases such as RA and T1D. In contrast, rs2235233 has no available data for specific associations, studies, or mapped genes, although it shares gene-level associations with 69 traits across 38 studies, with additional autoimmune connections to RA and T1D. For PSO, variant rs41269255 in POM121L2 gene is also a moderate-impact missense change, supported by 2 association records from 2 studies, but without known links to other autoimmune diseases. Nevertheless, the gene as a whole is associated with 80 traits across 41 studies and linked to multiple autoimmune conditions.

H1-1 (chromosome 6) is identified in both CED and SLE via the same variant (rs16891235), supported by 3 studies per disease. Although H1-1 is annotated with 3 studies with

3 traits, it shows no broader AID connections on variant and gene level. IL23R (chromosome 1), by contrast, is implicated in both CD and UC through rs11209026 (23 associations across 20 studies) and displays broad AID overlap. IL23R is connected to a range of inflammatory and autoimmune conditions, including PSO, ankylosing Spondylitis, RA, and IBD reported in 71 studies with 29 different traits. PTPN22, also on chromosome 1, demonstrates the highest degree of autoimmune involvement. It is linked to both RA and T1D through rs2476601, with over 100 variant-level associations in 102 studies and 48 traits. PTPN22 association include SLE, Graves' disease, Hashimoto thyroiditis, and myasthenia gravis. On GWAS catalog PTPN22 is associated with 54 traits in 109 studies. SH2B3 (chromosome 12) also displays associated with both CED and T1D through rs3184504, a variant with high study and association (more than 250) counts. SH2B3 further connects to other AIDs such as Autoimmune Hepatitis, MS, and Autoimmune Thyroid Disease. At the gene level, a single association is reported in one study in the context of CED. In contrast, for T1D, the GWAS catalog documents 395 associations involving 201 distinct traits across 362 studies. In contrast, SLC17A2 and ZNF322 (both on chromosome 6) are observed in CED, SLE, and T1D, but lack specific variant-level association and study. However, SLC17A2, show associations in both CED (70 associations across 47 traits, 69 studies) and SLE (63 associations across 42 traits, 62 studies), with additional autoimmune disease links including PSO and SLE itself. ZNF322 also demonstrates notable involvement, particularly in CED (84 associations across 45 traits, 65 studies) and T1D (71 associations across 38 traits, 55 studies), while for SLE no association and studies in available on GWAS catalog. The shared genes between AIDs are shown in Figure$\sim$ 3.8.

Network Propagation Identifies Key Biological Pathways related to Common Identified Genes

Autoimmune diseases frequently share genetic risk factors that converge on shared biological processes. However, the specific pathways mediating these overlaps remain elusive. Network propagation disseminates signals from identified disease-associated genes across interaction networks, facilitating the identification of pivotal biological pathways that underpin the shared pathophysiology of multiple autoimmune conditions.

Sharpening the shared genetic signal across autoimmune diseases motivates network propagation on genes common to several such diseases, drawing on existing interaction data, the initial overlapping genes, and hallmark gene sets from the Molecular Signatures Database (MSigDB); network propagation then expands the overlap into a broader disease module by adding neighbouring genes likely acting in the same pathways and, aided by the hallmark gene sets, exposes shared pathways or key hub

proteins. Table 3.8 lists the genes supplied as seed nodes for the network-propagation analysis, and Figure 3.9 depicts the Hallmark gene-set propagation results.

Among the 50 evaluated pathways , CED shows positive scores for 42 pathways (pathways related to immune and inflammatory signalling, cell-cycle control and proliferative/oncogenic programmes, metabolic and bio-energetic pathways, stress-adaptation and damage-response modules, developmental and lineage-specification pathways, cellular architecture and polarity), the highly positive (more than 0.50) pathways are related to Proliferation, Immune and Development and negative scores for 7 pathways (core energy metabolism, prosthetic-group/iron handling, oxygen-sensing response, lineage differentiation, secretory machinery, oxidative-stress defence). CD records 32 positive and 18 negative pathways, positive assignments include pathway related to hormone responses, metabolism, immune and inflammatory signaling, growth-factor and oncogenic signaling, Genome Integrity and Stress, Differentiation and Morphogenesis; all remaining pathways are negative mostly related to cell architecture and polarity, cell cycle and proliferation, signal transduction. PSO exhibits positive scores for 9 pathways related to cell cycle and proliferation, genome integrity and stress, secretory pathways, differentiation and morphogenesis and negative for 35 pathways mostly related to immune and inflammatory, metabolism and detoxification, signaling and development, cell cycle and stress response. RA displays 22 positive pathways related to signaling and development, immune and inflammation, tissue structure and function, cell cycle, stress and metabolism and 25 negative pathways related to cell cycle and genome integrity, metabolism and homeostasis, signaling and stress response, differentiation and tissue function. Similarly, SLE has positive status for 39 pathways, highest positive (more than 0.55) are metabolic pathway. SLE has negative status for 11, which includes signaling and development, cell cycle and proliferation, stress and quality control, secretion and specialized function. T1D yields 42 positive pathways related to metabolism and homeostasis, immune and vascular, signaling and stress, structure and differentiation, there are 8 negative pathways mainly related to cell cycle and proliferation, genome integrity and repair, proteostasis and stress response, differentiation and tissue function. UC registers positive for 31 pathways immune and inflammation, metabolism and homeostasis, cell cycle and stress response, development and differentiation and negative for 18 pathways including signaling, development, cell cycle, proliferation, metabolism, detoxification, structure, secretion and stress response. Details of the network propagation scores are provided in Suppl. Table S27.

Pathway enrichment analysis reveals both highly shared and distinct biological mechanisms across autoimmune diseases. The IL6-JAK-STAT3 signaling pathway is highly

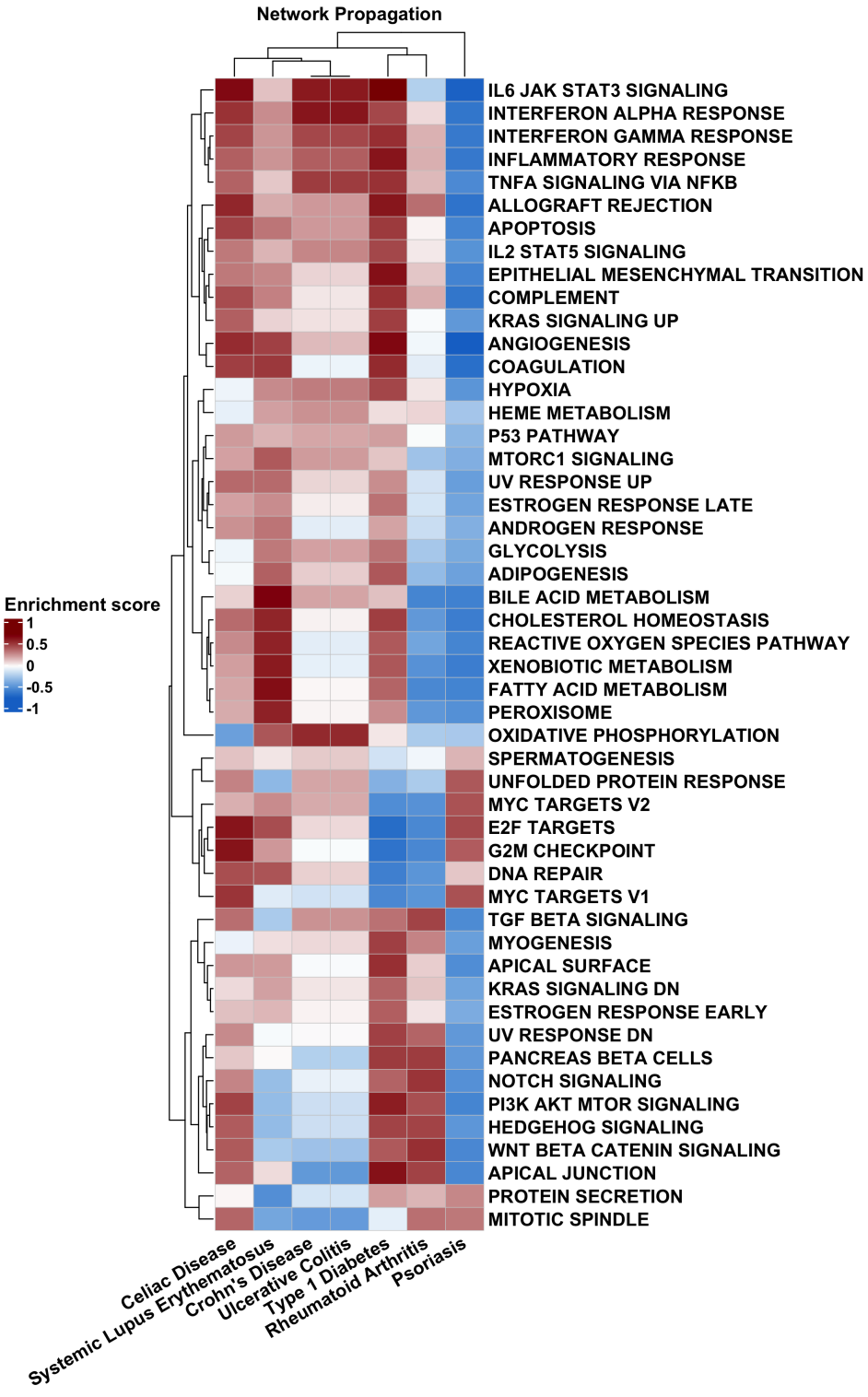


Figure 3.9: The heatmap plot visualizes the results of network propagation, emphasizing associations between pathways and diseases. This analysis is based on genes identified through impact filtration. Rows correspond to pathways, while columns represent diseases. Hierarchical clustering is applied, with dendrograms on the top and left sides to reveal relationships between diseases and pathways. The colour gradient reflects the intensity of variant effects on pathways, with enrichment scores ranging from -1 to 1. Negatively enriched pathways are less influenced by variants in genes, whereas positively enriched pathways are more likely to play a crucial role in the disease process.

upregulated in CED, CD, T1D, and UC, but highly downregulated in PSO. Cell proliferation pathways (G2M checkpoint and E2F targets) are highly activated in CED but suppressed in T1D. PSO is characterized by highly downregulation of immune, metabolic, and developmental pathways, while T1D shows activation across several pathways, including immune, cellular component and developmental pathways.
Subsequently focusing on, 15 pathways related to autoimmune diseases classified into the following categories: immune response, metabolism, developmental processes, signal transduction, cellular components, and fundamental biological pathways. Genes related to CED is positive for 13 pathways allograft rejection, androgen response, angiogenesis, apical junction, apoptosis, bile acid metabolism, cholesterol homeostasis, coagulation, IL6 jak stat3 signalling, inflammatory response, interferon gamma response, TGF-beta signalling and the reactive oxygen species pathway. CED genes are negatively enriched for adipogenesis and oxidative phosphorylation. CD shows 10 positive assignments in adipogenesis, allograft rejection, angiogenesis, apoptosis, bile acid metabolism, IL6 jak stat3 signalling, inflammatory response, interferon gamma response, TGF-beta signalling and oxidative phosphorylation. There are 5 negatively enriched pathways (androgen response, apical junction, cholesterol homeostasis, coagulation, reactive oxygen species) are related to genes identified for CD. PSO records two high negative values (angiogenesis, IL6 jak stat3 signalling) and negative status for the remaining 13 pathways, with no positive enrichment. RA has 5 positively enriched pathways (allograft rejection, apical junction, inflammatory response, interferon gamma response, TGF-beta signalling) and 10 negatively enriched pathways. Systemic lupus erythematosus is positive for 14 pathways (all except TGF-beta signalling) and negative for that single pathway. T1D is positive for 14 pathways, including one high-positive (IL-6 jak stat3 signalling), with no negative enrichment. UC registers 10 positive pathways (adipogenesis, allograft rejection, angiogenesis, apoptosis, bile acid metabolism, IL-6 jak stat3 signalling, inflammatory response, interferon gamma response, TGF-beta signalling, oxidative phosphorylation) and 5 negatively enriched pathways (androgen response, apical junction, cholesterol homeostasis, coagulation, reactive oxygen species).

3.3 Comorbidity Patterns in Pemphigus Across Racial and Ethnic Groups Using TriNetX

Understanding disease susceptibility in individuals with pemphigus across different race and ethnic groups motivates this comorbidity focused analysis, utilizing patient-level data from the TriNetX (TNX) platform, and presenting the results of the comorbidity focused analysis to illuminate patterns of susceptibility. This part of study en-

compasses 92 autoimmune diseases classified by their ICD-10-CM codes, as derived from Samuels et al., [245] (Suppl. Table 28). Pemphigus exhibited higher prevalence in women (56%) compared to men (43%). The majority are white (52%), while only a small portion of the cohort identified as Black/African American, Asian (8% each) or Hispanic (9%). Other ethnicities, including those categorized as just “Unknown” or “Not Hispanic or Latino”, are excluded from comorbidity study.

Subsequently, the risk of developing one or more of 74 distinct autoimmune diseases is evaluated among pemphigus patients. The patient count, percentage of each cohort, and cumulative incidence after a specified time window (from one day after the diagnosis of pemphigus, ICD10CM:L10, to any subsequent time) are outlined in Suppl. Table S29. Bullous pemphigoid had the highest patient count (1535), followed by cicatricial pemphigoid (438), PSO (424), and other RA (356). Conversely, diseases with the smallest patient counts included eosinophilic esophagitis (31), sarcoidosis (36), alopecia areata (39), and immune thrombocytopenic purpura (41) (Suppl. Table S29). Comorbidity is defined through the variable “Cumulative incidence at the end of the time window”. It refers to the number of individuals experiencing a specific event by the conclusion of a defined time frame. In this analysis, this window extends from the day following the index event to any subsequent time thereafter. The findings indicate that bullous pemphigoid (13.57%), other RA (6.77%), PSO (6.44%), and cicatricial pemphigoid (5.59%) exhibit the highest occurrence rates following pemphigus.
Propensity matching is conducted, as outlined in the “Methods” section, to minimize bias arising from confounding factors. Suppl. Table S30 presents the baseline characteristics of patients before and after propensity score matching. Overall, there is considerable variation in ethnicity frequencies across most categories, highlighting the necessity of separately considering these groups. Particularly noteworthy is the observation that Hispanic and White ethnicities represent the youngest and oldest groups, respectively, with the earliest and latest onset of disease (46.9 ± 19.3; $P \leq 0.001$ vs. 62.1 ± 18.3;$P \leq 0.001$), which might stem from the large change in the Hispanics demographics within the United States of America during the last 20 years. Furthermore, although all groups exhibit a sex bias favouring women, the Black group demonstrates the highest female predominance, with 705 patients (65.8%) compared to 367 male patients (34.2%).
The analysis comparing outcomes reveals the risk of developing comorbid AIDs among patients diagnosed with pemphigus, delineated by ethnicity, as presented in Suppl. Table S31. Pemphigus patients of White ethnicity exhibited a significantly elevated risk of developing bullous pemphigoid [HR 145.343; OR 57.347; $P \leq 0.001$]. Moreover, for this group, a significantly increased risk is noted for lichen planus [HR 6.477; OR 5.993; $P \leq 0.001$], systemic involvement of connective tissue [HR 4.614; OR 4.250;

$P \leq 0.001$], discoid lupus erythematosus [HR 3.912; OR 3.632; $P \leq 0.001$], antiphospholipid syndrome [HR 3.287; OR 1.504; P 0.015], and myasthenia gravis [HR 2.893; OR 1.603; P 0.02]. Notably, cicatricial pemphigoid exhibited the most pronounced and significant risk [HR 222.167; OR 20.728; $P \leq 0.001$] for the White group. Additionally, a significant risk of other AIDs, including SLE, neutropenia, Sjögren's syndrome, other RA, and PSO, is identified.

On the other hand, Black individuals diagnosed with pemphigus exhibited a significantly higher risk of developing lichen planus [HR 12.723; OR 1.217; P 0.002] and discoid lupus erythematosus [HR 10.176; OR 1.012; P 0.006] compared to their counterparts in the White group. Notably, among Black pemphigus patients, the highest risk is observed for thyrotoxicosis with diffuse goiter [HR 9.740; OR 1.005; P 0.008] compared to other ethnic groups. Similarly, elevated risks are noted for PSO [HR 3.864; OR 1.760; P 0.004] and RA [HR 2.595; OR 2.076; P 0.014]. Similarly, Hispanic patients diagnosed with pemphigus showed an increased risk of developing lichen planus [HR 16.083; OR 1.415; $P \leq 0.001$] and discoid lupus erythematosus [HR 6.418; OR 1.208; P 0.005]. Elevated risks are also observed for other RA [HR 7.065; OR 3.101; $P \leq 0.001$] and Sjögren's syndrome [HR 4.838; OR 1.313; P 0.007]. Interestingly, among Asian pemphigus patients, the highest risks are identified for SLE [HR 20.460; OR 3.977; $P \leq 0.001$] and antiphospholipid syndrome [HR 16.869; OR 1.513; $P \leq 0.001$]. In contrast, a comparatively lower risk is observed for lichen planus [HR 8.466; OR 1.655; P 0.001], PSO [HR 6.596; OR 4.527; $P \leq 0.001$], and Sjögren's syndrome [HR 6.187; OR 5.690; $P \leq 0.001$] among Asian pemphigus patients. As depicted in Figure$\sim$ 3.10, individuals diagnosed with pemphigus manifested a significantly heightened susceptibility to developing bullous pemphigoid and cicatricial pemphigoid across all ethnic groups. A similar trend is observed for other diseases, with elevated prevalence among pemphigus patients in most subgroups, although statistical significance may not be reached due to limited data. In the White group, only two diseases had fewer than 10 patients available, while the highest number of diseases are observed among Asian patients (15 diseases), with both the Black and Hispanic groups having 12 diseases each. Notably, individuals across all ethnic backgrounds demonstrated a significantly increased likelihood of developing discoid lupus erythematosus, lichen planus, and systemic involvement of connective tissue. Antiphospholipid syndrome exhibited significance among White and Asian populations, whereas PSO showed increased prevalence in all groups, although statistical significance is not reached in Hispanics. Myasthenia gravis is found to be significant in White and Asian populations but not in the other two groups.

In response to expert feedback concerning potential misclassification of pemphigus cases under ICD-10 code L10, I refined the case definition to improve diagnostic specificity. Although this adjustment led to a reduction in the number of identified cases, the

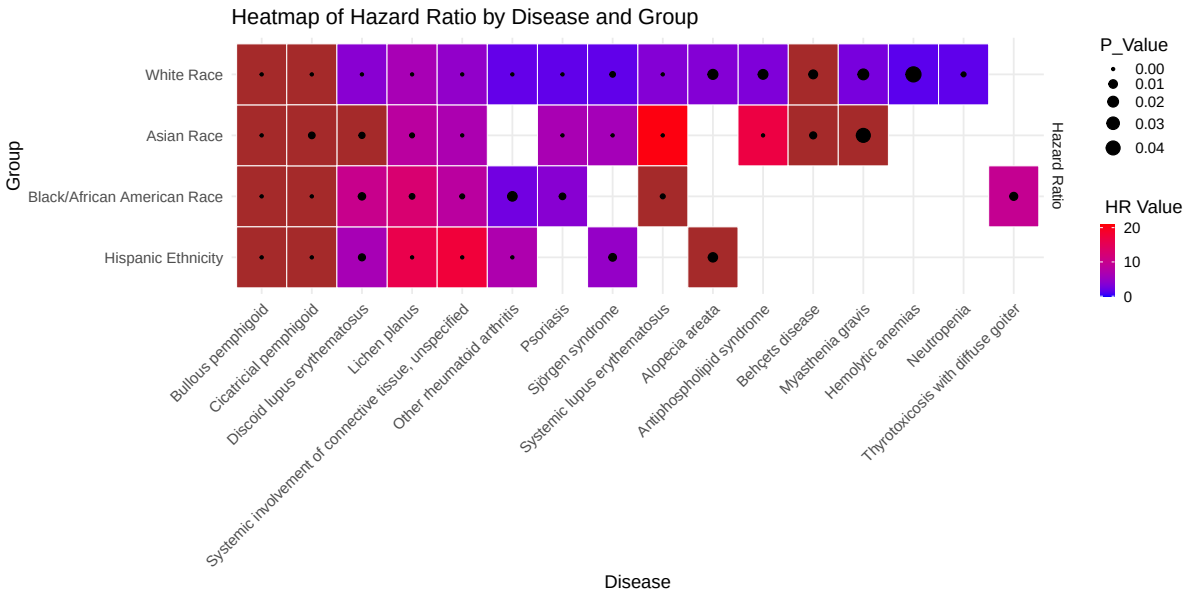


Figure 3.10: Comorbid AIDs with a significantly elevated HR in at least one group. The colour indicates the strength of the increase, the size of the black dot indicates the degree of significance. Only significant relationships are shown.

overall direction of the results remained consistent. Odds ratios across most conditions showed minimal change, though reduced statistical significance is observed in several associations particularly for hazard ratio estimates due to the smaller sample size (Table$\sim$ A.1). Importantly, the correlation between the original and refined OR are strongly positive (Pearson correlation = 0.95, $P < 0.001$), indicating that the selection of a more specific pemphigus definition did not substantially alter the pattern of comorbidities.

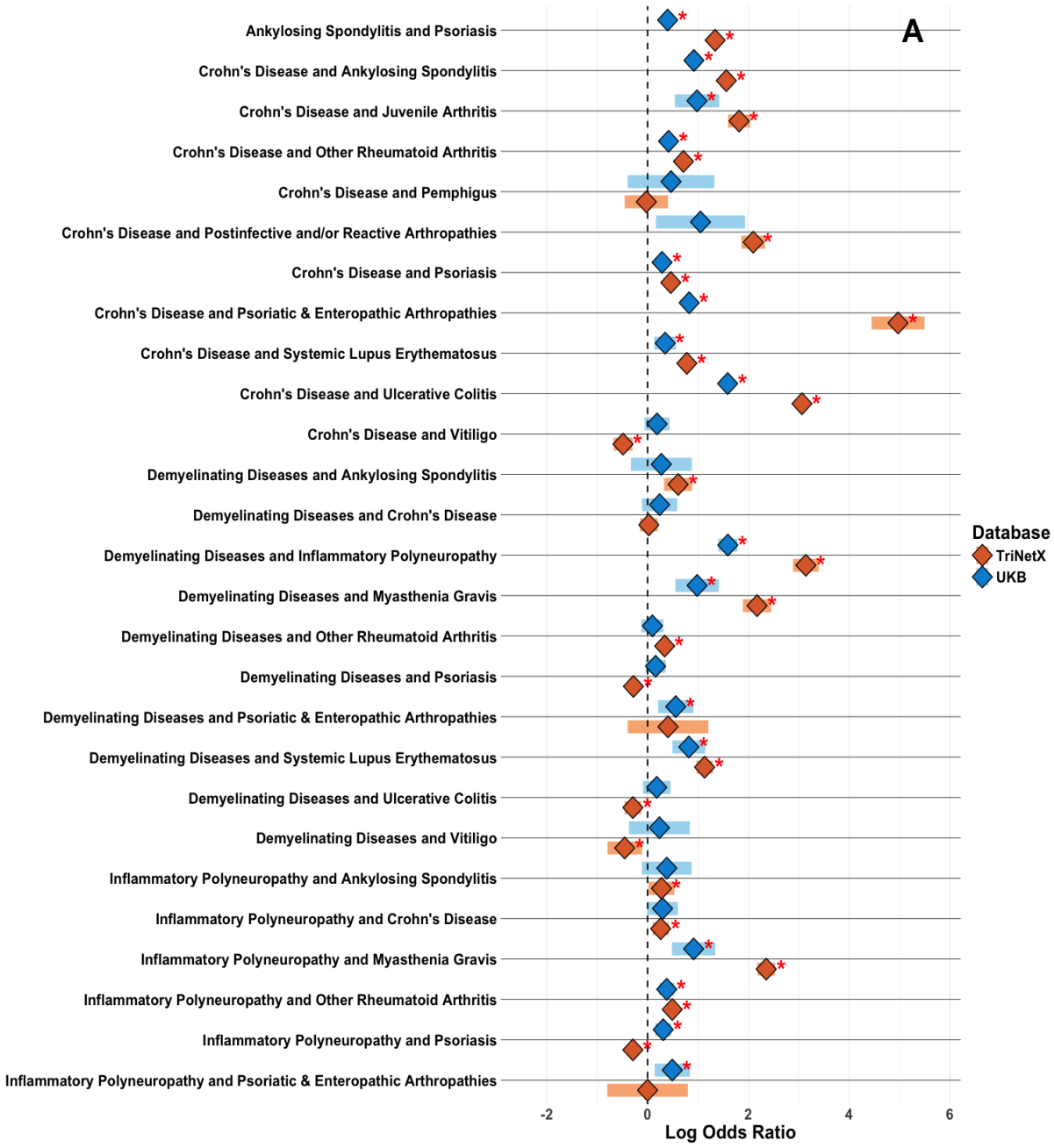
3.4 Analysis of Cross-Database Patterns and Trends of Disease Comorbidity

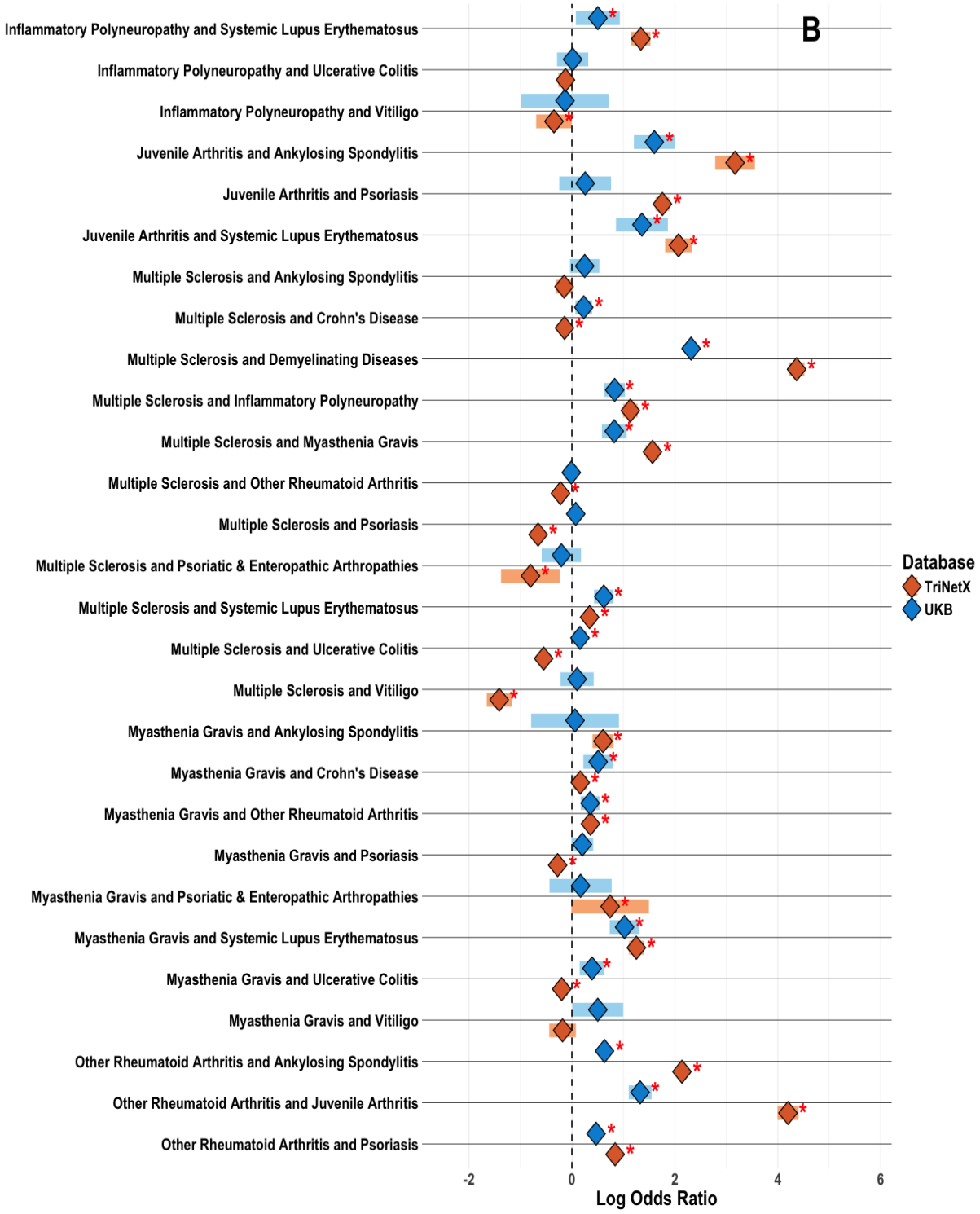
Strengthening comorbidity signal validity and highlighting biases unique to each database motivates this analysis, drawing on recurring disease co-occurrence patterns from the UK Biobank and TriNetX datasets, and comparing odds ratios obtained from samples of both databases to explore replicated and differential co-occurrence of AIDs.
Comorbidity in the TNX and UKB databases is quantified by odds ratios for the AIDs. The terminology used to describe AIDs is largely the same in both databases. However, two diseases have different names in each database. A potential method to standardize disease names across databases is to create a mapping table that links the differing names to a common identifier, such as an ICD code. The ICD-10 codes and names of autoimmune diseases are compared between both databases to ensure consistency and accuracy. In the TNX terminology, myasthenia gravis and en-

teropathic arthropathies are referred to as “myasthenia gravis and other myoneural disorders” and “psoriatic and enteropathic arthropathies,” respectively, on the UKB. However, the ICD10 codes for both datasets are identical, i.e. G70.0 and M07 respectively; Suppl. Table S32. In further analysis the term defined by UKB is used to refer both the diseases. There is a discrepancy on the ICD10 code for one AID; TNX reports M02 for postinfective and reactive arthropathies, whereas UKB reports M03. On TNX, the demographics of the different AID cohorts are analysed to gather information about the minimum, maximum, and mean ages, standard deviations, sex, patient count, and percentage of each cohort (Suppl. Table S33). There is an average age of 65 with the onset of the disease among all age groups, from 1 to 90 years of age. Men had a higher prevalence of ankylosing spondylitis, inflammatory polyneuropathy, postinfective and reactive arthropathies (53%, 54%, 51%, respectively) compared to females. Females are more likely to suffer from AIDs such as juvenile arthritis (70%), MS (73%), demyelinating diseases (66%), RA (72%), and psoriatic and enteropathic arthropathies (63%). In all TNX analyses, Whites have been chosen as the sample population. Then evaluated the comorbidity among 15 different AIDs. Other RA had the highest patient count (395,932) followed by PSO (320,554), UC (177,016), CD (173,481). Conversely, diseases with the smallest patient counts included psoriatic and enteropathic arthropathies (4,527), pemphigus (5,593), postinfective and reactive arthropathies (9,137), juvenile arthritis (29,413) providing insight into the distribution of cases within the studied cohorts.
In this study, 502,371 patients from the UKB were used, of which 229,069 were male and 273,302 were female. PSO is the most prevalent RA (13,110), followed by other RA (12,708), UC (6,522), and postinfective and reactive arthropathies (14) have the fewest patients, followed by pemphigus (51), juvenile arthritis (65), and myasthenia gravis (416). It is ankylosing spondylitis that has the highest male prevalence (68%), and SLE has the highest female prevalence (84%). The age range of these participants is 1934 - 1970. Participant ages were filtered based on age at onset to restrict the age range for each individual, Suppl. Table S34.

There is a similar pattern in the odds ratios obtained from the two databases: extremely high odds ratios ($>$ 20) for few diseases, and statistically significant. The odds ratio of MS with other demyelinating diseases in both datasets is the highest (UKB: 205.2, TNX: 78.5), followed by ankylosing spondylitis with juvenile arthritis (UKB: 39.8, TNX: 23.7), other demyelinating disease with inflammation polyneuropathy (UKB: 39.1, TNX: 23.1), juvenile arthritis and other RA (UKB: 21.1, TNX: 66.5) (Figure$\sim$ 3.11(A, B, C)). The discrepancy of odds ratio has been observed in both the databases in few diseases combination and the odds ratios from TNX is significant and from UKB is non-significant. Dermatological and gastrointestinal diseases

have the highest number of discrepancies in both databases. A few examples include PSO with MS (UKB: 1.18, TNX: 0.52), demyelinating diseases (UKB: 1.44, TNX: 0.74), myasthenia gravis (UKB: 1.59, TNX: 0.75) and vitiligo with CD (UKB: 1.54, TNX: 0.61), demyelinating diseases (UKB: 1.72, TNX: 0.64). The analysis reveals significant differences in disease-pair associations between the two groups, particularly where odds ratios shift directionality. Notably, pairs such as MS with UC (UKB:1.43, TNX: 0.57) with CD (UKB:1.70, TNX: 0.87), inflammatory polyneuropathy with PSO (UKB:2.04, TNX: 0.75), myasthenia gravis with UC (UKB:2.46, TNX: 0.82). Exhibit similar log odds ratios, with statistically significant p-values in both groups. For instance, ankylosing spondylitis with psoriatic and enteropathic arthropathies show a strong association in TNX (23.18) but a weak positive link in UKB (4.42). Consistent positive and significant association is observed across both database in CD with UC (TNX: 21.4; UKB: 38.8), with psoriatic and enteropathic arthropathies (TNX: 144; UKB: 6.68), MS with demyelinating diseases (TNX: 78.5; UKB: 205.2) and RA with juvenile arthritis (TNX: 66.5; UKB: 21.1). Several non-significant associations are observed—including between CD and pemphigus ($p > 0.05$ in both cohorts) and many other pairings likewise fall into this category. The odds ratio information for all disease combination is presented in Suppl. Table S35.


A
Ankylosing Spondylitis and Psoriasis
Crohn's Disease and Ankylosing Spondylitis
Crohn's Disease and Juvenile Arthritis
Crohn's Disease and Other Rheumatoid Arthritis
Crohn's Disease and Pemphigus
Crohn's Disease and Postinfective and/or Reactive Arthropathies
Crohn's Disease and Psoriasis
Crohn's Disease and Psoriatic & Enteropathic Arthropathies
Crohn's Disease and Systemic Lupus Erythematosus
Crohn's Disease and Ulcerative Colitis
Crohn's Disease and Vitiligo
Demyelinating Diseases and Ankylosing Spondylitis
Demyelinating Diseases and Crohn's Disease
Demyelinating Diseases and Inflammatory Polyneuropathy
Demyelinating Diseases and Myasthenia Gravis
Demyelinating Diseases and Other Rheumatoid Arthritis
Demyelinating Diseases and Psoriasis
Demyelinating Diseases and Psoriatic & Enteropathic Arthropathies
Demyelinating Diseases and Systemic Lupus Erythematosus
Demyelinating Diseases and Ulcerative Colitis
Demyelinating Diseases and Vitiligo
Inflammatory Polyneuropathy and Ankylosing Spondylitis
Inflammatory Polyneuropathy and Crohn's Disease
Inflammatory Polyneuropathy and Myasthenia Gravis
Inflammatory Polyneuropathy and Other Rheumatoid Arthritis
Inflammatory Polyneuropathy and Psoriasis
Inflammatory Polyneuropathy and Psoriatic & Enteropathic Arthropathies
-2
0
2
4
6
Log Odds Ratio
Database
TriNetX
UKB


B
Inflammatory Polyneuropathy and Systemic Lupus Erythematosus
Inflammatory Polyneuropathy and Ulcerative Colitis
Inflammatory Polyneuropathy and Vitiligo
Juvenile Arthritis and Ankylosing Spondylitis
Juvenile Arthritis and Psoriasis
Juvenile Arthritis and Systemic Lupus Erythematosus
Multiple Sclerosis and Ankylosing Spondylitis
Multiple Sclerosis and Crohn's Disease
Multiple Sclerosis and Demyelinating Diseases
Multiple Sclerosis and Inflammatory Polyneuropathy
Multiple Sclerosis and Myasthenia Gravis
Multiple Sclerosis and Other Rheumatoid Arthritis
Multiple Sclerosis and Psoriasis
Multiple Sclerosis and Psoriatic & Enteropathic Arthropathies
Multiple Sclerosis and Systemic Lupus Erythematosus
Multiple Sclerosis and Ulcerative Colitis
Multiple Sclerosis and Vitiligo
Myasthenia Gravis and Ankylosing Spondylitis
Myasthenia Gravis and Crohn's Disease
Myasthenia Gravis and Other Rheumatoid Arthritis
Myasthenia Gravis and Psoriasis
Myasthenia Gravis and Psoriatic & Enteropathic Arthropathies
Myasthenia Gravis and Systemic Lupus Erythematosus
Myasthenia Gravis and Ulcerative Colitis
Myasthenia Gravis and Vitiligo
Other Rheumatoid Arthritis and Ankylosing Spondylitis
Other Rheumatoid Arthritis and Juvenile Arthritis
Other Rheumatoid Arthritis and Psoriasis
-2
0
2
4
6
Log Odds Ratio
Database
TriNetX
UKB

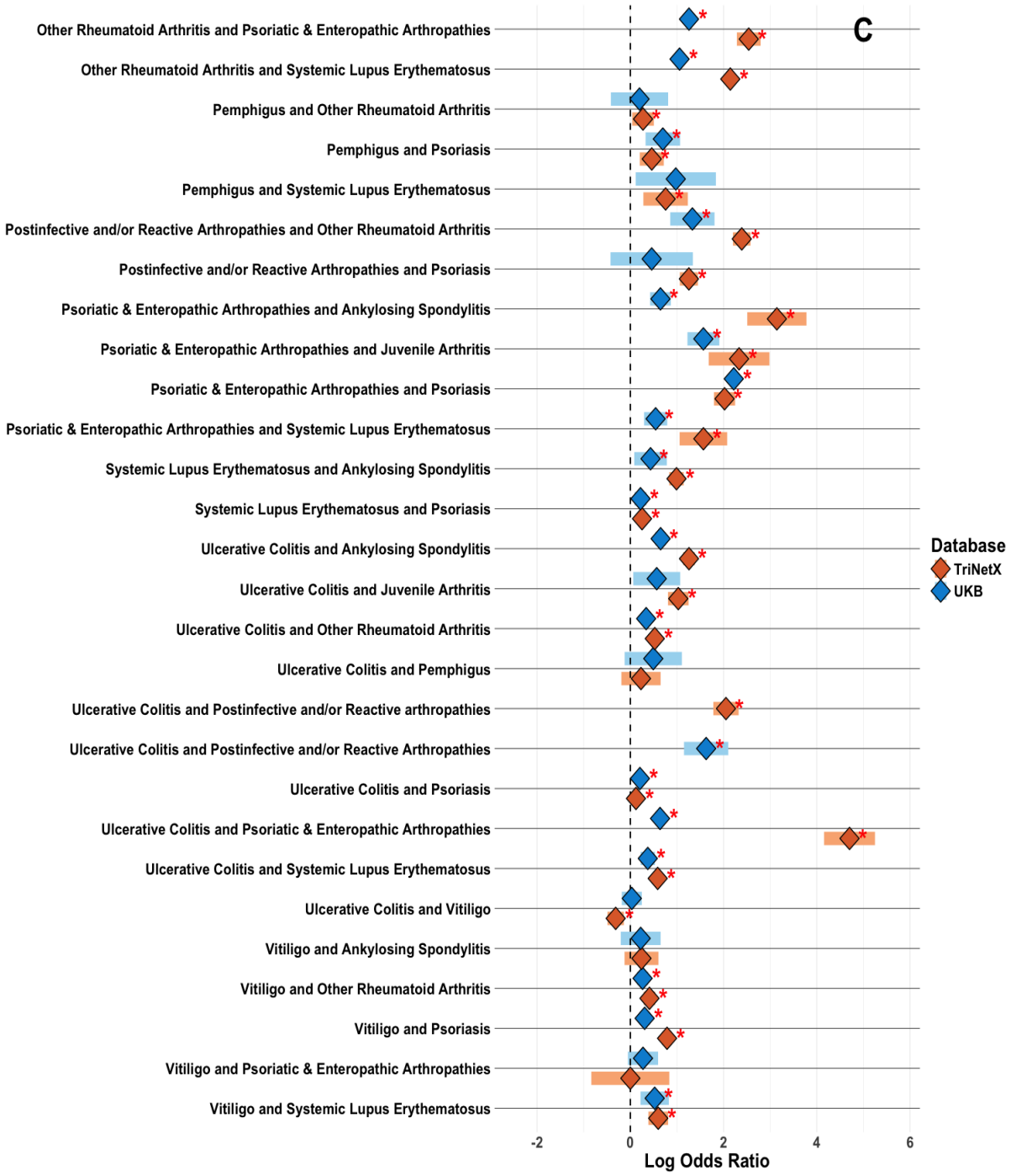


Figure 3.11: Combined figure A, B and C presents log odds ratios (OR) for combinations of diseases, ranging from -2 to 6. Orange bars denote odds ratios obtained from TNX, while blue bars represent those from the UKB. A star symbol denotes statistically significant OR, and the lines surrounding each OR indicate the corresponding confidence intervals

Chapter 4

Discussion

4.1 Interpretation of EFO Guided GWAS PGS and UKB Analyses in Autoimmune Disease Research

Initial survey in this study reveals that a comprehensive catalog of autoimmune Genome-Wide Association Studies (GWAS) and their associations has been compiled over more than a decade of research. There are relatively few polygenic score (PGS) for autoimmune diseases (AIDs), which are very new and have mostly been developed within the last three years. Accordingly, in the polygenic disease genetics field, research efforts go into two related directions: (i) unravelling specific functional effects of variants and (ii) combining effect estimates for a better personalized risk prediction. To achieve the first goal, computer-based approaches are used to correlate risk alleles with molecular traits [258] and identify functional variants via a technique known as fine-mapping [219]. PGS often struggle to reveal how genetic variants actually affect biology, because each variant's impact is very small and subtle. Despite ongoing advances, we are still far from turning these statistical associations into clear molecular mechanisms or effective treatments [259]. To achieve the second goal, PGS continue to be improved, for example by considering rare variants [260] or by including functional information [261]. Optimizing prediction performance is non-trivial, since machine learning models need to be calibrated to generalize to unseen data, i.e., overfitting of training data is prevented. The Area Under the Receiver Operating Characteristic (AUROC) of current AID PGS varies widely (range 0.56-0.99; typically, 0.6-0.8). Further evaluation of PGS in more cohorts and systematic comparisons, facilitated by the PGS catalog, will help gain further insights into PGS predictive performance for individual AIDs. Prospectively, PGS can be amended with other biomedical, clinical and behavioral data. Combined data sources together with recent developments in artificial intelligence promise to improve prediction of disease and personalized treatment options [262]. Finally, study highlight how PGS can be used to investigate interactions

between genetic and environmental factors a particularly relevant approach for AIDs, where environmental triggers play a key role [263].

4.2 Interpretation of Findings from Phenotypic, Genetic, and Risk Score Analyses Using UKB Data

4.2.1 Informative Insights into Phenotypic Convergence and Comorbidity Dynamics in Autoimmune Disorders

Analysis of sample overlap based on phenotypic data reveals both shared and distinct patterns among AIDs. A high proportion of unshared samples underscores the phenotypic and potentially etiological distinctiveness of individual conditions within the cohort. In contrast, sample overlap observed between certain AIDs suggests shared immunological pathways or common pathophysiological mechanisms. These patterns of overlap and exclusivity not only highlight important inter-disease relationships but also contribute to defining the unique genetic and molecular landscapes underlying each autoimmune condition.

Although Crohn's disease (CD) and ulcerative colitis (UC) are typically regarded as mutually exclusive disorders, rare clinical reports document patients exhibiting both conditions. For example, White et al. (1983) [264] detailed an 11 years history of CD and UC in a single 29 years old woman, underscoring their distinct but potentially overlapping immunological bases. Similarly, Castro et al. (1996) [265] reported two preadolescent sisters, one with CD and the other with UC developing their respective diseases within a year. These cases illustrate the complexity of autoimmune presentations and suggest that, while CD and UC are separate entities, shared immunological underpinnings may exist but don't imply the same disorder. The observed overlap of samples between CD and UC cohorts, despite the traditional view that these conditions do not coexist in the same individual at the same time, may arise from several factors. First, both diseases share some common immunopathogenic mechanisms, meaning that certain biomarkers or immunological signatures may be present in both conditions. This can lead to overlapping sample profiles even when the clinical manifestations are distinct. Second, the diagnostic criteria for inflammatory bowel diseases (IBD) can sometimes lead to ambiguous classifications. In clinical practice, a subset of patient is often categorized under "indeterminate colitis" due to overlapping clinical, endoscopic, or histopathological features. This diagnostic uncertainty may contribute to the sample overlap observed between CD and UC cohorts.

The variation in sample sizes across different AIDs within the UK Biobank (UKB) can be influenced by a combination of epidemiological, clinical, and methodological factors.

Foremost, disease prevalence plays a central role, common AIDs such as rheumatoid arthritis (RA) and psoriasis (PSO) are naturally more represented due to their higher incidence in the general population. Although systemic lupus erythematosus (SLE), CD, and multiple sclerosis (MS) are not rare conditions in the general population, they are relatively underrepresented in the UKB. This discrepancy may be attributable to the age range of participants at recruitment (40–69 years), which does not align with the typical age of onset for these conditions, often occurring in early adulthood.

The analysis of odds ratios (OR) among various autoimmune disease pairs demonstrates a predominance of positive and statistically significant associations, indicating substantial comorbidity across these conditions. This pattern suggests shared genetic and immunological mechanisms underlying their co-occurrence. A small number of disease pairs showed non-significant positive associations, likely due to limited sample sizes or distinct etiopathological features. I observed virtually no statistically significant inverse associations. Overall, these results underscore the intertwined nature of autoimmune diseases and the crucial need to account for comorbidity in both research and clinical practice

4.2.2 Interpretation of Correlations and Genetic Risk Patterns in Autoimmune Diseases Using PGS

The examination of correlations between risk scores and disease presence yields critical insights into shared pathophysiological mechanisms. Type 1 diabetes (T1D) and RA, as well as SLE and celiac disease (CED), exhibit statistically significant positive associations, reflecting shared immunological and genetic determinants that contribute to their co-occurrence. These associations, initially identified through clinical phenotype data, are further validated by PGS correlations, underscoring a convergent genetic architecture. The comorbidity between MS and SLE is particularly notable, as both conditions demonstrate a significantly elevated OR in conjunction with a strongly positive PGS correlation. This finding reinforces the hypothesis of a common genetic basis influencing susceptibility to both diseases. Similarly, PSO and RA show a positive OR and a modestly positive PGS correlation, suggesting partial overlap in genetic risk factors, albeit to a lesser extent. In contrast, the relationship between PSO and SLE is marked by a non-significant positive OR but a negative PGS correlation, suggesting that while there may be some overlap in clinical features, their underlying genetic architectures are likely distinct and potentially opposing in effect.

The contrasting risk-score patterns for MS, PSO, SLE, and RA highlight how genetic and environmental factors combine to produce distinct disease profiles. Risk score

observations emphasize the partially shared but largely independent genetic architectures of these autoimmune conditions and agree with previous reports of limited pleiotropy among them. Such distributions visually and quantitatively confirm a modest yet meaningful inverse genetic relationship among these conditions. Similar trend is observed in MS and RA, which show modest positive trends of RA mean along the x-axis and MS mean along y axis. Notably, in the quadrant analysis individuals with higher genetic risk for MS tend to have lower risk for PSO [266] and RA [267], and vice versa, reinforcing the inverse relationship reflected in the overall PGS correlation. In summary, the distribution and orientation of disease-specific risk scores provide evidence of a weak negative genetic correlation, highlighting the distinct risk profiles of MS and RA despite their shared classification as autoimmune diseases.

Although AIDs exhibit opposing trends in their mean PGS reflected by their concentration in opposite quadrants. the fact that over 50% of the samples are distributed across all quadrants for both combination of diseases indicates substantial overlap in individual risk profiles. This suggests that, while there may be modest differences in average genetic risk between the two diseases, the overall genetic architectures are not entirely distinct. The broad distribution of samples implies that many individuals carry PGS that confer susceptibility to more than one autoimmune condition, consistent with the concept of poly-autoimmunity and shared genetic predisposition. Therefore, the quadrant separation of group means highlights general directional trends, but the widespread sample overlap underscores the complexity and partial commonality of autoimmune disease risk, reflecting both shared and disease-specific genetic contributions.

The slightly orthogonal mean of PSO and SLE reveals, minimal difference between their genetic architectures [268]. SLE samples are predominantly clustered in the bottom-right quadrant (44%), suggesting a consistent trend toward higher scores along the X-axis, and reflecting a polarized distribution. In contrast, PSO samples are more evenly distributed across all quadrants, indicating greater variability in individual risk profiles and less directional clustering. This contrast in distribution patterns supports the interpretation that PSO and SLE are genetically divergent, with little shared genetic risk, despite some overlapping clinical features.

The distribution and correlation patterns between SLE and MS suggest a moderate degree of shared genetic risk [269]. The mean of PGS for both diseases trend positively along both the x and y axes, indicating a linear positive correlation between their genetic risk profiles. This relationship is further supported by the quadrant analysis, where the top right quadrant contains the highest concentration of samples for both diseases, reflecting a substantial overlap in individuals with elevated risk for both conditions. MS and SLE differ in their risk score distributions for more than 50% of sample, the alignment of their means and the shared high density region suggest a meaningful genetic correlation and potential comorbidity between the two diseases.

The distribution of PGS between PSO and RA reveals both shared and distinct genetic risk patterns, offering important insights into their comorbidity. Although many samples overlap between the two diseases, the divergence in group means where PSO samples tend to have high PSO risk but low RA risk, and vice versa, suggests partially independent genetic architectures. Notably, individuals with both diseases exhibit higher mean PGS for both conditions, indicating an additive genetic burden associated with co-occurrence. The quadrant distribution further emphasizes this complexity while a significant proportion of PSO samples are found in quadrants associated with both high and low RA risk, RA samples are more concentrated in the bottom-right quadrant, reflecting high RA but low PSO risk. Most co-affected individuals are located in the top-right quadrant, consistent with high genetic risk for both diseases. These findings highlight the heterogeneous nature of genetic risk in autoimmune comorbidity and underscore the importance of considering both shared and disease-specific risk profiles in genetic analyses, reflecting loci that are partly shared yet still distinct across conditions.

The analysis of PGS for SLE and RA reveals that individuals with both diseases tend to have intermediate genetic risk for each [270], suggesting that comorbidity is driven by moderate, overlapping genetic factors rather than extreme risk for either condition alone. SLE shows greater variability in genetic risk, while RA displays a more concentrated distribution. The majority of comorbid cases cluster in the high-risk quadrant for both diseases, indicating a shared genetic contribution. Overall, these findings highlight the complex, partially overlapping genetic architecture underlying RA-SLE comorbidity.

Across every AID pair, more than half of the PGS distributions defy a clear, uniform pattern underscoring the considerable genetic complexity and heterogeneity that characterises autoimmunity. Many AIDs share overlapping genetic risk factors particularly within the HLA region, which can lead to positively correlated or indistinct PGS distributions, masking any divergence seen in specific disease pairs. Comorbidities, misclassification of diagnoses, and non-genetic factors such as environment and epigenetics introduce further variability.

The pronounced female predominance observed in AIDs, with approximately 60% of cases being female, reinforces the well-established role of sex-based biological factors in AIDs susceptibility. This imbalance is particularly relevant for conditions such as SLE and RA, where hormonal influences, especially the immunostimulatory effects of estrogen, contribute to heightened immune reactivity in females [271] [272]. Additionally, the involvement of X-linked genes and the potential for incomplete X-chromosome inactivation may lead to increased expression of immune-related genes, further predisposing women to autoimmune conditions [273]. In contrast, PSO displays a more balanced sex ratio, reflecting a distinct pathogenic profile that may be less influenced

by X-chromosome specific factors [274].

4.2.3 Insights into AIDs Genetic Variants and Their Link to Pathways in the White British Unrelated UK Biobank Subgroup

Interpreting Functional Impact and Genetic Architecture Across Autoimmune Diseases

After removing the HLA region, PSO, T1D and CED retain a high number of risk variants due to their polygenic structure, strong associations with non-HLA variants, and more genetically homogeneous profiles. In contrast, RA, SLE, UC and CD rely more on variants in the HLA region, resulting in fewer detectable associations. This highlights the importance of considering disease-specific genetic architectures when interpreting GWAS and PGS analyses. Excluding the HLA region sharply reduces the number of significant MS variants, underscoring its concentrated genetic risk at that locus. In contrast to other autoimmune diseases with broader non-HLA contributions, MS's architecture is dominated by HLA-associated factors, highlighting antigen presentation as its central pathogenic mechanism. Annotating variants for predicted functional impact using VEP and restricting gene assignments to those within 500 kb enriches for plausible biological effects but excludes many others, especially those in linkage disequilibrium blocks spanning beyond the 500 kb window or representing distal regulatory elements. This dual strategy increases interpretability but reduces the total variant under consideration for both common and rare variants. The enrichment of high and moderate impact variant genes in diseases like CED, T1D, SLE, and PSO reflects a higher functional variant burden, while RA, CD, and UC show fewer such genes, potentially due to different genetic mechanisms.

Implications of Shared and Disease Specific Genes in Autoimmune Pathogenesis

Shared and disease specific genes study investigates the genetic landscape of AIDs by identifying gene variant associations through comprehensive filtering of common, rare, and functionally impactful variants. The relative scarcity of genes defined solely by common or rare alleles underscores the need to analyse the full spectrum of variant frequencies. The 11 of the 14 shared genes are identified on the basis of functional impact indicates that predicted effect size is a more reliable marker of shared autoimmune risk. Finally, identification of a single rare variant shared between T1D and CED, located in a gene not previously associated with autoimmunity in the GWAS Catalog, highlights how infrequent yet functionally important alleles can uncover previously unrecognized mechanisms underlying autoimmune disease. Its low allele frequency

likely limits its detection in large-scale GWAS, which are primarily powered to identify common variants, contributing to the under-representation of such rare variant associations in autoimmune research.

Importantly, this approach led to the identification of both known and novel gene–disease associations. Several genes, such as IL23R, PTPN22, SH2B3, and MAGI3, have been previously reported in the GWAS catalog for their associations with multiple autoimmune conditions, which serves to validate the analytical pipeline and highlight the robustness of the current findings. In contrast, genes such as ZNF322, POM121L2, and LINC02356 emerge as novel candidates may be due to the identification of previously unreported variants. While these genes have been associated with certain AID in prior studies, the current analysis reveals new associations with additional autoimmune conditions, thereby expanding their known disease relevance. Conversely, BTN1A1, H1-1, and ZNF391 have not been previously linked to AIDs in existing literature or GWAS databases; however, this study provides novel evidence supporting their association with AIDs. These novel findings highlight previously underrecognized but functionally important variants such as intronic, non-coding transcript, missense, and splice donor variants that may influence gene expression, splicing, or protein function. Unlike traditional GWAS that prioritize exonic and common variants, this study is impact-based filtering approach enabled the detection of biologically relevant variants with diverse functional effects, offering new insights beyond those captured in standard GWAS catalogs.

In addition, the identification of multiple butyrophilin (BTN) family genes BTN2A1, BTN3A1, BTN3A2, and BTN1A1 across distinct autoimmune diseases extends their known roles in immune regulation and suggests a broader relevance across the autoimmune spectrum than previously appreciated [275]. The detection of genes like MAGI3, which has known associations with some AIDs but not specific variant-level autoimmune links, also reveals the gaps in current GWAS annotations and highlights how functionally annotated variants can bridge these gaps.

Biological Insights from Network Based Pathway Mapping of Common Genes

The observed network propagation pattern reflects how genetic risk variants influence disease-specific biological processes through interconnected functional pathways. Stringent edge filtering, hub removal, Laplacian diffusion, and permutation-based testing largely suppress random noise, meaning the proteins that remain significant are unlikely to be artefactual and instead probably represent true network-level associations. Nevertheless, functional validation or independent datasets will be necessary to confirm biological causality. The predominance of positive enrichment in diseases like CED, SLE, and UC suggests widespread pathway activation, indicating

that disease-associated genes in these conditions are functionally central and propagate influence across a broad network of biological functions particularly immune, developmental, and metabolic systems [276] [277]. This implies a systemic, multi-pathway dysregulation characteristic of these diseases. By contrast, PSO exhibits strong negative enrichment, suggesting its risk genes exert localized or repressive effects. This pattern points to disrupted or silenced pathways rather than overactive ones, in line with previous research [278]. The differential directionality (positive vs. negative scores) likely captures the nature of the disease's pathogenesis whether it involves aberrant activation (e.g., inflammation, proliferation) or functional suppression (e.g., reduced metabolic activity or immune tolerance). Furthermore, shared positive pathways (e.g., IL6-JAK-STAT3 signaling) point to core autoimmune mechanisms, while disease-specific patterns reflect unique pathophysiological features. The network propagation method thus reveals not just which pathways are involved, but how disease genes integrate into the broader biological system either as amplifiers or dampeners of function offering a functional layer to genetic association data.

4.3 Understanding Comorbidity Dynamics in Pemphigus Among Ancestral Groups Using TriNetX

In support of molecular findings, clinical evidence from multiple epidemiological studies further emphasises the interconnected nature of autoimmune disorders. The recurrent presence of certain genes across multiple disease categories underscores their potential as pivotal determinants of disease susceptibility. Overall, the integration of computational impact predictions with allele frequency data from diverse populations supports the existence of a heterogeneous yet interrelated network of genetic variants that warrant further functional and clinical evaluation.

Several investigations have indicated that individuals diagnosed with pemphigus are predisposed to developing concurrent autoimmune disorders, including PSO, various forms of RA, lichen planus, Sjögren syndrome, SLE, and alopecia areata [224] [279]. Comparable patterns are observed within the White group; however, statistical significance is not reached for Sjörgen syndrome and alopecia areata in the Black group, and for other forms of RA and alopecia areata in the Asian group. Similarly, in the Hispanic group, PSO and SLE did not demonstrate significance. Since almost all of these relationships show a consistent trend across all categories, merely not reaching significance in some subgroups, they are plausibly an issue of limited data. While Kridin et al., reported an association between UC and pemphigus, but not CD, the study did not find any correlation between pemphigus and either UC or CD in any ethnic group. Both of these diseases are variants of IBDs and may not even be autoimmune, merely

immune system related [280], and as a result a lack of relationship with pemphigus is quite plausible. Additionally, no associations are found with autoimmune thyroiditis, CED, eosinophilic esophagitis, immune thrombocytopenic purpura, or sarcoidosis. MS is not linked with pemphigus in previous studies, and this analysis revealed the absence of association across all ethnic groups [281]. Particularly, no evidence of anti-correlations or protective effects between the investigated AIDs is found.

These findings hold significant relevance for clinical practice, underscoring the necessity of screening for comorbidities to ensure comprehensive care for pemphigus patients. Many of the most prominent comorbid AIDs affect the skin, which could be overlooked by physicians when a patient already presents with pemphigus. Among the strongest non-pemphigoid diseases are discoid and SLE. Both of these can add further complications that need to be considered, such as photosensitivity, renal or blood disorders and which might require an increase in immunosuppressive drug doses such as prednisone [158]. These two further belong to a large cluster of autoimmune connective tissue diseases associated with pemphigus also including RA, Sjörgen and undifferentiated connective tissue disease, which is named as “Systemic involvement of connective tissue, unspecified” in TriNetX (TNX). An exhaustive list of possible complications and further required treatment is not feasible here, but it includes among others joint problems requiring physiotherapy, dryness of soft tissues which preclude certain medications, stress triggers requiring psychotherapy or complications between the osteoporosis caused by autoimmune symptoms and glucosteroids. Conversely, certain associations, such as those with antiphospholipid syndrome, neutropenia, and thyroid toxicosis, may come as unexpected. To date, evidence of elevated levels of antiphospholipid antibodies in pemphigus patients has been limited to a small study [282]. While prior theories have suggested a potential link between neutropenia and pemphigus treatment with immunosuppressives, investigations have shown a low prevalence [283]. Much of the existing research has primarily focused on single-institutional studies, leaving a gap for a comprehensive, multi-institutional investigation on a global scale. This specific study aimed to fill this gap by conducting a detailed quantitative analysis, with a particular emphasis on the influence of race and ethnicity on the susceptibility of pemphigus patients to secondary AIDs, which may manifest as comorbidities. The investigation thoroughly examined the risk of secondary AIDs among pemphigus patients across diverse racial and ethnic groups. The results consistently demonstrated robust associations between pemphigus and other autoimmune conditions, notably pemphigoid. These findings align with previous research indicating a correlation between various autoimmune disorders [215] [216]. In comparison to another recent review utilizing TNX data [237], The results indicate consistent associations among AIDs, although certain methodological differences are observed. While Kasperskiewicz et al., additionally adjusted for obesity and nicotine dependence and

separately examined associations with bullous Pemphigoid, they analyzed a smaller cohort of pemphigus patients ($n \sim 8600$), a reduced subset of 25 autoimmune disorders, and only differentiated between White and Black/African American racial categories, excluding Asians and Hispanics. Additionally, a minor issue observed in their supplementary materials is the inadequate exclusion of group sizes with $n \leq$10, which could lead to both false positives and false negatives. For instance, in supplementary table 9 of Kasperskiewicz et al. [237], the risk difference (0.272, line 2) may be inflated due to rounding up in the case group, while the risk difference (0.055, line 3) may be deflated due to rounding up in the control group. On line 4, the direction of bias is indeterminable since both sides are rounded. The last case is the most common on TNX, and as a result there is a notable bias towards insignificance for rarer states on the platform if not controlled for. In contrast, this study excludes OR and related measures, such as risk difference, when any value is 10 or less to mitigate potential estimation errors.

4.4 Interpreting Comorbidity Patterns and Underlying Insights Across TNX and UKB Databases

The disparity in terminology and ICD-10 coding between TNX and UKB databases for the same AID poses a critical issue for data standardisation and interoperability. This inconsistency can compromise patient cohort selection and misclassification in research and healthcare applications and can also affect data analysis, research findings, epidemiological studies, and clinical decision-making. The observed differences in the prevalence of AIDs in UKB between males and females within the white population, particularly the higher prevalence of men in ankylosing spondylitis and the increased number of women samples in other demyelinating disease, MS and RA, underscore the substantial role of sex as a biological variable in the development of AIDs. These findings align with established trends in AID epidemiology [284]. Sex hormones, such as oestrogen, can enhance immune responses, potentially increasing the susceptibility of females to autoimmune disorders. Genetic factors, including genes located on the X chromosome, and epigenetic changes mediated by oestrogen may also play a role [284]. It is crucial to acknowledge the potential for diagnosis bias. As Fairweather et al. suggests, the higher prevalence of certain AIDs in women may be attributed to diagnostic biases or fundamental biological differences between males and females. Findings in this study demonstrate robust associations between MS and other demyelinating diseases, as well as substantial connections between ankylosing spondylitis and juvenile arthritis. These findings align with existing research that proposes shared genetic and immunological pathways among these conditions [285]

[286]. The clustering of autoimmune and inflammatory diseases in individuals further supports the hypothesis of overlapping pathophysiological mechanisms, including genetic predisposition, immune dysregulation, and environmental triggers. Discrepancies in odds ratios between the UKB and TNX datasets, exceeding 1 in UKB and below 1 in TNX for certain conditions, notably dermatological and gastrointestinal may reflect differences in population characteristics, diagnostic criteria, or data collection methods. Variability in the clinical presentation and classification of diseases may contribute to these inconsistencies [287]. Additionally, disparities in sample size, study design, and inclusion criteria likely influence the observed differences. Discrepancies in odds ratios (OR $>$ 1) for multi-organ autoimmune diseases significant in one dataset but not in the other, highlight associations that warrant deeper investigation. In the TNX electronic health record database, a larger and more heterogeneous population produces OR $>$ 1 without reaching statistical significance, likely due to unmeasured confounders, variable data granularity, and a higher proportion of female participants. By contrast, the UKB's deeply phenotype cohort yields significant OR $>$ 1 findings through rigorous co-variate adjustment and greater phenotypic consistency; yet some associations remain non-significant because of smaller case counts and stricter inclusion criteria. Differences in disease ascertainment and male to female ratios between TNX and UKB further influence statistical power and significance, especially for conditions with sex biased prevalence. Together, these observations underscore the need for larger, more homogeneous cohorts or meta-analyses that harmonize sample size, population characteristics, and confounding to ensure robust interpretation of epidemiological associations across diverse datasets. Some disease pairs show opposite log odds patterns in TNX versus UKB, appearing protective in TNX but risk enhancing in UKB; yet remain statistically significant in both. This consistency implies that factors like genetic predisposition, environmental exposures, and database-specific practices shape these associations. Again, differences in sample size, inclusion criteria, and healthcare delivery between the two cohorts likely also contribute to these divergent findings. The consistent and significant odds ratio more than one across both databases indicate that certain diseases share common risk factors, reinforcing their biological or epidemiological connections. These findings emphasize the importance of careful interpretation when comparing results across different databases and suggest that aligning methodological approaches could enhance the reliability of association studies. Further investigation is needed to identify the factors contributing to these discrepancies and to develop strategies that improve the consistency of disease-pair analyses across diverse datasets.

Chapter 5

Limitations

This study presents a range of limitations arising from dataset structure, disease representation, analytical methodology, and inherent challenges in autoimmune disease (AID) research. One major limitation is the uneven representation of autoimmune diseases in the UK Biobank (UKB), where case ascertainment depends on a combination of hospital inpatient data, primary care records, and self-reported diagnoses each with varying levels of completeness and diagnostic accuracy. Complex clinical presentations and diagnostic delays in certain autoimmune diseases further risk misclassifications or omission, leading to sample sizes that may not reflect true disease prevalence and potentially limiting statistical power and generalizability. Odds ratio based comorbidity analyses provide valuable signals but are constrained to concurrent disease presence and do not infer temporal ordering or causality. Polygenic score (PGS) correlations indicate genetic overlap between diseases but lack functional and clinical interpretability, as they do not elucidate the underlying biological mechanisms. PGS analyses are further affected by sample size imbalances, such as female predominance in diseases like systemic lupus erythematosus, and overlapping cases across diseases with divergent genetic profiles, which can lead to ambiguous or conflicting risk interpretations. Moreover, PGS is primarily derived from genome-wide association studies (GWAS) based on individuals of European ancestry, limiting cross-population applicability. These scores capture only additive effects of common variants, excluding rare variants, gene-gene interactions, and environmental influences that play critical roles in autoimmune pathogenesis. Limitations in variant level analyses also warrant attention. Variant prioritization relies heavily on in silico functional predictions, which, while informative, require experimental validation to confirm biological relevance. The inclusion of rare variants deepens analysis but introduces statistical challenges due to low allele frequencies, making replication and interpretation difficult. Some newly associated genes are poorly characterized in autoimmune contexts, and the functional significance of these associations remains speculative without further evidence. Additionally, the GWAS Catalog only archives genome-wide signifi-

cant variants from curated studies, meaning many potentially relevant variants including those in strong linkage disequilibrium with reported variants are excluded if not submitted or statistically prioritized. Enrichment analyses using hallmark gene sets from MSigDB, although biologically meaningful, may overlook context-dependent or disease-specific functions and assume static gene-to-pathway relationships, ignoring factors like tissue specificity, gene regulation dynamics, or disease progression stages. The interpretation of positive and negative enrichment scores, although suggestive of functional propagation, does not provide clear insights into causal directionality or regulatory mechanisms. Network propagation methods further present interpretability challenges, while effective for identifying enriched pathways, they do not pinpoint which genes drive enrichment and may amplify signals due to network topology rather than biological relevance, especially when using large or overlapping gene sets. The comorbidity analysis of pemphigus introduces several unique limitations. Relying on electronic health records introduces the risk of diagnostic inaccuracy, particularly in distinguishing pemphigus from clinically similar conditions such as pemphigoid. The analysis is inherently limited to individuals with access to healthcare, potentially biasing results and reducing generalizability. Data access constraints in TriNetX (TNX) prevent raw data analysis, restricting the application of advanced statistical methods and exploratory modeling. The large sample size necessitated stratification by race and ethnicity for accurate propensity score matching; however, individuals labeled as "unknown" for key demographic variables were excluded, and subgroup sizes for non-White populations were often too small for reliable statistics. Additionally, TNX applies rounding practices to protect patient privacy, sometimes leading to control groups with zero recorded cases, making odds ratio or hazard ratio calculation infeasible despite apparent significance. These limitations, combined with the inherent constraints of observational data, complicate efforts to establish causality even when attempts were made to exclude reverse temporal relationships. Comparative comorbidity analyses between TNX and UKB face further challenges. Differences in ICD-10 coding conventions and terminology between the two platforms hinder data standardization and reduce cross-database interoperability. The inability to access raw TNX data limits reproducibility, as future queries cannot retrieve identical cohorts. The UKB cohort itself reflects a healthy volunteer bias participants are generally healthier, better educated, and less socioeconomically deprived than the broader population further limiting external generalizability. Baseline measures and self-reported data were collected only once, providing a static view rather than a longitudinal perspective. Recall dependent and touchscreen-based responses may introduce misclassification, and substantial missing phenotype data reduce analytical depth. Finally, linking UKB to additional health or administrative records remains constrained by regulatory and ethical challenges, limiting the enrichment of clinical phenotypes and the integration of real-world

health outcomes. Collectively, these limitations highlight the need for more diverse, longitudinal, and deeply phenotyped datasets, standardized coding frameworks, and robust experimental follow-up to validate findings and better understand AID etiology and comorbidity.

Despite these limitations, this study provides several valuable contributions to the understanding of AID etiology and comorbidity. Leveraging large-scale datasets from the UKB and TNX, it offers one of the most comprehensive assessments to date of genetic overlap and clinical co-occurrence patterns across multiple autoimmune conditions. The study highlights key signals of polygenic sharing and potential biological pathways, serving as a valuable resource for hypothesis generation and future mechanistic studies. It also identifies methodological challenges and biases inherent to current data sources, offering a critical framework that can guide the design of more diverse, longitudinal, and deeply phenotype cohorts. This work lays the groundwork for improved data integration strategies and more precise modeling in future autoimmune disease research. Ultimately, it contributes to a growing body of knowledge that seeks to unravel the complex genetic and clinical architecture of autoimmune diseases.

Chapter 6

Conclusion

Autoimmune diseases share a fundamental breach of immune tolerance, yet the forces that drive their diverse clinical phenotypes vary in scope and intensity. By integrating four complementary lines of evidence (i) a synthesis of genetic databases, (ii) cross-disease polygenic-risk analyses, (iii) an ancestry-stratified study of pemphigus comorbidity, and (iv) a cross-cohort comparison of electronic health-record resources this work draws primary conclusions. This study highlights the evolving landscape of autoimmune disease genetics, focussing on both the functional impact of genetic variants and improving polygenic risk prediction. Genome-wide association studies have provided a comprehensive catalog of variant associations, but translating these findings into mechanistic understanding and clinical applications remains incomplete. Polygenic scores offer promising avenues for personalised risk prediction, but current models show variable performance and require refinement. As machine learning and integrative approaches advance, combining genetic, clinical, and environmental data could enhance disease prediction and inform precision medicine strategies for autoimmune diseases. The review also emphasises the potential of PGS as tools for risk estimation and dissecting gene-environment interactions, crucial for understanding autoimmune disease aetiology. The potential of polygenic risk leads this study to applies these approaches to real-world genomic data and detailed analyses of autoimmune samples using sample overlap, odds ratios, and polygenic score distributions reveal genetic and phenotypic relationships among autoimmune diseases. While many diseases have distinct phenotypes and genotypes, some share immunopathological pathways. Consistent positive odds ratios reinforce the interconnected nature of autoimmune diseases and emphasise the importance of comorbidity in clinical and research settings. Correlations between polygenic score and disease presence show shared and divergent genetic architectures. Some disease pairs have moderate to strong positive correlations, while others like multiple sclerosis and rheumatoid arthritis have inverse associations, reflecting genetic antagonism. Quadrant-based risk score distributions confirm these trends but also highlight individual variability, with over 50%

of samples falling into all risk quadrants. This supports the concept of polyautoimmunity and suggests the complex, non-binary nature of genetic risk across diseases. Pairwise autoimmune disease risk score analyses reveal shared and distinct genetic patterns similar as correlation study. Individuals with multiple autoimmune diseases often have additive risk burdens, suggesting that co-occurrence is not driven by extreme risk in one disease alone, but by moderate overlapping susceptibility. While directional trends in group means indicate genetic divergence or convergence, the wide risk score distributions suggest partial genetic overlap, not complete genetic overlap, which explains the comorbidity. The $\sim$ 60% female predominance across autoimmune diseases aligns with evidence of sex-based biological influence in autoimmune susceptibility. This consistent pattern emphasises the need to consider sex as a critical covariate in genetic analysis and clinical practice. Differential reliance on the HLA region across autoimmune diseases highlights the need to account for disease-specific genetic architectures in association studies. Diseases like psoriasis, type 1 diabetes, and celiac disease retain strong signals after HLA exclusion, indicating robust non-HLA contributions. In contrast, conditions like rheumatoid arthritis, systemic lupus erythematosus, and ulcerative colitis depend more on HLA-linked variants, which may limit risk detection when removed. These differences underscore the importance of tailored analytic strategies in autoimmune genetics. Integrated analysis of common, rare, and impactful variants identifies known and novel gene-disease associations, demonstrating the value of comprehensive variant filtering. Shared rare variants, especially those not previously catalogued, expand understanding of genetic contributions to autoimmunity and highlight the limitations of genome-wide association studies focused primarily on common variants. The identification of functionally impactful variants in previously unassociated genes offers new avenues for biological investigation and therapeutic development. Network propagation analysis reveals how disease-associated variants influence interconnected biological pathways. Conditions like systemic lupus erythematosus, ulcerative colitis, and celiac disease show broad pathway activation, indicating systemic dysregulation across immune, metabolic, and developmental systems. In contrast, diseases like psoriasis exhibit negative enrichment, suggesting localised or repressive regulatory dynamics. These findings offer a functional framework for interpreting genetic risk and highlight the utility of systems-level analyses in understanding autoimmune pathogenesis. This analysis extends the earlier investigation to rarer autoimmune phenotypes and evaluates comorbidity patterns by race and ancestry. It confirms and extends earlier observations that pemphigus frequently co-occurs with various autoimmune diseases. Robust associations are detected with pemphigoid, discoid lupus erythematosus, systemic lupus erythematosus, rheumatoid arthritis subtypes, and Sjögren syndrome across all racial and ethnic groups. Where statistical significance is not achieved for specific disorders in certain cohorts, point estimates

parallel those in White patients, suggesting limited subgroup sample sizes rather than true absence of effect. Conversely, the consistent lack of association with inflammatory bowel diseases, autoimmune thyroiditis, celiac disease, multiple sclerosis, and other conditions suggests that pemphigus shares pathogenic pathways with only a subset of immune-mediated diseases. This highlights the need for systematic screening of pemphigus patients for connective-tissue and cutaneous autoimmune comorbidities, as these may be masked by the primary disease. Identifying concomitant disorders like systemic lupus erythematosus is essential for adjusting immunosuppressive regimens and monitoring for organ-specific complications. Methodologically, the study improves on prior single-centre reports by using a large, racially diverse cohort and enforcing strict cell-count to minimise spurious odds-ratio estimates. Unexpected links between antiphospholipid syndrome, neutropenia, and thyroid toxicosis warrant replication in independent datasets and mechanistic exploration to determine if they reflect biological overlap or treatment-related phenomena. This study also refines the comorbidity landscape of pemphigus, highlights ethnicity-specific data gaps, and provides an evidence base for integrated multidisciplinary care. Future work should combine longitudinal follow-up with molecular profiling to uncover shared immunological drivers and guide precision-based management of pemphigus and allied autoimmune disorders. This cross-cohort evaluation focuses on the practical challenges of reproducing signals across heterogeneous health-record resources, beyond the earlier synthesis of shared autoimmune architecture and comorbidity. Comparing the TriNetX and UK Biobank cohorts reveals three key themes. First, inconsistent terminology and ICD-10 coding for identical autoimmune diseases hinder data interoperability, cohort definition, and multi-database analyses. Standardised disease ontologies are crucial for reliable cross-study comparisons and clinical translation. Second, sex is a critical biological variable. In the white UK Biobank population, women predominate in demyelinating disorders like multiple sclerosis and rheumatoid arthritis, while men are more affected by ankylosing spondylitis. These patterns align with established epidemiology and likely reflect immune-endocrine mechanisms, X-chromosome gene dosage, and diagnostic bias. Third, disease-pair associations differ by dataset. Strong links between multiple sclerosis and demyelinating diseases, and between ankylosing spondylitis and juvenile arthritis, are replicated, supporting shared genetic and immunological pathways. However, odds-ratio magnitude and significance often diverge for dermatological and gastrointestinal conditions due to differences in case definitions, population structure, sex ratios, and covariate adjustment. In TriNetX, broad heterogeneity and larger sample sizes yield elevated but non-significant odds ratios, while the more deeply phenotype UK Biobank cohort produces significant findings despite smaller case counts. The opposite log-odds directions of the same disease pair in both resources highlight the influence of database-specific factors like healthcare

delivery context, environmental exposures, and data granularity. These observations suggest harmonised coding practices, larger ancestry-balanced cohorts, and meta-analytic strategies that normalise for sampling and methodological differences. Standardisation is crucial for interpreting epidemiological signals and translating them into robust insights on autoimmune disease aetiology and comorbidity.

Overall, this study offers a comprehensive, data-standardised resource for autoimmune research by curating key genetic databases, mapping polygenic-risk profiles that reveal both shared and opposing liabilities across common autoimmune diseases, charting ancestry-specific comorbidity patterns for the rare blistering disorder pemphigus, and showing that harmonised diagnostic coding is crucial for reproducible cross-cohort analyses. Taken together, these contributions link genetic architecture, comorbidity, and methodological rigour in an ancestry-aware framework that clarifies where autoimmune diseases converge and diverge, thereby laying the groundwork for precision-medicine strategies that can leverage common pathways while respecting disease-specific nuances.

Funding
I gratefully acknowledge funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (EXC 22167-390884018) and by Research Training Group 2633 "Autoimmune Pre-Disease," project A9 (GRK 2633/9 – 2021). Open Access funding was enabled and organized by Projekt DEAL.

Supplementary material
The supplementary material can be found online on Zenodo at:
https://doi.org/10.5281/zenodo.15478926

Generative-AI Disclosure Statement
In this thesis writing I made limited, careful use of the generative AI system solely to suggest alternative phrasings that sharpened the clarity and concision of few sentences that I had already drafted, it has helped me to debugs the R codes for plot generation and conversion of abstract into German language. All AI outputs were critically reviewed, fact-checked, and re-written as needed; the reasoning, structure, and final wording remain entirely my own.

Abbreviations

Abbreviation	Definition
AF	Allele frequency
AID	Autoimmune disease
ASPR	Age-standardised prevalence rate
AUROC	Area Under the Receiver-Operating Characteristic Curve
CD	Crohn's disease
CI	Confidence interval
ClinGen	Clinical Genome Resource
CNV	Copy number variations
dbSNP	Database for single nucleotide polymorphisms
DMARD	Disease-modifying antirheumatic drugs
EBI	European Bioinformatics Institute
EFO	Experimental Factor Ontology
EMBL	European Molecular Biology Laboratory
GDPR	General Data Protection Regulation
gnomAD	Genome Aggregation Database
GWAS	Genome-wide association studies
HIPAA	Health Insurance Portability and Accountability Act
HLA	Human leukocyte antigen
HOCs	Health Care organizations
HR	Hazard ratio
IBD	Inflammatory bowel disease
ICD	International Classification of Diseases
Indels	Insertions/deletions
MAS	Multiple autoimmune syndrome
MHC	Major histocompatibility complex
MS	Multiple sclerosis
NHGRI	National Human Genome Research Institute
OR	Odds ratio
PCA	Principal component analysis

PE	Pemphigus erythematosus
PF	Pemphigus foliaceus
PRS	Polygenic Risk Scores
PGS	Polygenic score
PPM	PGS Performance Metric
PV	Pemphigus vulgaris
RA	Rheumatoid arthritis
ROS	Reactive oxygen species
RR	Risk ratio
RWR	Random Walk with Restart
SLE	Systemic lupus erythematosus
SNV	Single nucleotide variants
SNP	Single nucleotide polymorphism
T1D	Type 1 diabetes
TNX	TriNetX
UC	Ulcerative colitis
UCSC	University of California Santa Cruz
UKB	UK Biobank
UTR	Untranslated region
UTRs	Untranslated regions
VEP	Variant effect predictor
WES	Whole exome sequencing
WGS	Whole genome sequencing
WHO	World Health Organization

Appendix A

Appendix

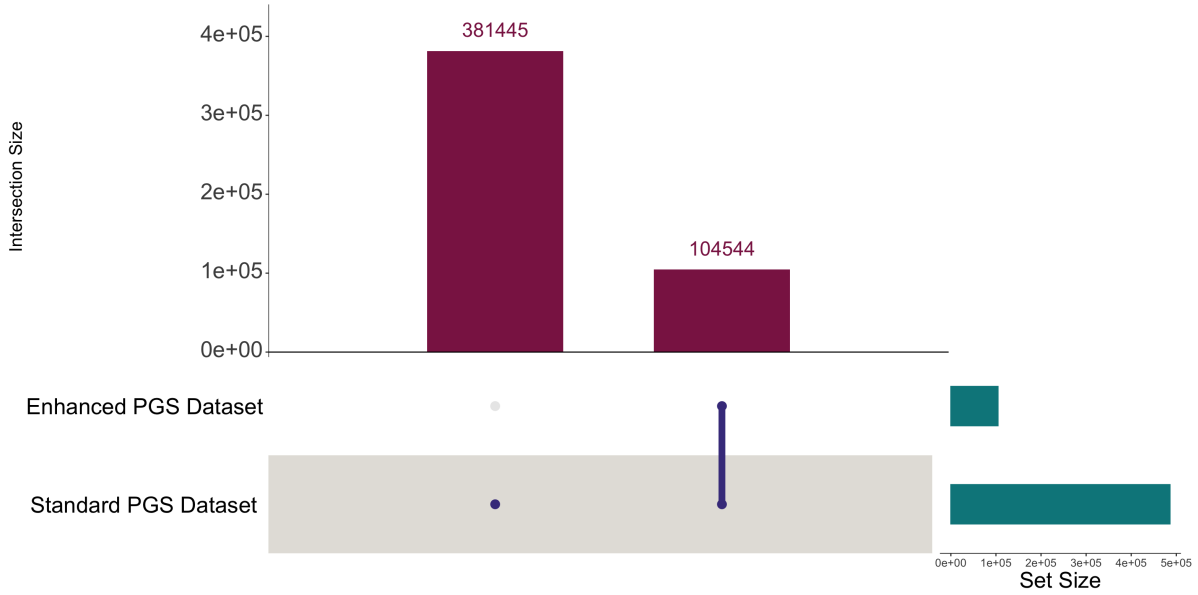


Figure A.1: An Upset plot representing the number of overlap samples between standard and enhanced dataset. The bar chart (dark red) shows the size of each intersection indicating how many samples are shared among the two sets. Each row represents one of the original sets and each column represents a specific intersection. In the matrix, filled dots (blue) indicate the sets that contribute to that particular intersection, while empty spaces denote the absence of a set in that combination. Additionally, the separate section (dark green) displays the overall size of each individual set for reference. The plot illustrates that the enhanced dataset is a subset of the standard dataset.

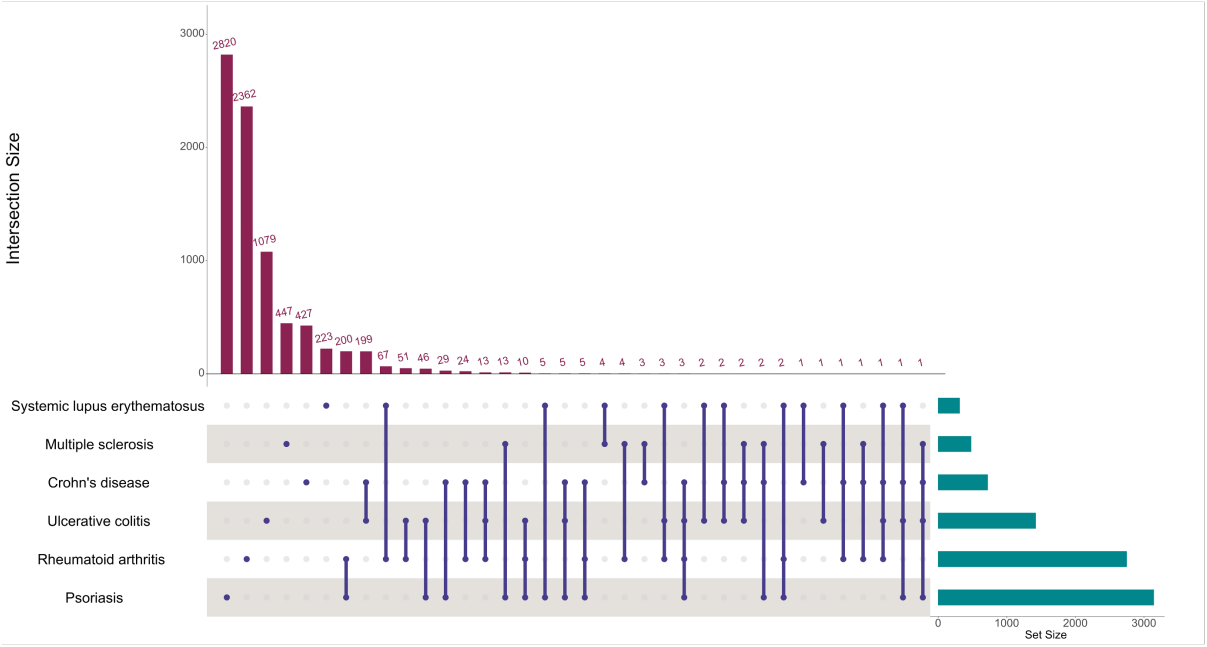


Figure A.2: This upset plot visualizes the overlap of samples among AIDs in the enhanced dataset. The dark red bar chart represents the size of each intersection, with the numbers above the bars indicating the sample count shared between sets. Each row corresponds to an original set, while each column represents a specific intersection. In the matrix, blue filled dots signify the presence of a set within an intersection, whereas empty spaces indicate its absence. Additionally, a separate dark green section highlights the total size of each individual set for reference.

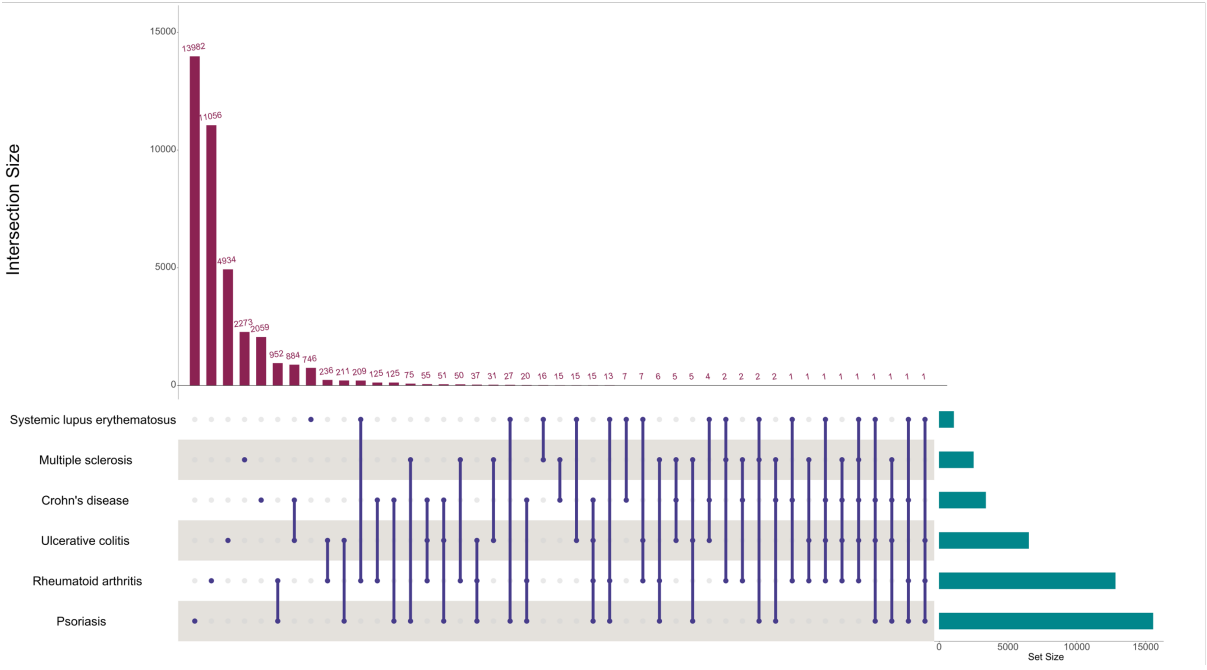


Figure A.3: This upset plot visualizes the overlap of samples among AIDs in the standard dataset. The dark red bar chart represents the size of each intersection, with the numbers above the bars indicating the sample count shared between sets. Each row corresponds to an original set, while each column represents a specific intersection. In the matrix, blue filled dots signify the presence of a set within an intersection, whereas empty spaces indicate its absence. Additionally, a separate dark green section highlights the total size of each individual set for reference.

Table A.1: Odds and Hazard Ratios for Autoimmune Outcomes Associated with Pemphigus (Excluding Other Blistering Diseases) in White Patients Using TriNetX

Outcome Name	Patient Count	OR	P-value	HR	P-value
Alopecia areata	12	1.201	0.669	2.843	0.349
Antiphospholipid syndrome	20	2.008	0.067	4.414	0.932
Autoimmune thyroiditis	27	0.494	0.002	0.567	0.673
Behçet's disease	10	1	1	3.144	0.576
Bullous pemphigoid	560	72.084	0	132.89	0.010
Celiac disease	17	1.310	0.464	1.442	0.725
Cicatricial pemphigoid	287	–	–	–	–
Crohn's disease	29	1.533	0.147	1.641	0.842
Discoid lupus erythematosus	34	3.433	0	12.566	0.052
Eosinophilic esophagitis	10	1	1	1.477	0.256
Hemolytic anemias	22	0.956	0.881	1.13	0.103
Immune thrombocytopenic purpura	13	1.302	0.531	3.063	0.378
Lichen planus	55	5.602	0	6.802	0.695
Multiple sclerosis	22	2.211	0.033	2.645	0.702
Myasthenia gravis	13	1.302	0.531	2.343	0.451
Neutropenia	60	1.637	0.018	1.839	0.490
Other rheumatoid arthritis	81	1.317	0.107	1.443	0.640
Psoriasis	84	1.637	0.005	1.843	0.573
Sarcoidosis	10	1	1	2.088	0.033
Sjörgen syndrome	53	2.557	0	2.854	0.027
Systemic connective tissue, unspecified	37	3.741	0	5.294	0.574
Systemic lupus erythematosus	36	3.306	0	3.656	0.083
Thyrotoxicosis with diffuse goiter	13	0.866	0.705	0.973	0.511
Type 1 diabetes	73	1.313	0.129	1.43	0.847
Ulcerative colitis	35	1.169	0.532	1.306	0.914
Vitiligo	10	1	1	5.537	0.323

Bibliography

[1] Curtis Edward Margo and Lynn E. Harman. “Autoimmune disease: Conceptual history and contributions of ocular immunology.” In: *Survey of ophthalmology* 61 5 (2016), pp. 680–8. URL: https://api.semanticscholar.org/CorpusID:25072192.

[2] Diana Chang et al. “Accounting for eXentricities: Analysis of the X Chromosome in GWAS Reveals X-Linked Genes Implicated in Autoimmune Diseases”. In: *PLoS ONE* 9 (2014). URL: https://api.semanticscholar.org/CorpusID:180660.

[3] Wared Nour-Eldine et al. “A glimpse into the history of description of the antiphospholipid syndrome”. In: *Lupus* 29 (2020), pp. 1493 –1502. URL: https://api.semanticscholar.org/CorpusID:220940099.

[4] Yun R Li et al. “Genetic sharing and heritability of paediatric age of onset autoimmune diseases”. In: *Nature communications* 6.1 (2015), p. 8442.

[5] Jorge Cárdenas-Roldán, Adriana Rojas-Villarraga, and Juan-Manuel Anaya. “How do autoimmune diseases cluster in families? A systematic review and meta-analysis”. In: *BMC medicine* 11 (2013), pp. 1–22.

[6] Corinne Richard-Miceli and Lindsey A Criswell. “Emerging patterns of genetic overlap across autoimmune disorders”. In: *Genome medicine* 4 (2012), pp. 1–9.

[7] Ze Xiu Xiao, Joseph S Miller, and Song Guo Zheng. “An updated advance of autoantibodies in autoimmune diseases.” In: *Autoimmunity reviews* (2020), p. 102743. URL: https://api.semanticscholar.org/CorpusID:229317154.

[8] Nilofor Padhy. “Rheumatoid Arthritis: Chronic Inflammatory Autoimmune Disease”. In: *Journal of Arthritis* 10 (2021), pp. 1–1. URL: https://api.semanticscholar.org/CorpusID:237961814.

[9] Jeffrey A. Sparks. “Rheumatoid Arthritis”. In: *Annals of Internal Medicine* 170 (2019), ITC1–ITC16. URL: https://api.semanticscholar.org/CorpusID:56191178.

[10] Watson W. Buchanan et al. “Rheumatoid arthritis.” In: *Inflammopharmacology* (2023). URL: https://api.semanticscholar.org/CorpusID:258743108.

[11] Anca D. Askanase, Katrina Shum, and Hal J. Mitnick. “Systemic Lupus Erythematosus: An Overview”. In: *Social Work in Health Care* 51 (2012), pp. 576–586. URL: https://api.semanticscholar.org/CorpusID:205474400.

[12] Renaud Felten et al. “THE HISTORY OF LUPUS THROUGHOUT THE AGES.” In: *Journal of the American Academy of Dermatology* (2020). URL: https://api.semanticscholar.org/CorpusID:218557029.

[13] Bernard Zalc. “One hundred and fifty years ago Charcot reported multiple sclerosis as a new neurological disease”. In: *Brain* 141.12 (2018), pp. 3482–3488.

[14] Pascal Joly and Noémie Litrowski. “Pemphigus group (vulgaris, vegetans, foliaceus, herpetiformis, brasiliensis).” In: *Clinics in dermatology* 29 4 (2011), pp. 432–6. URL: https://api.semanticscholar.org/CorpusID:11421217.

[15] Gina Chacon, Alex G. Ortega-Loayza, and Ronald M. Cyr. “Historical notes on endemic pemphigus in South America”. In: *International Journal of Dermatology* 51 (2012). URL: https://api.semanticscholar.org/CorpusID:40947056.

[16] Anna V Kagramanova, Oleg V Knyazev, and Asfold I Parfenov. “Crohn disease: before and after 1932 year”. In: *Terapevticheskii arkhiv* 95.2 (2023), pp. 193–197.

[17] Oleg V Knyazev, Anna V Kagramanova, and Asfold I Parfenov. “Ulcerative colitis. To the 180th anniversary of the description by Karl Rokytansky”. In: *Terapevticheskii arkhiv* 93.12 (2021), pp. 1564–1568.

[18] Stefano Guandalini and Asaad A. Assiri. “Celiac disease: a review.” In: *JAMA pediatrics* 168 3 (2014), pp. 272–8. URL: https://api.semanticscholar.org/CorpusID:22056710.

[19] Ciarán P Kelly et al. “Advances in diagnosis and management of celiac disease”. In: *Gastroenterology* 148.6 (2015), pp. 1175–1186.

[20] Monty S. Losowsky. “A History of Coeliac Disease”. In: *Digestive Diseases* 26 (2008), pp. 112 –120. URL: https://api.semanticscholar.org/CorpusID:10062937.

[21] Michael Mahler et al. “Current Concepts and Future Directions for the Assessment of Autoantibodies to Cellular Antigens Referred to as Anti-Nuclear Antibodies”. In: *Journal of Immunology Research* 2014 (2014). URL: https://api.semanticscholar.org/CorpusID:1109763.

[22] Mohan S. Maddur et al. “Autoimmunity as a Predisposition for Infectious Diseases”. In: *PLoS Pathogens* 6 (2010). URL: https://api.semanticscholar.org/CorpusID:18412272.

[23] Katharina Hofmann, Ann-Katrin Clauder, and Rudolf Armin Manz. “Targeting B Cells and Plasma Cells in Autoimmune Diseases”. In: *Frontiers in Immunology* 9 (2018). URL: https://api.semanticscholar.org/CorpusID:5021859.

[24] Christiane S. Hampe. “B Cells in Autoimmune Diseases”. In: *Scientifica* 2012 (2012). URL: https://api.semanticscholar.org/CorpusID:2230201.

[25] Eliana Mariño and Shane T. Grey. “B cells as effectors and regulators of autoimmunity”. In: *Autoimmunity* 45 (2012), pp. 377 –387. URL: https://api.semanticscholar.org/CorpusID:28014139.

[26] Brad Bolon. “Cellular and Molecular Mechanisms of Autoimmune Disease”. In: *Toxicologic Pathology* 40 (2012), pp. 216 –229. URL: https://api.semanticscholar.org/CorpusID:23007198.

[27] Wen-Tao Ma et al. “The Role of Monocytes and Macrophages in Autoimmune Diseases: A Comprehensive Review”. In: *Frontiers in Immunology* 10 (2019). URL: https://api.semanticscholar.org/CorpusID:162183323.

[28] Bhagirath Singh, Kelly L. Summers, and Steven M. Kerfoot. “Novel regulatory Th17 cells and regulatory B cells in modulating autoimmune diseases.” In: *Cellular immunology* 339 (2019), pp. 29–32. URL: https://api.semanticscholar.org/CorpusID:52812931.

[29] Giulio Fortuna and Michael T. Brennan. “Systemic lupus erythematosus: epidemiology, pathophysiology, manifestations, and management.” In: *Dental clinics of North America* 57 4 (2013), pp. 631–55. URL: https://api.semanticscholar.org/CorpusID:3139745.

[30] Ron Milo and Ariel Miller. “Revised diagnostic criteria of multiple sclerosis.” In: *Autoimmunity reviews* 13 4-5 (2014), pp. 518–24. URL: https://api.semanticscholar.org/CorpusID:41271979.

[31] Antony Raharja, Satveer K Mahil, and Jonathan N Barker. “Psoriasis: a brief overview”. In: *Clinical Medicine* 21.3 (2021), pp. 170–173.

[32] Michael Kasperkiewicz et al. “Pemphigus”. In: *Nature reviews Disease primers* 3.1 (2017), pp. 1–18.

[33] Fatima Z. Syed. “Type 1 Diabetes Mellitus”. In: *Annals of Internal Medicine* 175 (2022), ITC33 –ITC48. URL: https://api.semanticscholar.org/CorpusID:247293765.

[34] Martin W Laaß, Dirk Roggenbuck, and Karsten Conrad. “Diagnosis and classification of Crohn’s disease.” In: *Autoimmunity reviews* 13 4-5 (2014), pp. 467–71. URL: https://api.semanticscholar.org/CorpusID:19783400.

[35] Kevin D. Deane and V Michael Holers. “The Natural History of Rheumatoid Arthritis.” In: *Clinical therapeutics* (2019). URL: https://api.semanticscholar.org/CorpusID:189817679.

[36] Nikolaos Grigoriadis and Vincent Van Pesch. “A basic overview of multiple sclerosis immunopathology”. In: *European journal of neurology* 22 (2015), pp. 3–13.

[37] Inna S. Afonina, Elien Van Nuffel, and Rudi Beyaert. “Immune responses and therapeutic options in psoriasis”. In: *Cellular and Molecular Life Sciences* 78 (2021), pp. 2709 –2727. URL: https://api.semanticscholar.org/CorpusID:230107665.

[38] Katharina Boch et al. “Mortality in eight autoimmune bullous diseases: A global large-scale retrospective cohort study”. In: *Journal of the European Academy of Dermatology and Venereology* 37.4 (2023), e535–e537.

[39] Enno Schmidt, Michael Kasperkiewicz, and Pascal Joly. “Pemphigus”. In: *The Lancet* 394.10201 (2019), pp. 882–894.

[40] Hayato Takahashi et al. “Autoimmunity and immunological tolerance in autoimmune bullous diseases.” In: *International immunology* (2019). URL: https://api.semanticscholar.org/CorpusID:83462566.

[41] Volker Spindler and Jens Waschke. “Pemphigus—a disease of desmosome dysfunction caused by multiple mechanisms”. In: *Frontiers in immunology* 9 (2018), p. 136.

[42] Justin M. Gregory, D. Moore, and Jill H. Simmons. “Type 1 diabetes mellitus.” In: *Pediatrics in review* 34 5 (2013), pp. 203–15. URL: https://api.semanticscholar.org/CorpusID:8436110.

[43] Antonio Di Sabatino et al. “New insights into immune mechanisms underlying autoimmune diseases of the gastrointestinal tract.” In: *Autoimmunity reviews* 14 12 (2015), pp. 1161–9. URL: https://api.semanticscholar.org/CorpusID:7387645.

[44] Thaddeus S Stappenbeck et al. “Crohn disease: a current perspective on genetics, autophagy and immunity”. In: *Autophagy* 7.4 (2011), pp. 355–374.

[45] Joseph D Feuerstein and Adam S Cheifetz. “Ulcerative colitis: epidemiology, diagnosis, and management”. In: *Mayo Clinic Proceedings*. Vol. 89. 11. Elsevier. 2014, pp. 1553–1563.

[46] Raffaella Nenna et al. “Coeliac disease”. In: *Autoimmune Dis* 2014 (2014), p. 623784.

[47] Balid Albarbar. “A Review on Autoimmune Diseases: Recent Advances and Future Perspectives”. In: *AlQalam Journal of Medical and Applied Sciences* (2024). URL: https://api.semanticscholar.org/CorpusID:271722130.

[48] Sterling West. “Clinical Overview of Rheumatoid Arthritis”. In: *Lung Disease in Rheumatoid Arthritis*. Ed. by Aryeh Fischer and Joyce S. Lee. Cham: Springer International Publishing, 2018, pp. 1–18. ISBN: 978-3-319-68888-6. DOI: 10.1007/978-3-319-68888-6_1. URL: https://doi.org/10.1007/978-3-319-68888-6_1.

[49] R Chandrasekar and Sivagami Chandrasekar. “Natural herbal treatment for rheumatoid arthritis-a review”. In: *International Journal of Pharmaceutical Sciences and Research* 8.2 (2017), p. 368.

[50] Karim Raza, Caroline M. Cardy, and Elizabeth Ann Justice. “Rheumatoid arthritis”. In: *Oxford Medicine Online* (2018). URL: https://api.semanticscholar.org/CorpusID:240308872.

[51] Bassem I. Yamout and Raed Alroughani. “Multiple Sclerosis”. In: *Seminars in Neurology* 38 (2018), pp. 212 –225. URL: https://api.semanticscholar.org/CorpusID:43931137.

[52] Kalpana Pandey and Nimisha. “An Overview on Promising Nanotechnological Approaches for the Treatment of Psoriasis.” In: *Recent patents on nanotechnology* (2020). URL: https://api.semanticscholar.org/CorpusID:211025172.

[53] Ghofran Noor Mohammad Qorban et al. “Rheumatoid Arthritis, Pathophysiology and Management”. In: *The Egyptian Journal of Hospital Medicine* 70 (2018), pp. 1898–1903. URL: https://api.semanticscholar.org/CorpusID:80632243.

[54] Young-Chang Kwon et al. “Update on the genetics of systemic lupus erythematosus: genome-wide association studies and beyond”. In: *Cells* 8.10 (2019), p. 1180.

[55] Jiwon Oh, Ángela Vidal-Jordana, and Xavier Montalban. “Multiple sclerosis: clinical aspects”. In: *Current Opinion in Neurology* 31 (2018), 752–759. URL: https://api.semanticscholar.org/CorpusID:6103857.

[56] Sk Shahriar Ahmed et al. “Biologics and biosimilars in psoriasis”. In: *Indian Journal of Dermatology* 68.3 (2023), pp. 282–295.

[57] Yen Loo Lim et al. “Autoimmune pemphigus: latest advances and emerging therapies”. In: *Frontiers in Molecular Biosciences* 8 (2022), p. 808536.

[58] Stephen M. Adams, Elizabeth Denby Close, and Aparna P. Shreenath. “Ulcerative Colitis: Rapid Evidence Review.” In: *American family physician* 105 4 (2022), pp. 406–411. URL: https://api.semanticscholar.org/CorpusID:248180598.

[59] Kartikeya Tripathi and Joseph D Feuerstein. “New developments in ulcerative colitis: latest evidence on management, treatment, and maintenance”. In: *Drugs in context* 8 (2019).

[60] Joseph D Feuerstein et al. “AGA clinical practice guidelines on the medical management of moderate to severe luminal and perianal fistulizing Crohn’s disease”. In: *Gastroenterology* 160.7 (2021), pp. 2496–2508.

[61] Omar Nadhem et al. “Review and practice guidelines for celiac disease in 2014”. In: *Postgraduate Medicine* 127 (2015), pp. 259 –265. URL: https://api.semanticscholar.org/CorpusID:25325537.

[62] RS Nithyashree and R Deveswaran. “A comprehensive review on rheumatoid arthritis”. In: *Journal of Pharmaceutical Research International* (2020).

[63] Kenji Oku and Tatsuya Atsumi. “Systemic lupus erythematosus: nothing stale her infinite variety”. In: *Modern Rheumatology* 28 (2018), pp. 758 –765. URL: https://api.semanticscholar.org/CorpusID:49428531.

[64] Ron Milo and Esther Kahana. “Multiple sclerosis: geoepidemiology, genetics and the environment.” In: *Autoimmunity reviews* 9 5 (2010), A387–94. URL: https://api.semanticscholar.org/CorpusID:29097478.

[65] Marijana Vičić et al. “Current concepts of psoriasis immunopathogenesis”. In: *International Journal of Molecular Sciences* 22.21 (2021), p. 11574.

[66] Marie Cerna. “Epigenetic regulation in etiology of type 1 diabetes mellitus”. In: *International journal of molecular sciences* 21.1 (2019), p. 36.

[67] Charles W. Randall et al. “From historical perspectives to modern therapy: a review of current and future biological treatments for Crohn’s disease”. In: *Therapeutic Advances in Gastroenterology* 8 (2015), pp. 143 –159. URL: https://api.semanticscholar.org/CorpusID:9216314.

[68] Qingdong Guan. “A comprehensive review and update on the pathogenesis of inflammatory bowel disease”. In: *Journal of immunology research* 2019.1 (2019), p. 7247238.

[69] Naiyana Gujral, Hugh J Freeman, and Alan BR Thomson. “Celiac disease: prevalence, diagnosis, pathogenesis and treatment”. In: *World journal of gastroenterology: WJG* 18.42 (2012), p. 6036.

[70] Na Deng et al. “Single nucleotide polymorphisms and cancer susceptibility”. In: *Oncotarget* 8 (2017), pp. 110635 –110649. URL: https://api.semanticscholar.org/CorpusID:30230041.

[71] Greg Shaw. “Polymorphism and single nucleotide polymorphisms (SNPs)”. In: *BJU International* 112 (2013). URL: https://api.semanticscholar.org/CorpusID:23413964.

[72] Anthony M. DeAngelis, Meaghan Roy-O’Reilly, and Annabelle Rodriguez. “Genetic Alterations Affecting Cholesterol Metabolism and Human Fertility1”. In: *Biology of Reproduction*. 2014. URL: https://api.semanticscholar.org/CorpusID:5337626.

[73] Renu Chaudhary et al. “Role of single nucleotide polymorphisms in pharmacogenomics and their association with human diseases”. In: *Drug Metabolism Reviews* 47 (2015), pp. 281 –290. URL: https://api.semanticscholar.org/CorpusID:46863177.

[74] Aitor Nogales and Marta L DeDiego. “Host Single Nucleotide Polymorphisms Modulating Influenza A Virus Disease in Humans”. In: *Pathogens* 8 (2019). URL: https://api.semanticscholar.org/CorpusID:203639036.

[75] Wouter J. Venema et al. “A cis-regulatory element regulates ERAP2 expression through autoimmune disease risk SNPs”. In: *Cell Genomics* 4 (2023). URL: https://api.semanticscholar.org/CorpusID:257379791.

[76] Kaoru Yamagata, Shingo Nakayamada, and Yoshiya Tanaka. “Critical roles of super-enhancers in the pathogenesis of autoimmune diseases”. In: *Inflammation and Regeneration* 40 (2020). URL: https://api.semanticscholar.org/CorpusID:221614230.

[77] Xiaoxiao Liu, Zhijun Han, and Cheng-Jian Yang. “Associations of microRNA single nucleotide polymorphisms and disease risk and pathophysiology”. In: *Clinical Genetics* 92 (2017). URL: https://api.semanticscholar.org/CorpusID:13936267.

[78] Raghavan Chinnadurai et al. “From Single Nucleotide Polymorphisms to Constant Immunosuppression: Mesenchymal Stem Cell Therapy for Autoimmune Diseases”. In: *BioMed Research International* 2013 (2013). URL: https://api.semanticscholar.org/CorpusID:6572365.

[79] Szilvia Fiatal and Róza Ádány. “Application of Single-Nucleotide Polymorphism-Related Risk Estimates in Identification of Increased Genetic Susceptibility to Cardiovascular Diseases: A Literature Review”. In: *Frontiers in Public Health* 5 (2018). URL: https://api.semanticscholar.org/CorpusID:3377234.

[80] Karen A Hunt et al. “Negligible impact of rare autoimmune-locus coding-region variants on missing heritability”. In: *Nature* 498.7453 (2013), pp. 232–235.

[81] Jody Ye, Kathleen M Gillespie, and Santiago Rodriguez. “Unravelling the roles of susceptibility loci for autoimmune diseases in the post-GWAS era”. In: *Genes* 9.8 (2018), p. 377.

[82] Simon H Jiang, Maurice Stanley, and Carola G Vinuesa. “Rare genetic variants in systemic autoimmunity”. In: *Immunology and Cell Biology* 98.6 (2020), pp. 490–499.

[83] Adil Harroud and David A. Hafler. “Common genetic factors among autoimmune diseases”. In: *Science* 380 (2023), pp. 485 –490. URL: https://api.semanticscholar.org/CorpusID:258486613.

[84] Paula S. Ramos, Andrew M. Shedlock, and Carl D. Langefeld. “Genetics of autoimmune diseases: insights from population genetics”. In: *Journal of Human Genetics* 60 (2015), pp. 657 –664. URL: https://api.semanticscholar.org/CorpusID:5458305.

[85] Fulvia Ceccarelli, Nancy Agmon-Levin, and Carlo Perricone. “Genetic Factors of Autoimmune Diseases”. In: *Journal of Immunology Research* 2016 (2016). URL: https://api.semanticscholar.org/CorpusID:31919759.

[86] Kazuhiko Yamamoto and Yukinori Okada. “Shared genetic factors and their causality in autoimmune diseases”. In: *Annals of the Rheumatic Diseases* 78 (2019), pp. 1449 –1451. URL: https://api.semanticscholar.org/CorpusID:73484859.

[87] Michael Olbrich et al. “Genetics and omics analysis of autoimmune skin blistering diseases”. In: *Frontiers in immunology* 10 (2019), p. 2327.

[88] M Yu Zakharova et al. “The contribution of major histocompatibility complex class II genes to an association with autoimmune diseases”. In: *Acta Naturae ( )* 11.4 (43) (2019), pp. 4–12.

[89] Hiroaki Hatano and Kazuyoshi Ishigaki. “Functional Genetics to Understand the Etiology of Autoimmunity”. In: *Genes* 14 (2023). URL: https://api.semanticscholar.org/CorpusID:257238204.

[90] Chris Cotsapas et al. “Pervasive sharing of genetic effects in autoimmune disease”. In: *PLoS genetics* 7.8 (2011), e1002254.

[91] David Ellinghaus et al. “Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci”. In: *Nature genetics* 48.5 (2016), pp. 510–518.

[92] Xiao-Feng Chen et al. “Multiomics dissection of molecular regulatory mechanisms underlying autoimmune-associated noncoding SNPs”. In: *JCI Insight* 5 (2020). URL: https://api.semanticscholar.org/CorpusID:221478520.

[93] Anne M Hocking and Jane H. Buckner. “Genetic basis of defects in immune tolerance underlying the development of autoimmunity”. In: *Frontiers in Immunology* 13 (2022). URL: https://api.semanticscholar.org/CorpusID:251201513.

[94] Dermot P. B. McGovern, Subra Kugathasan, and Judy H. Cho. “Genetics of Inflammatory Bowel Diseases.” In: *Gastroenterology* 149 5 (2015), 1163–1176.e2. URL: https://api.semanticscholar.org/CorpusID:36730196.

[95] Jonathan Flint. “Gwas”. In: *Current Biology* 23.7 (2013), R265–R266.

[96] Emil Uffelmann et al. “Genome-wide association studies”. In: *Nature Reviews Methods Primers* 1.1 (2021), p. 59.

[97] Judy H Cho. “Genome-wide association studies: present status and future directions”. In: *Gastroenterology* 138.5 (2010), pp. 1668–1672.

[98] Shing Wan Choi, Timothy Shin Heng Mak, and Paul F. O’Reilly. “Tutorial: a guide to performing polygenic risk score analyses”. In: *Nature Protocols* 15 (2020), pp. 2759 –2772. URL: https://api.semanticscholar.org/CorpusID:256839147.

[99] Ahmad Mohammad Alqudah et al. “GWAS: Fast-forwarding gene identification and characterization in temperate Cereals: lessons from Barley – A review”. In: *Journal of Advanced Research* 22 (2019), pp. 119 –135. URL: https://api.semanticscholar.org/CorpusID:210708013.

[100] Kouichi Ozaki et al. “Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction”. In: *Nature genetics* 32.4 (2002), pp. 650–654.

[101] Jacqueline Milet et al. “Mixed logistic regression in genome-wide association studies”. In: *BMC bioinformatics* 21 (2020), pp. 1–17.

[102] Stephen T Sherry et al. “dbSNP: the NCBI database of genetic variation”. In: *Nucleic acids research* 29.1 (2001), pp. 308–311.

[103] Carlo Bonferroni. “Teoria statistica delle classi e calcolo delle probabilita”. In: *Pubblicazioni del R istituto superiore di scienze economiche e commericiali di firenze* 8 (1936), pp. 3–62.

[104] Priya Duggal et al. “Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies”. In: *BMC genomics* 9 (2008), pp. 1–8.

[105] Cathryn M. Lewis and Evangelos Vassos. “Polygenic risk scores: from research tools to clinical instruments”. In: *Genome Medicine* 12 (2020). URL: https://api.semanticscholar.org/CorpusID:218682590.

[106] JR Shaffer, E Feingold, and ML Marazita. “Genome-wide association studies: Prospects and challenges for oral health”. In: *Journal of Dental Research* 91.7 (2012), pp. 637–641.

[107] Laura Buzdugan et al. “Assessing statistical significance in multivariable genome wide association analysis”. In: *Bioinformatics* 32 (2016), pp. 1990 –2000. URL: https://api.semanticscholar.org/CorpusID:16355659.

[108] Gang Shi et al. “Mining gold dust under the genome wide significance level: a two-stage approach to analysis of GWAS”. In: *Genetic Epidemiology* 35 (2011). URL: https://api.semanticscholar.org/CorpusID:21629124.

[109] Lingjie Weng et al. “SNP-based pathway enrichment analysis for genome-wide association studies”. In: *BMC Bioinformatics* 12 (2011), pp. 99 –99. URL: https://api.semanticscholar.org/CorpusID:5065425.

[110] Bahareh Rabbani, Mustafa Tekin, and Nejat Mahdieh. “The promise of whole-exome sequencing in medical genetics”. In: *Journal of human genetics* 59.1 (2014), pp. 5–15.

[111] Eleanor G Seaby, Reuben J Pengelly, and Sarah Ennis. “Exome sequencing explained: a practical guide to its clinical application”. In: *Briefings in functional genomics* 15.5 (2016), pp. 374–384.

[112] Robert M. Geraghty, Eric G. Olinger, and John A. Sayer. “Whole exome sequencing of large populations: identification of loss of function alleles and implications for inherited kidney diseases.” In: *Kidney international* (2021). URL: https://api.semanticscholar.org/CorpusID:231864976.

[113] Frederik Otzen Bagger et al. “Whole genome sequencing in clinical practice”. In: *BMC medical genomics* 17.1 (2024), p. 39.

[114] TM Aune et al. “Expression of long non-coding RNAs in autoimmunity and linkage to enhancer function and autoimmune disease risk genetic variants”. In: *Journal of autoimmunity* 81 (2017), pp. 99–109.

[115] Guillaume Lettre and John D Rioux. “Autoimmune diseases: insights from genome-wide association studies”. In: *Human molecular genetics* 17.R2 (2008), R116–R121.

[116] Jonathan Massey and Steve Eyre. “Rare variants and autoimmune disease”. In: *Briefings in functional genomics* 13.5 (2014), pp. 392–397.

[117] Fei Chen et al. “Whole-genome sequencing of a monozygotic twin discordant for systemic lupus erythematosus”. In: *Molecular medicine reports* 17.6 (2018), pp. 8391–8396.

[118] Matt A Field. “Detecting pathogenic variants in autoimmune diseases using high-throughput sequencing”. In: *Immunology and Cell Biology* 99.2 (2021), pp. 146–156.

[119] 1000 Genomes Project Consortium Corresponding authors Auton Adam adam. auton@ gmail. com 1 b Abecasis Gonçalo R. goncalo@ umich. edu 2 c et al. “A global reference for human genetic variation”. In: *Nature* 526.7571 (2015), pp. 68–74.

[120] Anders Bergström et al. “Insights into human genetic variation and population history from 929 diverse genomes”. In: *Science* 367.6484 (2020), eaay5012.

[121] Konrad J Karczewski et al. “The mutational constraint spectrum quantified from variation in 141,456 humans”. In: *Nature* 581.7809 (2020), pp. 434–443.

[122] Daniel Taliun et al. “Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program”. In: *Nature* 590.7845 (2021), pp. 290–299.

[123] William McLaren et al. “The ensembl variant effect predictor”. In: *Genome biology* 17 (2016), pp. 1–14.

[124] James Renwick Beattie and Francis W. L. Esmonde-White. “Exploration of Principal Component Analysis: Deriving Principal Component Analysis Visually Using Spectra”. In: *Applied Spectroscopy* 75 (2021), pp. 361 –375. URL: https://api.semanticscholar.org/CorpusID:230486086.

[125] Aaron S. Hess and John R Hess. “Principal component analysis”. In: *Transfusion* 58 (2018). URL: https://api.semanticscholar.org/CorpusID:8475494.

[126] Roger Watson. “Polygenic risk scores.” In: *Journal of advanced nursing* (2019). URL: https://api.semanticscholar.org/CorpusID:96436251.

[127] Paul S. de Vries. “Polygenic risk, lifestyle and the lifetime risk of coronary artery disease”. In: *Heart* 109 (2023), pp. 730 –731. URL: https://api.semanticscholar.org/CorpusID:256698457.

[128] Benjamin Cross, Richard M. Turner, and Munir Pirmohamed. “Polygenic risk scores: An overview from bench to bedside for personalised medicine”. In: *Frontiers in Genetics* 13 (2022). URL: https://api.semanticscholar.org/CorpusID:253449488.

[129] Chachrit Khunsriraksakul et al. “Construction and Application of Polygenic Risk Scores in Autoimmune Diseases”. In: *Frontiers in Immunology* 13 (2022). URL: https://api.semanticscholar.org/CorpusID:250106921.

[130] Tuan V. Nguyen and John A. Eisman. “Post-GWAS Polygenic Risk Score: Utility and Challenges”. In: *JBMR Plus* 4 (2020). URL: https://api.semanticscholar.org/CorpusID:224952643.

[131] Mika Ala-Korpela and Michael V. Holmes. “Polygenic risk scores and the prediction of common diseases.” In: *International journal of epidemiology* (2019). URL: https://api.semanticscholar.org/CorpusID:209328735.

[132] Mohammad Dehestani, Hui Liu, and Thomas Gasser. “Polygenic Risk Scores Contribute to Personalized Medicine of Parkinson’s Disease”. In: *Journal of Personalized Medicine* 11 (2021). URL: https://api.semanticscholar.org/CorpusID:239239716.

[133] Jack Euesden, Cathryn M. Lewis, and Paul F. O’Reilly. “PRSice: Polygenic Risk Score software”. In: *Bioinformatics* 31 (2014), pp. 1466 –1468. URL: https://api.semanticscholar.org/CorpusID:15976950.

[134] Naomi R. Wray et al. “From Basic Science to Clinical Application of Polygenic Risk Scores: A Primer.” In: *JAMA psychiatry* (2020). URL: https://api.semanticscholar.org/CorpusID:222169651.

[135] Lucía Santiago-Lamelas et al. “Utility of polygenic risk scores to aid in the diagnosis of rheumatic diseases.” In: *Best practice & research. Clinical rheumatology* (2024), p. 101973. URL: https://api.semanticscholar.org/CorpusID:271120460.

[136] Anca Gabriela Pavel et al. “Cumulative Effect Assessment of Common Genetic Variants on Prostate Cancer: Preliminary Studies”. In: *Biomedicines* 10 (2022). URL: https://api.semanticscholar.org/CorpusID:253292771.

[137] Carla Lluís-Ganella et al. “Additive effect of multiple genetic variants on the risk of coronary artery disease.” In: *Revista espanola de cardiologia* 63 8 (2010), pp. 925–33. URL: https://api.semanticscholar.org/CorpusID:37065576.

[138] Samuel A. Lambert, Gad Abraham, and Michael Inouye. “Towards clinical utility of polygenic risk scores.” In: *Human molecular genetics* (2019). URL: https://api.semanticscholar.org/CorpusID:198999041.

[139] Matthew A. Brown and Zhixiu Li. “Polygenic risk scores and rheumatic diseases”. In: *Chinese Medical Journal* 134 (2021), pp. 2521 –2524. URL: https://api.semanticscholar.org/CorpusID:239052485.

[140] Tatiane Yanes et al. “The Emerging Field of Polygenic Risk Scores and Perspective for Use in Clinical Care.” In: *Human molecular genetics* (2020). URL: https://api.semanticscholar.org/CorpusID:220336127.

[141] Alicia R. Martin et al. “Clinical use of current polygenic risk scores may exacerbate health disparities”. In: *Nature Genetics* 51 (2019), pp. 584 –591. URL: https://api.semanticscholar.org/CorpusID:85567228.

[142] Jana Schwarzerová et al. “A perspective on genetic and polygenic risk scores—advances and limitations and overview of associated tools”. In: *Briefings in Bioinformatics* 25 (2024). URL: https://api.semanticscholar.org/CorpusID:269927536.

[143] Song Zhai et al. “Pharmacogenomics polygenic risk score for drug response prediction using PRS-PGx methods”. In: *Nature Communications* 13 (2022). URL: https://api.semanticscholar.org/CorpusID:252160792.

[144] Adebowale Mary K. Deanna R. Segun Palmira Chani J. Michael M Adeyemo Balaconis Darnes Fatumo Granados Moreno Ho et al. “Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps”. In: *Nature Medicine* 27 (2021), pp. 1876 –1884. URL: https://api.semanticscholar.org/CorpusID:244131258.

[145] Maria Cerezo et al. “The NHGRI-EBI GWAS Catalog: standards for reusability, sustainability and diversity”. In: *Nucleic acids research* 53.D1 (2025), pp. D998–D1005.

[146] Annalisa Buniello et al. “The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019”. In: *Nucleic acids research* 47.D1 (2019), pp. D1005–D1012.

[147] Eugenio López-Cortegano and Armando Caballero. “Inferring the Nature of Missing Heritability in Human Traits Using Data from the GWAS Catalog”. In: *Genetics* 212 (2019), pp. 891 –904. URL: https://api.semanticscholar.org/CorpusID:163166805.

[148] Elliot Sollis et al. “The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource”. In: *Nucleic Acids Research* 51 (2022), pp. D977 –D985. URL: https://api.semanticscholar.org/CorpusID:253419413.

[149] Tanya B. Horwitz et al. “A Decade in Psychiatric GWAS Research”. In: *Molecular psychiatry* 24 (2018), pp. 378 –389. URL: https://api.semanticscholar.org/CorpusID:49414226.

[150] Maria Cerezo et al. “The NHGRI-EBI GWAS Catalog: standards for reusability, sustainability and diversity”. In: *Nucleic Acids Research* 53 (2024), pp. D998 –D1005. URL: https://api.semanticscholar.org/CorpusID:274059401.

[151] Ramiro Magno and Ana-Teresa Maia. “gwasrapidd: an R package to query, download and wrangle GWAS catalog data”. In: *Bioinformatics* 36 (2019), pp. 649 –650. URL: https://api.semanticscholar.org/CorpusID:181371662.

[152] Tianze Cao, Anshui Li, and Yuexia Huang. “pandasGWAS: a Python package for easy retrieval of GWAS catalog data”. In: *BMC genomics* 24.1 (2023), p. 238.

[153] Anastasia L. Wise, Lin Gyi, and Teri A. Manolio. “eXclusion: toward integrating the X chromosome in genome-wide association analyses.” In: *American journal of human genetics* 92 5 (2013), pp. 643–7. URL: https://api.semanticscholar.org/CorpusID:9471035.

[154] Thomas Battram et al. “The EWAS Catalog: a database of epigenome-wide association studies”. In: *Wellcome Open Research* 7 (2022). URL: https://api.semanticscholar.org/CorpusID:246587206.

[155] Danielle Welter et al. “The NHGRI GWAS Catalog, a curated resource of SNP-trait associations”. In: *Nucleic Acids Research* 42 (2013), pp. D1001 –D1006. URL: https://api.semanticscholar.org/CorpusID:215529634.

[156] Samuel A Lambert et al. “The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation”. In: *Nature Genetics* 53.4 (2021), pp. 420–425.

[157] Hannah Wand et al. “Improving reporting standards for polygenic scores in risk prediction studies”. In: *Nature* 591 (2020), pp. 211 –219. URL: https://api.semanticscholar.org/CorpusID:218538816.

[158] Cong Yu, M Eric Gershwin, and Christopher Chang. “Diagnostic criteria for systemic lupus erythematosus: a critical review”. In: *Journal of autoimmunity* 48 (2014), pp. 10–13.

[159] Renata Dwornicka and Jacek Pietraszek. “The outline of the expert system for the design of experiment”. In: *Production Engineering Archives* 20 (2018), pp. 43 –48. URL: https://api.semanticscholar.org/CorpusID:58542750.

[160] Hanna Ćwiek-Kupczyńska et al. “Semantic concept schema of the linear mixed model of experimental observations”. In: *Scientific Data* 7 (2020). URL: https://api.semanticscholar.org/CorpusID:211525726.

[161] Sirarat Sarntivijai et al. “Linking rare and common disease: mapping clinical disease-phenotypes to ontologies in therapeutic target validation”. In: *Journal of Biomedical Semantics* 7 (2016). URL: https://api.semanticscholar.org/CorpusID:10675398.

[162] Edison Ong et al. “Comparison, alignment, and synchronization of cell line information between CLO and EFO”. In: *BMC Bioinformatics* 18 (2017). URL: https://api.semanticscholar.org/CorpusID:6187014.

[163] Fina A. S. Kurreeman et al. “Genetic basis of autoantibody positive and negative rheumatoid arthritis risk in a multi-ethnic cohort derived from electronic health records.” In: *American journal of human genetics* 88 1 (2011), pp. 57–69. URL: https://api.semanticscholar.org/CorpusID:17551897.

[164] Konstantina Charmpi et al. “Optimizing network propagation for multi-omics data integration”. In: *PLoS Computational Biology* 17.11 (2021), e1009161.

[165] Hadas Biran et al. “WebPropagate: A Web Server for Network Propagation.” In: *Journal of molecular biology* 430 15 (2018), pp. 2231–2236. URL: https://api.semanticscholar.org/CorpusID:3836319.

[166] Daniel E Carlin et al. “Network propagation in the cytoscape cyberinfrastructure”. In: *PLoS computational biology* 13.10 (2017), e1005598.

[167] Peng Ji et al. “Signal propagation in complex networks”. In: *Physics Reports* (2023). URL: https://api.semanticscholar.org/CorpusID:257997615.

[168] Lenore Cowen et al. “Network propagation: a universal amplifier of genetic associations”. In: *Nature Reviews Genetics* 18.9 (2017), pp. 551–562.

[169] Giovanni Visonà et al. “Network propagation for GWAS analysis: a practical guide to leveraging molecular networks for disease gene discovery”. In: *Briefings in bioinformatics* 25.2 (2024), bbae014.

[170] Hadas Biran, Martin Kupiec, and Roded Sharan. “Comparative analysis of normalization methods for network propagation”. In: *Frontiers in genetics* 10 (2019), p. 4.

[171] Aristo Vojdani, K. Michael Pollard, and Andrew W. Campbell. “Environmental Triggers and Autoimmunity”. In: *Autoimmune Diseases* 2014 (2014). URL: https://api.semanticscholar.org/CorpusID:7850574.

[172] Hanane Touil, Kristin Mounts, and Philip Lawrence De Jager. “Differential impact of environmental factors on systemic and localized autoimmunity”. In: *Frontiers in Immunology* 14 (2023). URL: https://api.semanticscholar.org/CorpusID:258824564.

[173] Kenneth Michael Pollard. “Environment, Autoantibodies, and Autoimmunity”. In: *Frontiers in Immunology* 6 (2015). URL: https://api.semanticscholar.org/CorpusID:18297194.

[174] Jennifer M. P. Woo et al. “The role of environmental exposures and gene–environment interactions in the etiology of systemic lupus erythematous”. In: *Journal of Internal Medicine* 291 (2022), pp. 755 –778. URL: https://api.semanticscholar.org/CorpusID:246700145.

[175] M. Firoze Khan and Hui Wang. “Environmental Exposures and Autoimmune Diseases: Contribution of Gut Microbiome”. In: *Frontiers in Immunology* 10 (2020). URL: https://api.semanticscholar.org/CorpusID:210130619.

[176] Sandra Dedrick et al. “The Role of Gut Microbiota and Environmental Factors in Type 1 Diabetes Pathogenesis”. In: *Frontiers in Endocrinology* 11 (2020). URL: https://api.semanticscholar.org/CorpusID:211472610.

[177] Lars Klareskog et al. “The importance of differences; On environment and its interactions with genes and immunity in the causation of rheumatoid arthritis”. In: *Journal of Internal Medicine* 287 (2020), pp. 514 –533. URL: https://api.semanticscholar.org/CorpusID:212728498.

[178] Marina Arleevskaya et al. “Interplay of Environmental, Individual and Genetic Factors in Rheumatoid Arthritis Provocation”. In: *International Journal of Molecular Sciences* 23 (2022). URL: https://api.semanticscholar.org/CorpusID:251079052.

[179] Biola M. Javierre, Henar Hernando, and Esteban Ballestar. “Environmental triggers and epigenetic deregulation in autoimmune disease.” In: *Discovery medicine* 12 67 (2011), pp. 535–45. URL: https://api.semanticscholar.org/CorpusID:25730067.

[180] Marina I. Arleevskaya et al. “Editorial: Microbial and Environmental Factors in Autoimmune and Inflammatory Diseases”. In: *Frontiers in Immunology* 8 (2017). URL: https://api.semanticscholar.org/CorpusID:10270574.

[181] Leonidas H. Duntas. “Environmental factors and thyroid autoimmunity.” In: *Annales d’endocrinologie* 72 2 (2011), pp. 108–13. URL: https://api.semanticscholar.org/CorpusID:205515859.

[182] Stefanie Jörg et al. “Environmental factors in autoimmune diseases and their role in multiple sclerosis”. In: *Cellular and Molecular Life Sciences: CMLS* 73 (2016), pp. 4611 –4622. URL: https://api.semanticscholar.org/CorpusID:16403403.

[183] Cathie Sudlow et al. “UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age”. In: *PLoS medicine* 12.3 (2015), e1001779.

[184] William E.R. Ollier, Tim Sprosen, and Tim C Peakman. “UK Biobank: from concept to reality.” In: *Pharmacogenomics* 6 6 (2005), pp. 639–46. URL: https://api.semanticscholar.org/CorpusID:331938.

[185] Rishi Caleyachetty et al. “United Kingdom Biobank (UK Biobank): JACC Focus Seminar 6/8.” In: *Journal of the American College of Cardiology* 78 1 (2021), pp. 56–65. URL: https://api.semanticscholar.org/CorpusID:235710140.

[186] Megan C. Conroy et al. “UK Biobank: a globally important resource for cancer research”. In: *British Journal of Cancer* 128 (2022), pp. 519 –527. URL: https://api.semanticscholar.org/CorpusID:253707652.

[187] Clare Bycroft et al. “The UK Biobank resource with deep phenotyping and genomic data”. In: *Nature* 562.7726 (2018), pp. 203–209.

[188] Geoff Watts. “UK Biobank opens its data vaults to researchers”. In: *BMJ : British Medical Journal* 344 (2012). URL: https://api.semanticscholar.org/CorpusID:37705856.

[189] Paul Elliott and Tim C Peakman. “The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine.” In: *International journal of epidemiology* 37 2 (2008), pp. 234–44. URL: https://api.semanticscholar.org/CorpusID:12166388.

[190] Frank M. Sullivan et al. “How could primary care meet the informatics needs of UK Biobank? A Scottish proposal.” In: *Informatics in primary care* 11 3 (2003), pp. 129–35. URL: https://api.semanticscholar.org/CorpusID:344016.

[191] Paul M. Matthews and Cathie L. M. Sudlow. “The UK Biobank.” In: *Brain : a journal of neurology* 138 Pt 12 (2015), pp. 3463–5. URL: https://api.semanticscholar.org/CorpusID:3056195.

[192] Thomas J. Littlejohns et al. “The UK Biobank imaging enhancement of 100,000 participants:rationale, data collection, management and future directions”. In: *Nature Communications* 11 (2020). URL: https://api.semanticscholar.org/CorpusID:218878328.

[193] Joshua D Backman et al. “Exome sequencing and analysis of 454,787 UK Biobank participants”. In: *Nature* 599.7886 (2021), pp. 628–634.

[194] Rochi Saurabh et al. “A survey of genome-wide association studies, polygenic scores and UK Biobank highlights resources for autoimmune disease genetics”. In: *Frontiers in immunology* 13 (2022), p. 972107.

[195] Bjarni V Halldorsson et al. “The sequences of 150,119 genomes in the UK Biobank”. In: *Nature* 607.7920 (2022), pp. 732–740.

[196] Noel R. Rose. “Prediction and Prevention of Autoimmune Disease in the 21st Century: A Review and Preview.” In: *American journal of epidemiology* 183 5 (2016), pp. 403–6. URL: https://api.semanticscholar.org/CorpusID:329150.

[197] Scott M Hayter and Matthew C. Cook. “Updated assessment of the prevalence, spectrum and case definition of autoimmune disease.” In: *Autoimmunity reviews* 11 10 (2012), pp. 754–65. URL: https://api.semanticscholar.org/CorpusID:24881814.

[198] Shaye Kivity and Michael Ehrenfeld. “Can we explain the higher prevalence of autoimmune disease in women?” In: *Expert Review of Clinical Immunology* 6 (2010), pp. 691 –694. URL: https://api.semanticscholar.org/CorpusID:37818130.

[199] Allan Gibofsky. “Overview of epidemiology, pathophysiology, and diagnosis of rheumatoid arthritis.” In: *The American journal of managed care* 18 13 Suppl (2012), S295–302. URL: https://api.semanticscholar.org/CorpusID:33563052.

[200] Paul P Smith and Caroline Gordon. “Systemic lupus erythematosus: clinical presentations.” In: *Autoimmunity reviews* 10 1 (2010), pp. 43–5. URL: https://api.semanticscholar.org/CorpusID:1570789.

[201] Khalid Al Johani et al. “Multiple Sclerosis—A Demyelinating Disorder and Its Dental Considerations—A Literature Review with Own Case Report”. In: *Brain Sciences* 13.7 (2023), p. 1009.

[202] Ángela Vidal-Jordana and Xavier Montalban. “Multiple Sclerosis: Epidemiologic, Clinical, and Therapeutic Aspects.” In: *Neuroimaging clinics of North America* 27 2 (2017), pp. 195–204. URL: https://api.semanticscholar.org/CorpusID:4721874.

[203] Yash Arvind Hete and Mr. Dipak Tonchar. “Autoimmune Disorder An Overview”. In: *International Journal of Advanced Research in Science, Communication and Technology* (2024). URL: https://api.semanticscholar.org/CorpusID:274974340.

[204] Gude Himabindhu. “A Brief Account on Autoimmune Disorders”. In: *Endocrinology and Metabolic Syndrome* 9 (2020), pp. 1–1. URL: https://api.semanticscholar.org/CorpusID:226566822.

[205] Frances Rees et al. “The worldwide incidence and prevalence of systemic lupus erythematosus: a systematic review of epidemiological studies”. In: *Rheumatology* 56 (2017), 1945–1961. URL: https://api.semanticscholar.org/CorpusID:205310254.

[206] Antoni Sisó-Almirall et al. “AB1343 Prevalence of autoimmune diseases in catalonia: a population based study using a public big data analytics (PADRIS)”. In: *Annals of the Rheumatic Diseases* 77 (2018), pp. 1760 –1760. URL: https://api.semanticscholar.org/CorpusID:80723056.

[207] Fan Cao et al. “Temporal trends in the prevalence of autoimmune diseases from 1990 to 2019.” In: *Autoimmunity reviews* (2023), p. 103359. URL: https://api.semanticscholar.org/CorpusID:258772601.

[208] Andrea T Borchers et al. “The geoepidemiology of systemic lupus erythematosus”. In: *Autoimmunity reviews* 9.5 (2010), A277–A287.

[209] Yinon Shapira, Nancy Agmon-Levin, and Yehuda Shoenfeld. “Defining and analyzing geoepidemiology and human autoimmunity.” In: *Journal of autoimmunity* 34 3 (2010), J168–77. URL: https://api.semanticscholar.org/CorpusID:44890373.

[210] Dario Didona et al. “Pemphigus: current and future therapeutic strategies”. In: *Frontiers in immunology* 10 (2019), p. 1418.

[211] Caitlin N Suire and Mangesh D Hade. “Extracellular vesicles in type 1 diabetes: A versatile tool”. In: *Bioengineering* 9.3 (2022), p. 105.

[212] Cristiano Pagnini and Fabio Cominelli. “Tumor necrosis factor’s pathway in Crohn’s disease: potential for intervention”. In: *International journal of molecular sciences* 22.19 (2021), p. 10273.

[213] Bruno César Da Silva et al. “Epidemiology, demographic characteristics and prognostic predictors of ulcerative colitis”. In: *World Journal of Gastroenterology: WJG* 20.28 (2014), p. 9458.

[214] Juan-Manuel Anaya et al. “The kaleidoscope of autoimmunity: multiple autoimmune syndromes and familial autoimmunity”. In: *Expert review of clinical immunology* 3.4 (2007), pp. 623–635.

[215] Adriana Rojas-Villarraga et al. “Introducing polyautoimmunity: secondary autoimmune diseases no longer exist”. In: *Autoimmune diseases* 2012.1 (2012), p. 254319.

[216] Emily C Somers et al. “Are individuals with an autoimmune disease at higher risk of a second autoimmune disorder?” In: *American journal of epidemiology* 169.6 (2009), pp. 749–755.

[217] Emily C Somers et al. “Autoimmune diseases co-occurring within individuals and within families: a systematic review”. In: *Epidemiology* 17.2 (2006), pp. 202–217.

[218] Nathalie Conrad et al. “Incidence, prevalence, and co-occurrence of autoimmune disorders over time and by age, sex, and socioeconomic status: a population-based cohort study of 22 million individuals in the UK”. In: *The Lancet* 401.10391 (2023), pp. 1878–1890.

[219] Minal Caliskan, Christopher D Brown, and Joseph C Maranville. “A catalog of GWAS fine-mapping efforts in autoimmune disease”. In: *The American Journal of Human Genetics* 108.4 (2021), pp. 549–563.

[220] Jordan T Russell et al. “Genetic risk for autoimmunity is associated with distinct changes in the human gut microbiome”. In: *Nature communications* 10.1 (2019), p. 3621.

[221] Erin B Taylor. “The complex role of adipokines in obesity, inflammation, and autoimmunity”. In: *Clinical Science* 135.6 (2021), pp. 731–752.

[222] Yun-Wen Chiu et al. “Comorbid autoimmune diseases in patients with pemphigus: a nationwide case-control study in Taiwan”. In: *European Journal of Dermatology* 27 (2017), pp. 375–381.

[223] Khalaf Kridin et al. “Association between pemphigus and psoriasis: a population-based large-scale study”. In: *Journal of the American Academy of Dermatology* 77.6 (2017), pp. 1174–1175.

[224] Khalaf Kridin et al. “Ulcerative colitis associated with pemphigus: a population-based large-scale study”. In: *Scandinavian journal of gastroenterology* 52.12 (2017), pp. 1360–1364.

[225] Jolanda M Denham and Ivor D Hill. “Celiac disease and autoimmunity: review and controversies”. In: *Current allergy and asthma reports* 13 (2013), pp. 347–353.

[226] Yael A Leshem et al. “Autoimmune diseases in patients with pemphigus and their first-degree relatives”. In: *International journal of dermatology* 50.7 (2011), pp. 827–831.

[227] A Parameswaran et al. “Identification of a new disease cluster of pemphigus vulgaris with autoimmune thyroid disease, rheumatoid arthritis and type I diabetes”. In: *British Journal of Dermatology* 172.3 (2015), pp. 729–738.

[228] Khalaf Kridin. “Pemphigus group: overview, epidemiology, mortality, and comorbidities”. In: *Immunologic research* 66.2 (2018), pp. 255–270.

[229] DY Hsu et al. “Comorbidities and inpatient mortality for pemphigus in the USA”. In: *British Journal of Dermatology* 174.6 (2016), pp. 1290–1298.

[230] R Wang et al. “Prevalence of myasthenia gravis and associated autoantibodies in paraneoplastic pemphigus and their correlations with symptoms and prognosis”. In: *British Journal of Dermatology* 172.4 (2015), pp. 968–975.

[231] Inga Koneczny et al. “Common denominators in the immunobiology of IgG4 autoimmune diseases: what do glomerulonephritis, pemphigus vulgaris, myasthenia gravis, thrombotic thrombocytopenic purpura and autoimmune encephalitis have in common?” In: *Frontiers in immunology* 11 (2021), p. 605214.

[232] Matvey B Palchuk et al. “A global federated real-world data and analytics platform for research”. In: *JAMIA open* 6.2 (2023), ooad035.

[233] Matvey B. Palchuk et al. “A global federated real-world data and analytics platform for research”. In: *JAMIA Open* 6 (2023). URL: https://api.semanticscholar.org/CorpusID:258676779.

[234] Issam D Moussa et al. “The NCDR CathPCI Registry: a US national perspective on care and outcomes for percutaneous coronary intervention”. In: *Heart* 99 (2013), pp. 297 –303. URL: https://api.semanticscholar.org/CorpusID:20524899.

[235] Umit Topaloglu and Matvey B Palchuk. “Using a federated network of real-world data to optimize clinical trials operations”. In: *JCO clinical cancer informatics* 2 (2018), pp. 1–10.

[236] C. Trocchia et al. “552 Using the TriNetX database to describe the prevalence of VTE and gastrointestinal disease in children with cystic fibrosis”. In: *Journal of Cystic Fibrosis* (2023). URL: https://api.semanticscholar.org/CorpusID:264318418.

[237] Michael Kasperkiewicz et al. “Risk of comorbid autoimmune diseases in patients with immunobullous disorders: a global large-scale cohort study”. In: *Journal of the American Academy of Dermatology* 89.6 (2023), pp. 1269–1271.

[238] Deborah J Thompson et al. “A systematic evaluation of the performance and properties of the UK Biobank Polygenic Risk Score (PRS) Release”. In: *Plos one* 19.9 (2024), e0307270.

[239] Simon Jupp et al. “A new Ontology Lookup Service at EMBL-EBI.” In: *SWAT4LS* 2 (2015), pp. 118–119.

[240] Laurence R. Meyer et al. “The UCSC Genome Browser database: extensions and updates 2013”. In: *Nucleic Acids Research* 41 (2012), pp. D64 –D69. URL: https://api.semanticscholar.org/CorpusID:10341498.

[241] Brian J. Raney et al. “The UCSC Genome Browser database: 2024 update”. In: *Nucleic Acids Research* 52 (2023), pp. D1082 –D1088. URL: https://api.semanticscholar.org/CorpusID:265149175.

[242] Adam Frankish et al. “GENCODE reference annotation for the human and mouse genomes”. In: *Nucleic Acids Research* 47 (2018), pp. D766 –D773. URL: https://api.semanticscholar.org/CorpusID:53028895.

[243] Arthur Liberzon et al. “The molecular signatures database hallmark gene set collection”. In: *Cell systems* 1.6 (2015), pp. 417–425.

[244] Damian Szklarczyk et al. “The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest”. In: *Nucleic acids research* 51.D1 (2023), pp. D638–D646.

[245] Hadas Samuels et al. “Autoimmune disease classification based on PubMed text mining”. In: *Journal of Clinical Medicine* 11.15 (2022), p. 4345.

[246] Manuel J Amador-Patarroyo, Alberto Rodriguez-Rodriguez, and Gladis Montoya-Ortiz. “How does age at onset influence the outcome of autoimmune diseases?” In: *Autoimmune diseases* 2012.1 (2012), p. 251730.

[247] Junyang Qian et al. “A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank”. In: *PLoS genetics* 16.10 (2020), e1009141.

[248] Bjarni J Vilhjálmsson et al. “Modeling linkage disequilibrium increases accuracy of polygenic risk scores”. In: *The american journal of human genetics* 97.4 (2015), pp. 576–592.

[249] Florian Privé, Julyan Arbel, and Bjarni J Vilhjálmsson. “LDpred2: better, faster, stronger”. In: *Bioinformatics* 36.22-23 (2020), pp. 5424–5431.

[250] Yong-Fei Wang et al. “Identification of 38 novel loci for systemic lupus erythematosus and genetic heterogeneity between ancestral groups”. In: *Nature communications* 12.1 (2021), p. 772.

[251] Leonid Padyukov. “Genetics of rheumatoid arthritis”. In: *Seminars in immunopathology*. Vol. 44. 1. Springer. 2022, pp. 47–62.

[252] International Multiple Sclerosis Genetics Consortium*† et al. “Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility”. In: *Science* 365.6460 (2019), eaav7188.

[253] Koldo Garcia-Etxebarria et al. “Local genetic variation of inflammatory bowel disease in Basque population and its effect in risk prediction”. In: *Scientific Reports* 12.1 (2022), p. 3386.

[254] Ying Jin et al. “Early-onset autoimmune vitiligo associated with an enhancer variant haplotype that upregulates class II HLA expression”. In: *Nature communications* 10.1 (2019), p. 391.

[255] Juliana Imgenberg-Kreuz et al. “Genetics and epigenetics in primary Sjögren’s syndrome”. In: *Rheumatology* 60.5 (2021), pp. 2085–2098.

[256] Shuang-Xia Zhao et al. “Robust evidence for five new Graves’ disease risk loci from a staged genome-wide association analysis”. In: *Human molecular genetics* 22.16 (2013), pp. 3347–3362.

[257] Lourdes Ortiz-Fernández and Amr H Sawalha. “Genetics of Behcet’s disease: functional genetic analysis and estimating disease heritability”. In: *Frontiers in Medicine* 8 (2021), p. 625710.

[258] Maria Gutierrez-Arcelus, Stephen S Rich, and Soumya Raychaudhuri. “Autoimmune diseases—connecting risk alleles with molecular traits of the immune system”. In: *Nature Reviews Genetics* 17.3 (2016), pp. 160–174.

[259] Simon Makin. “Cracking the genetic code of autoimmune disease”. In: *Nature* 595.7867 (2021), S57–S59.

[260] Ricky Lali et al. “Calibrated rare variant genetic risk scores for complex disease prediction using large exome sequence repositories”. In: *Nature Communications* 12.1 (2021), p. 5852.

[261] Omer Weissbrod et al. “Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores”. In: *Nature Genetics* 54.4 (2022), pp. 450–458.

[262] Imogen S Stafford et al. “A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases”. In: *NPJ digital medicine* 3.1 (2020), p. 30.

[263] Justine A Ellis, Andrew S Kemp, and Anne-Louise Ponsonby. “Gene–environment interaction in autoimmune disease”. In: *Expert Reviews in Molecular Medicine* 16 (2014), e4.

[264] CL White et al. “Crohn’s disease and ulcerative colitis in the same patient.” In: *Gut* 24.9 (1983), pp. 857–862.

[265] Massimo Castro et al. “Crohn’s disease and ulcerative colitis with onset in the same year in two sisters.” In: *Journal of pediatric gastroenterology and nutrition* 23 3 (1996), pp. 316–9. URL: https://api.semanticscholar.org/CorpusID:38357416.

[266] Hao Zhou et al. “Reverse causation between multiple sclerosis and psoriasis: a genetic correlation and Mendelian randomization study”. In: *Scientific Reports* 14.1 (2024), p. 8845.

[267] Akari Suzuki et al. “Insight from genome-wide association studies in rheumatoid arthritis and multiple sclerosis”. In: *FEBS letters* 585.23 (2011), pp. 3627–3632.

[268] María Teruel and Marta Eugenia Alarcón-Riquelme. “The genetic basis of systemic lupus erythematosus: What are the risk factors and what have we learned.” In: *Journal of autoimmunity* 74 (2016), pp. 161–175. URL: https://api.semanticscholar.org/CorpusID:24559773.

[269] Sophia Kerns et al. *705 Examination of the shared genetic architecture between multiple sclerosis and systemic lupus erythematosus facilitates discovery of novel lupus risk loci*. 2024.

[270] Haojie Lu et al. “Detection of genetic overlap between rheumatoid arthritis and systemic lupus erythematosus using GWAS summary statistics”. In: *Frontiers in genetics* 12 (2021), p. 656545.

[271] Sara K Tedeschi, Bonnie Bermas, and Karen H Costenbader. “Sexual disparities in the incidence and course of SLE and RA”. In: *Clinical Immunology* 149.2 (2013), pp. 211–218.

[272] Deena Khan and S Ansar Ahmed. “The immune system is a natural target for estrogen action: opposing effects of estrogen in two prototypical autoimmune diseases”. In: *Frontiers in immunology* 6 (2016), p. 635.

[273] Camille M. Syrett and Montserrat C. Anguera. “When the balance is broken: X-linked gene dosage from two X chromosomes and female-biased autoimmunity”. In: *Journal of Leukocyte Biology* 106 (2019), pp. 919 –932. URL: https://api.semanticscholar.org/CorpusID:164217767.

[274] David Hägg et al. “Severity of psoriasis differs between men and women: a study of the clinical outcome measure psoriasis area and severity index (PASI) in 5438 Swedish register patients”. In: *American journal of clinical dermatology* 18 (2017), pp. 583–590.

[275] Lucie Abeler-Dörner et al. “Butyrophilins: an emerging family of immune regulators.” In: *Trends in immunology* 33 1 (2012), pp. 34–41. URL: https://api.semanticscholar.org/CorpusID:13958226.

[276] Babajan Banaganapalli et al. “Exploring celiac disease candidate pathways by global gene expression profiling and gene network cluster analysis”. In: *Scientific reports* 10.1 (2020), p. 16290.

[277] Nick Huang and Andras Perl. “Metabolism as a target for modulation in autoimmune diseases”. In: *Trends in immunology* 39.7 (2018), pp. 562–576.

[278] Faheem Ahmed et al. “Drug repurposing in psoriasis, performed by reversal of disease-associated gene expression profiles”. In: *Computational and Structural Biotechnology Journal* 20 (2022), pp. 6097–6107.

[279] Francesco Moro et al. “Pemphigus: trigger and predisposing factors”. In: *Frontiers in Medicine* 10 (2023), p. 1326359.

[280] Jean-Laurent Casanova and Laurent Abel. “Revisiting Crohn’s disease as a primary immunodeficiency of macrophages”. In: *Journal of experimental medicine* 206.9 (2009), pp. 1839–1843.

[281] Khalaf Kridin et al. “Association between pemphigus and neurologic diseases”. In: *JAMA dermatology* 154.3 (2018), pp. 281–285.

[282] Takeshi Echigo et al. “Antiphospholipid antibodies in patients with autoimmune blistering disease”. In: *Journal of the American Academy of Dermatology* 57.3 (2007), pp. 397–400.

[283] Katharina Boch et al. “Low prevalence of late-onset neutropenia after rituximab treatment in patients with pemphigus”. In: *Journal of the American Academy of Dermatology* 83.6 (2020), pp. 1824–1825.

[284] DeLisa Fairweather et al. “Mechanisms underlying sex differences in autoimmunity”. In: *The Journal of clinical investigation* 134.18 (2024).

[285] Yikun Mou et al. “Clinical features in Juvenile-onset ankylosing spondylitis patients carrying different B27 subtypes”. In: *BioMed Research International* 2015.1 (2015), p. 594878.

[286] Michael D. Kornberg and Peter A. Calabresi. “Multiple Sclerosis and Other Acquired Demyelinating Diseases of the Central Nervous System.” In: *Cold Spring Harbor perspectives in biology* (2024). URL: https://api.semanticscholar.org/CorpusID:270094421.

[287] CRH Hedin et al. “Inflammatory bowel disease and psoriasis: modernizing the multidisciplinary approach”. In: *Journal of Internal Medicine* 290.2 (2021), pp. 257–278.