

**Γεφυρώνοντας μεγάλα ομικά, γενετικά και ιατρικά  
δεδομένα για την ευρεία εφαρμογή της Ιατρικής Ακριβείας  
στην Ελλάδα**

**ΠΑΡΑΔΟΤΕΟ Π8.1**

**«Τεχνική αναφορά με τις μεθοδολογίες και τα ενοποιημένα  
δεδομένα»**

<b>Φορέας</b>	Πανεπιστήμιο Θεσσαλίας
<b>Τύπος Παραδοτέου</b>	Έκθεση
<b>Ημερομηνία Υποβολής Παραδοτέου</b>	31 Δεκεμβρίου 2025
<b>Ενότητα Εργασίας</b>	Ενότητα Εργασίας 8: Βάση γενωμικών και βιοϊατρικών δεδομένων

1. Εισαγωγή.....	4
2. Βάση δεδομένων αιτιολογικών συσχετίσεων μέσω Μεντελιανής τυχαιοποίησης από GWAS .....	5
3. Βάση δεδομένων αιτιολογικών συσχετίσεων από συννοσηρότητες .....	8
4. Ανάπτυξη και Αξιολόγηση Μοντέλων Πρόβλεψης Κινδύνου Βιοχημικών Δεικτών.....	18
5. Συνθετικά γενετικά δεδομένα ασθενών .....	20
6. Ανάπτυξη του διαδικτυακού εργαλείου Flame (v2.0) για τη λειτουργική ανάλυση και ενοποίηση βιολογικών δεδομένων».....	22
7. Ανάπτυξη της βάσης δεδομένων neomefDB για την ταυτοποίηση και αξιοποίηση καρκινικών βιοδεικτών.....	22
8. Βιβλιογραφία .....	23



## Πίνακας Εικόνων

Εικόνα 1: Στιγμιότυπο της Βάσης Δεδομένων Μεντελιανής Τυχαιοποίησης.

Εικόνα 2: Δίκτυο ασθενειών των 48 κοινών αλληλεπιδράσεων (Common Hudine-Mendelian Network).

Εικόνα 3: Διάγραμμα διασποράς (scatter plot) που απεικονίζει τη συσχέτιση μεταξύ του μεγέθους επίδρασης (effect size) της Μενδελιανής Τυχαιοποίησης και της μερικής συσχέτισης (weightcor) της Ανάλυσης του Δικτύου Συννοσηρότητας.

## 1. Εισαγωγή

Στη συγκεκριμένη ΕΕ (ΕΕ8), σύμφωνα και με το τεχνικό παράρτημα του έργου, επιχειρήθηκε η ενοποίηση των δεδομένων από τα πακέτα εργασίας ΕΕ6 και ΕΕ7 τόσο με κλινικά δεδομένα ασθενών όσο και με δεδομένα τα οποία είναι ελεύθερα διαθέσιμα από άλλες πηγές (πχ γονιδιακής έκφρασης από GEO, ArrayExpress, TCGA). Σε αυτό το πλαίσιο, στο παραδοτέο Π8.1 πραγματοποιήσαμε μια συστηματική ανασκόπηση για την συγκέντρωση όλων των μελετών μεντελικής τυχαιοποίησης (**mendelian randomization**) για την εύρεση αιτιολογικών σχέσεων μεταξύ των ασθενειών (Ενότητα 2), ενώ αναπτύχθηκαν και μεθοδολογίες που επιτρέπουν την ενσωμάτωση επιδημιολογικών δεδομένων από συννοσηρότητες (**comorbidities**), οι οποίες είναι διαθέσιμες από μεγάλα δείγματα ασθενών, όπως το **HuDiNe** (Ενότητα 3). Η ανάλυση των συνοπτικών δεδομένων μας έδωσε ένα μεγάλο δείγμα συσχετίσεων που οργανώθηκαν σε βάση δεδομένων και μπορούν να χρησιμοποιηθούν στον ιατρικό φάκελο. Με τον τρόπο αυτό θα μπορεί να υπάρξει πληροφόρηση για την πιθανή εμφάνιση άλλων ασθενειών σε έναν ασθενή με δεδομένη διάγνωση και αυτή η πληροφορία θα μπορεί να δίνεται επιπλέον της απλής πρόβλεψης με βάση το γενετικό του προφίλ. Στην Ενότητα 4 περιγράφονται τα βιοχημικά δεδομένα που χρησιμοποιήθηκαν για την εισαγωγή στον ιατρικό φάκελο, και οι μέθοδοι ανάλυσης αυτών, ενώ στην Ενότητα 5 περιγράφονται τα γενετικά δεδομένα που χρησιμοποιήθηκαν με τις αναλύσεις της ΕΕ7 και το πως οργανώθηκαν σε βάση δεδομένων για την εισαγωγή στον ιατρικό φάκελο ασθενούς.

## 2. Βάση δεδομένων αιτιολογικών συσχετίσεων μέσω Μεντελιανής τυχαιοποίησης από GWAS

Στην παρούσα ενότητα του Έργου πραγματοποιήθηκε η συλλογή όλων των δημοσιευμένων άρθρα της βάσης δεδομένων Pubmed (Canese & Weis, 2013), σχετικά με τη Μεντελιανή Τυχαιοποίηση (Mendelian Randomization - MR). Ως κλειδί αναζήτησης στη βάση δεδομένων χρησιμοποιήθηκε το «mendelian randomization», ενώ τα άρθρα που ανακτήθηκαν από την αναζήτηση αφορούν το χρονολογικό εύρος μεταξύ 1999 και 2022.

Τα δημοσιευμένα άρθρα τα οποία συλλέχθηκαν περιέχουν συνοπτικά δεδομένα (summary statistics) από τη βάση δεδομένων GWAS Catalog (Welter et al., 2014). Πρόκειται για συνοπτικά αποτελέσματα συσχέτισης μεταξύ εκατομμυρίων γενετικών παραλλαγών (Single Nucleotide Polymorphisms - SNPs) και ενός φαινοτύπου, και περιλαμβάνουν συνήθως τις εκτιμήσεις των β-συντελεστών, τα standard errors, τα p-values, τις συχνότητες αλληλομόρφων και το sample size κάθε μελέτης. Αυτά τα δεδομένα, επειδή είναι απρόσωπα και δεν απαιτούν πρόσβαση σε ατομικού επιπέδου γενετικά αρχεία, έχουν επιτρέψει την ευρεία διάθεση και επαναχρησιμοποίησή τους σε ποικίλες δευτερογενείς αναλύσεις μεγάλης κλίμακας. Η χρήση των συνοπτικών δεδομένων έχει καταστεί θεμελιώδης στη σύγχρονη γενετική επιδημιολογία, επιτρέποντας την εφαρμογή μεθόδων όπως η Μεντελιανή Τυχαιοποίηση (Burgess & Thompson, 2015), η Ανάλυση Γενετικής Συσχέτισης (genetic correlation analysis) (Werme et al., 2022), οι Πολυγονιδιακοί Δείκτες Κινδύνου (Polygenic Risk Scores - PRS) (Chatterjee et al., 2016), η Λεπτομερής Γενετική Χαρτογράφηση (fine-mapping) (Spain & Barrett, 2015), καθώς και οι μετα-αναλύσεις GWAS (Hedges, 1992) σε πολυεθνικά δείγματα. Ωστόσο, η δυνατότητα συνδυασμού δεδομένων από ανεξάρτητες μελέτες αυξάνει δραματικά τη στατιστική ισχύ, μειώνει το κόστος και καθιστά εφικτή τη διερεύνηση σύνθετων φαινοτύπων, σπάνιων νοσημάτων ή υποομάδων του πληθυσμού.

Η Μεντελιανή Τυχαιοποίηση αποτελεί μια ισχυρή αναλυτική προσέγγιση που χρησιμοποιεί γενετικές παραλλαγές ως εργαλεία (instrumental variables) για να εκτιμήσει αιτιώδεις επιδράσεις μεταξύ ενός εκθέτη (exposure) και ενός αποτελέσματος (outcome), βασιζόμενη στην τυχαιοποίηση των αλληλομόρφων και στη σταθερότητά τους καθ' όλη τη διάρκεια της ζωής (Davies et al., 2018). Η ευρεία διαθεσιμότητα συνοπτικών δεδομένων από μεγάλες μελέτες GWAS έχει επιτρέψει την εφαρμογή της Μεντελιανής Τυχαιοποίησης χωρίς πρόσβαση σε ατομικά δεδομένα. Στην Μεντελιανή Τυχαιοποίηση ενός δείγματος (1-sample MR), οι γενετικές συσχετίσεις με τον εκθέτη και το αποτέλεσμα λαμβάνονται από το ίδιο δείγμα. Αντίθετα, στην Μεντελιανή Τυχαιοποίηση δύο δειγμάτων (2-sample MR), οι συσχετίσεις των πολυμορφισμών με τον εκθέτη και το αποτέλεσμα προέρχονται από δύο πλήρως ανεξάρτητα δείγματα. Αυτό το πλαίσιο επιτρέπει μεγαλύτερη στατιστική ισχύ, χρήση πολύ μεγαλύτερων δειγμάτων (meta-GWAS consortia) και μειώνει την πιθανότητα μεροληψίας από την επικάλυψη δειγμάτων (Sanderson et al., 2019). Η αιτιώδης εκτίμηση προκύπτει συνήθως μέσω μεθόδων όπως Inverse-Variance Weighting (IVW), MR-Egger, weighted median και mode-based estimators, οι οποίες εφαρμόζονται απευθείας στα συνοπτικά δεδομένα, λαμβάνοντας υπόψη την ανεξαρτησία των πολυμορφισμών - SNPs μέσω LD clumping ή LD matrices (Bowden et al., 2015).

Η μετα-ανάλυση στη Μεντελιανή Τυχαιοποίηση αποτελεί μια επέκταση, στην οποία τα αποτελέσματα διαφορετικών αναλύσεων - 2-sample MR ή 1-sample MR - συνδυάζονται, επιτρέποντας την εκτίμηση της αιτιότητας με μεγαλύτερη ακρίβεια και τη γενικευσιμότητα σε διαφορετικούς πληθυσμούς. Συνήθως, πραγματοποιείται η χρήση fixed ή random-effects μοντέλων, τα οποία ενσωματώνουν την ετερογένεια μεταξύ των συνόλων δεδομένων (datasets) και την πιθανή διακύμανση στις συσχετίσεις μεταξύ των SNP και των χαρακτηριστικών (traits).

Τέλος, στην περίπτωση αμφίδρομων αναλύσεων Μεντελιανής Τυχαιοποίησης (bidirectional MR) εξετάζουμε την αιτιότητα προς και τις δύο κατευθύνσεις, διεξάγοντας δύο ανεξάρτητες αναλύσεις Μεντελιανής Τυχαιοποίησης ( $A \rightarrow B$  και  $B \rightarrow A$ ) με διαφορετικά σετ πολυμορφισμών, επιλεγμένα από μελέτες GWAS για τον αντίστοιχο εκθέτη-exposure. Η μέθοδος αυτή είναι ιδιαίτερα χρήσιμη όταν ο εκθέτης και το αποτέλεσμα συνδέονται φαινοτυπικά, επιτρέποντας την αποσαφήνιση του αν οι παρατηρούμενες συσχετίσεις οφείλονται σε πραγματική αιτιότητα (reverse causation) ή σε υποκείμενη γενετική σύγχυση (genetic confounding) (Verbanck et al., 2018). Συνολικά, η χρήση συνοπτικών δεδομένων από μεγάλες GWAS έχει καταστήσει τη Μεντελιανή Τυχαιοποίηση ένα από τα πιο ισχυρά διαθέσιμα εργαλεία για την αιτιώδη κατανόηση βιολογικών μηχανισμών και παραγόντων κινδύνου σε ανθρώπινες νόσους.

Συγκεκριμένα, από τη βιβλιογραφική ανασκόπηση που πραγματοποιήθηκε, συλλέχθηκαν συνολικά 5.994 άρθρα, εκ των οποίων τα 3.993 αφαιρέθηκαν από την ανάλυση για ποικίλους λόγους (απουσία συνοπτικών δεδομένων, μεντελιανή κληρονομικότητα και όχι τυχαιοποίηση, κ.ά.). Στη συνέχεια, στα εναπομείναντα 2.001 άρθρα, έγινε μια εκτενής καταγραφή των ακόλουθων δεδομένων:

- pubmed id (μοναδικό αναγνωριστικό του άρθρου στη βάση δεδομένων Pubmed),
- doi,
- ονοματεπώνυμο πρώτου συγγραφέα,
- τίτλος άρθρου,
- περιοδικό δημοσίευσης,
- έτος δημοσίευσης άρθρου,
- φαινότυπος 1 (εκθέτης – exposure),
- icd10 για τον φαινότυπο 1,
- Phecode για τον φαινότυπο 1,
- Phecode κατηγορία για τον φαινότυπο 1,
- φαινότυπος 2 (αποτέλεσμα-outcome),
- icd10 για τον φαινότυπο 2,
- Phecode για τον φαινότυπο 2,
- Phecode κατηγορία για τον φαινότυπο 2,
- αριθμός πολυμορφισμών που συμμετέχουν στη συσχέτιση των δυο φαινοτύπων,
- πολυμορφισμοί,
- είδος μελέτης (1 sample MR, 2 sample MR, meta-analysis),
- p-value,
- z-score,
- odds ratios (μετασχηματισμός σε λογαριθμική κλίμακα),

- beta coefficients,
- διαστήματα εμπιστοσύνης (95% confidence intervals).

Επιπλέον, για κάθε μελέτη που αναλύθηκε προστέθηκαν και σχόλια σχετικά με την πηγή προέλευσης των δεδομένων (consortia, IVW method, κ.ο.κ.). Ενώ, η καταγραφή των δεδομένων έγινε σε αρχείο excel. Μετά την ολοκλήρωση της καταγραφής, πραγματοποιήθηκε η συλλογή όλων των καταγραφών για κάθε μελέτη, καταλήγοντας σε ένα σύνολο 19.625 συσχετίσεις φαινοτύπων, πολλές από τις οποίες όμως είναι καταγραφές της ίδιας συσχέτισης από διαφορετικές μελέτες.

Απώτερος σκοπός της παρούσας ανάλυσης είναι η δημιουργία ενός δικτύου ασθενειών (disease-disease associations). Για αυτό τον σκοπό αφαιρέθηκαν οι καταγραφές που δεν κατηγοριοποιούνται σύμφωνα με το πρότυπο ICD10, όπως χαρακτηριστικά που αναφέρονται σε ποσοτικές μετρήσεις για σωματομετρικά χαρακτηριστικά, πρωτεΐνες, κύτταρα, μεταβολίτες και μικροβίωμα. Η δημιουργία δικτύου ασθενειών με βάση τη Μεντελιανή Τυχαιοποίηση αφορά στην κατασκευή ενός γραφήματος δικτύου (network), στο οποίο οι κόμβοι αντιπροσωπεύουν τις ασθένειες ή τους φαινότυπους, ενώ οι ακμές αντιστοιχούν σε αιτιώδεις σχέσεις που εκτιμώνται χρησιμοποιώντας τα συνοπτικά δεδομένα από τις GWAS. Με την εφαρμογή της Μεντελιανής Τυχαιοποίησης για κάθε ζεύγος έκθεσης-φαινοτύπου, μπορούν να εντοπιστούν σημαντικές αιτιώδεις επιδράσεις μεταξύ ασθενειών (π.χ.  $A \rightarrow B$ ), οι οποίες στη συνέχεια μετατρέπονται σε κατευθυνόμενες ακμές. Το αποτέλεσμα είναι ένα αιτιώδες δίκτυο ασθενειών, το οποίο επιτρέπει την αναγνώριση κεντρικών παθοφαινοτύπων, κοινών βιολογικών μηχανισμών και μοτίβων πολυνοσηρότητας. Ένα τέτοιο δίκτυο μπορεί να ενισχυθεί με τη χρήση αμφίδρομων μεντελιανών τυχαιοποιήσεων (bidirectional MR), αναλύσεις ευαισθησίας (sensitivity analyses π.χ. Egger, weighted median, MR-PRESSO) και χρήσης κατωφλίων στα p-values ή στα causal effect estimates, ώστε να διατηρούνται μόνο οι πιο αξιόπιστες αιτιώδεις συνδέσεις.

Στην συνέχεια, επιλέχθηκαν οι μοναδικές συσχετίσεις, αφαιρώντας δηλαδή καταγραφές που μελετούσαν την ίδια συσχέτιση από διαφορετικές μελέτες και πραγματοποιήθηκε ανάλυση των μοναδικών στατιστικά σημαντικών συσχετίσεων, διατηρώντας ως κατώφλι  $<0.05$  στα p-values, κάτι που οδήγησε στην αξιοποίηση 987 συσχετίσεων φαινοτύπων από τις 3.141 που είχαμε αρχικά πριν την εφαρμογή του κατωφλίου. Από τις τελικές καταγραφές προέκυψαν 581 μοναδικά ζευγάρια ασθενειών.

Τέλος, αναπτύχθηκε μια ολοκληρωμένη βάση δεδομένων η οποία περιλαμβάνει το σύνολο των 3.141 φαινοτυπικών συσχετίσεων, ενσωματώνοντας όλες τις διαθέσιμες πληροφορίες από το αρχείο δεδομένων Excel. Η βάση δεδομένων είναι διαθέσιμη στην ιστοσελίδα <https://gomedprecision.gr/DB/mendelianDB.html>.

Καμία γρήγορη στην πρώτη ολοκληρωμένη φωνητική απόδοση απολογιστικών συστατικών που προέρχουν μόνο της μεθόδου της Μεντελιανής Τυχαιοποίησης (Mendelian Randomization - MR). Η παρούσα πλατφόρμα αποτελεί ένα εργαλείο αναζήτησης και ανάλυσης δεδομένων, με στόχο την κατανοήση των βιολογικών μηχανισμών και των παραγόντων κινδύνου που συνδέουν διαφορετικά ανθρώπινα σύνδεσμοι. Η βάση δεδομένων φιλοξενεί 3.141 καταγραφές γενετικών συστατικών, απόγονος αποκλειστικά σε όλη σχεδόν των έθνη κωδικοποιηθεί με το σύστημα ICD-10, επιτρέποντας τη δημιουργία ενός αλυσίδων "Δίκτυου Ασθενειών" (Disease Network).

ΔΕΔΟΜΕΝΑ

Pubmed ID	link	FSM	First_Author	Journal/Book	Publication_Year	Disease	Phenotype	Lead_SNP	Phenotype	trait2	ICD10	ICD10_S	S	Number_of_Participants	Type_of_Study	p-value	z	Cau.	Effect_Size	Effect_Size_Type	SE	LL	UL	Adjusted	Comm.	Phacode	PhacodeString1	PhacodeCategory1	Phacode	PhacodeString2	PhacodeCategory2	
329667	N. Car.	Ponaford	HJ	Circulation	2020	1.	Body Mass.	Body	Z68	Covid-19	Emer.	U07		3815	15MR	0.003	2.	YES	0.5596157	OR	1.	2.				IVWMe_EK_23	Underw.	Endocrine/Metab	ID_359.1	Sars-Co.	Infections	
329667	N. Car.	Ponaford	HJ	Circulation	2020	1.	Lifetime S.	Probl.	Z72	Covid-19	Emer.	U07		3815	15MR	0.031	2.	YES	1.3711807	OR	1.	1.				IVWMe_MB_28	Current	Mental	ID_359.1	Sars-Co.	Infections	
329667	N. Car.	Ponaford	HJ	Circulation	2020	1.	Type 2 Dia.	Type	E11	Covid-19	Emer.	U07		3815	15MR	0.907	0.	NO	0.0099503	OR	0.	1.				IVWMe_ID_091	Gangrene	Infections	ID_359.1	Sars-Co.	Infections	
329667	N. Car.	Ponaford	HJ	Circulation	2020	1.	Body Mass.	Body	Z68	Covid-19	Emer.	U07		900687	15MR	0.000	3.	YES	0.3822624	OR	1.	1.				IVWMe_EK_23	Underw.	Endocrine/Metab	ID_359.1	Sars-Co.	Infections	
329667	N. Car.	Ponaford	HJ	Circulation	2020	1.	Lifetime S.	Probl.	Z72	Covid-19	Emer.	U07		900687	15MR	0.000	4.	YES	1.4516138	OR	2.	0.				IVWMe_MB_28	Current	Mental	ID_359.1	Sars-Co.	Infections	
329667	N. Car.	Ponaford	HJ	Circulation	2020	1.	Type 2 Dia.	Type	E11	Covid-19	Emer.	U07		900687	15MR	1	0	NO	0	OR	0.	1.				IVWMe_ID_091	Gangrene	Infections	ID_359.1	Sars-Co.	Infections	
329437	N. Inve.	Yew	YH	Sci Rep	2020	1.	Body Mass.	Body	Z68	Atopic Der.	Atopl.	L20	941	rs_797275	25MR	0.015	2.	YES	0.0769610	OR	1.	1.				IVWMe_EK_23	Underw.	Endocrine/Metab	DE_668	Dermatol.	Dermatological	
338267	N. Cas.	Park	S	Heprol Dis.	2021	1.	Overall act.	Activ.	Y93	Chronic KI.	Chron.	R18	5	rs_749960	25MR	0.001	2.	YES	-0.713349	OR	0.	0.				IVWMe...			GU_382.2	Chronic	Gastroenterary	
329267	N. Cas.	Park	S	Heprol Dis.	2021	1.	Tv watching	Activ.	Y93	Chronic KI.	Chron.	R18	152	rs_749960	25MR	0.007	2.	YES	0.1822313	OR	1.	1.				IVWMe...			GU_382.2	Chronic	Gastroenterary	
338267	N. Cas.	Park	S	Heprol Dis.	2021	1.	Using com.	Activ.	Y93	Chronic KI.	Chron.	R18	37	rs_749960	25MR	0.849	0.	NO	0.0094939	OR	0.	1.				IVWMe...			GU_382.2	Chronic	Gastroenterary	
338217	N. Sise.	Caryena	HX	World J Biol.	2021	1.	Insomnia	Sleep.	G47	Attention	Atten.	F90	40	rs_508753	25MR	0.018	2.	YES	0.87	BETA	0.	1.				IVWMe...	NE_333	Sleep di.	Neurological	HL_304	Attention.	Mental
338217	N. Sise.	Caryena	HX	World J Biol.	2021	1.	Snoring	Abnor.	R06	Attention	Atten.	F90	30	rs_415290	25MR	0.366	0.	NO	0.096	BETA	0.	1.				IVWMe...	BE_488	Abnorm.	Respiratory	HL_304	Attention.	Mental
338217	N. Sise.	Caryena	HX	World J Biol.	2021	1.	Daytime S.	Probl.	Z72	Attention	Atten.	F90	43	rs_507445	25MR	0.722	0.	NO	-0.17	BETA	0.	1.				IVWMe...	MB_28	Current	Mental	HL_304	Attention.	Mental
338217	N. Sise.	Caryena	HX	World J Biol.	2021	1.	Attention	Atten.	F90	Insomnia	Sleep.	G47	10	rs_508753	25MR	0.822	0.	NO	0.004	BETA	0.	1.				IVWMe...	MB_304	Attention.	Mental	NE_333	Sleep di.	Neurological
338217	N. Sise.	Caryena	HX	World J Biol.	2021	1.	Attention	Atten.	F90	Snoring	Abnor.	R06	10	rs_415290	25MR	0.307	1.	NO	-0.034	BETA	0.	1.				IVWMe...	MB_304	Attention.	Mental	BE_488	Abnorm.	Respiratory
338217	N. Sise.	Caryena	HX	World J Biol.	2021	1.	Attention	Atten.	F90	Daytime S.	Probl.	Z72	10	rs_507445	25MR	0.678	0.	NO	0.003	BETA	0.	1.				IVWMe...	MB_304	Attention.	Mental	HL_382.1	Current	Mental
320347	N. Coff.	Qian	Y	Ann Neurol	2020	1.	Caffeine C.	Other.	F15	Stroke	Cereb.	363	3	rs_565250	meta-analysis	0.918	0	NO	0	OR	0.	1.				IVWMe...	MB_28	Stimula.	Mental	CV_431.11	Cerebral.	Cardiovascular
320347	N. Coff.	Qian	Y	Ann Neurol	2020	1.	Caffeine C.	Other.	F15	Ischemic C.	Cereb.	363	3	rs_562152	meta-analysis	0.344	0.	NO	-0.030459	OR	0.	1.				IVWMe...	MB_28	Stimula.	Mental	CV_431.11	Cerebral.	Cardiovascular
320347	N. Coff.	Qian	Y	Ann Neurol	2020	1.	Caffeine C.	Other.	F15	Large vess.	Cereb.	363	3	rs_532308	meta-analysis	0.567	0.	NO	0.0487901	OR	0.	1.				IVWMe...	MB_28	Stimula.	Mental	CV_431.11	Cerebral.	Cardiovascular
320347	N. Coff.	Qian	Y	Ann Neurol	2020	1.	Caffeine C.	Other.	F15	Small vess.	Cereb.	363	3	rs_533321	meta-analysis	0.056	0.	NO	-0.162518	OR	0.	1.				IVWMe...	MB_28	Stimula.	Mental	CV_431.11	Cerebral.	Cardiovascular
320347	N. Coff.	Qian	Y	Ann Neurol	2020	1.	Caffeine C.	Other.	F15	Stroke	Cereb.	363	8	rs_565250	meta-analysis	0.881	0.	NO	-0.102050	OR	0.	1.				IVWMe...	MB_28	Stimula.	Mental	CV_431.11	Cerebral.	Cardiovascular
320347	N. Coff.	Qian	Y	Ann Neurol	2020	1.	Caffeine C.	Other.	F15	Ischemic C.	Cereb.	363	8	rs_562152	meta-analysis	0.235	0.	NO	-0.061875	OR	0.	1.				IVWMe...	MB_28	Stimula.	Mental	CV_431.11	Cerebral.	Cardiovascular
320347	N. Coff.	Qian	Y	Ann Neurol	2020	1.	Caffeine C.	Other.	F15	Large vess.	Cereb.	363	8	rs_532308	meta-analysis	0.669	0.	NO	0.052689	OR	0.	1.				IVWMe...	MB_28	Stimula.	Mental	CV_431.11	Cerebral.	Cardiovascular
320347	N. Coff.	Qian	Y	Ann Neurol	2020	1.	Caffeine C.	Other.	F15	Small vess.	Cereb.	363	8	rs_533321	meta-analysis	0.149	0.	NO	-0.186329	OR	0.	1.				IVWMe...	MB_28	Stimula.	Mental	CV_431.11	Cerebral.	Cardiovascular
318213	N. Ho.	Wanberg	H	PLoS Med	2019	1.	Overweigh.	Over.	E66	Diabetes	Other.	E13	57	rs_212618	25MR	4.26	9.	YES	0.3074846	OR	1.	1.				IVWMe...	EK_236	Overwe.	Endocrine/Metab	ID_391	Gangrene	Infections
318213	N. Ho.	Wanberg	H	PLoS Med	2019	1.	Disease (BP.	Asthma	J45	Diabetes	Other.	E13	57	rs_190768	25MR	1.956	9.	YES	0.2231435	OR	1.	1.				IVWMe...	KE_475	Asthma	Respiratory	ID_391	Gangrene	Infections
318021	N. App.	Kameli	N	Int J Epide.	2020	1.	BMI	Body	Z68	Prostate C.	Malig.	C61	535	rs_821529	25MR	0.003	2.	YES	-0.105360	OR	0.	1.				IVWMe...	EK_23	Underw.	Endocrine/Metab	CA_107.2	Malig...	Neoplasms
318021	N. App.	Kameli	N	Int J Epide.	2020	1.	Alcohol	Alcohol.	F10	Prostate C.	Malig.	C61	77	rs_1081534	25MR	0.77	0.	NO	-0.040821	OR	0.	1.				IVWMe...	MB_28	Alcohol	Mental	CA_107.2	Malig...	Neoplasms

Εικόνα 1 : Στιγμιότυπο της Βάσης Δεδομένων Μεντελιανής Τυχαιοποίησης.

### 3. Βάση δεδομένων αιτιολογικών συσχετίσεων από συννοσηρότητες

Η κατανόηση των σχέσεων μεταξύ των ασθενειών είναι ζωτικής σημασίας για την αποκάλυψη αιτιολογικών μηχανισμών και τη βελτίωση των στρατηγικών πρόληψης. Παραδοσιακά, τα δίκτυα ασθενειών (diseasomes) βασίζονται στη συσχέτιση κατά Pearson για να συμπεράνουν αλληλεπιδράσεις μεταξύ ασθενειών. Ωστόσο, η μέθοδος αυτή αδυνατεί να λάβει υπόψη συγχυτικούς παράγοντες (confounders), καθιστώντας την ακατάλληλη για ασφαλή αιτιακή συμπερασματολογία. Η παρούσα μελέτη εισάγει μια καινοτόμο προσέγγιση χρησιμοποιώντας τη **μερική συσχέτιση (partial correlation)**, η οποία ποσοτικοποιεί τη σχέση μεταξύ δύο μεταβλητών ελέγχοντας την επίδραση των υπόλοιπων, προσφέροντας έτσι μια πιο ακριβή εικόνα των άμεσων και δυνητικά αιτιακών σχέσεων.

Η έρευνα χρησιμοποίησε δεδομένα από τη βάση Human Disease Network (HuDiNe), η οποία περιλαμβάνει πάνω από 291.000 συσχετίσεις μεταξύ 995 ασθενειών, βασισμένες σε ιατρικά αρχεία περισσότερων από 30 εκατομμυρίων ασθενών του συστήματος Medicare. Πραγματοποιήθηκε κωδικοποίηση των ονομάτων των ασθενειών με το πρότυπο ICD-10, ολοκληρώνοντας έτσι τη συνολική ανάλυση των χιλιάδων αλληλεπιδράσεων. Εφαρμόστηκε ένας εκτιμητής συρρίκνωσης τύπου James-Stein για τον υπολογισμό του πίνακα συνδιακύμανσης, επιτρέποντας τη διαχείριση μεγάλου όγκου δεδομένων («μικρό n, μεγάλο p»). Χρησιμοποιήθηκε το όριο FDR < 0,01 για τον προσδιορισμό των στατιστικά σημαντικών συσχετίσεων. Η ανάλυση και η οπτικοποίηση πραγματοποιήθηκαν με τη χρήση της γλώσσας R και του λογισμικού Cytoscape.

Το τελικό δίκτυο μερικής συσχέτισης ανέδειξε μια πιο «καθαρή» εικόνα σε σύγκριση με τα παραδοσιακά μοντέλα. Εντοπίστηκαν 7.894 σημαντικές συσχετίσεις μεταξύ 697 ασθενειών.

Το δίκτυο χαρακτηρίζεται ως αραιό (density 0.03) και παρουσιάζει μια τοπολογία «ελεύθερης κλίμακας» (scale-free), όπου λίγες ασθένειες λειτουργούν ως κεντρικοί κόμβοι (hubs). Το δίκτυο Pearson ήταν πολύ πιο πυκνό (23.424 συσχετίσεις), περιλαμβάνοντας όμως πολλές έμμεσες ή παραπλανητικές συνδέσεις που η μερική συσχέτιση κατάφερε να φιλτράρει. Η υπέρταση αναδείχθηκε ως ο πιο συνδεδεμένος κόμβος με 247 συσχετίσεις. Πολλές από αυτές τις συνδέσεις (π.χ. με την παχυσαρκία, τη χρόνια νεφρική νόσο, τον διαβήτη και την αθηροσκλήρωση) επιβεβαιώθηκαν μέσω μελετών Μεντελιανής Τυχαιοποίησης (MR). Η μέθοδος αποκάλυψε πιθανές συνδέσεις της υπέρτασης με παθήσεις όπως ο καρκίνος του μαστού, η ενδομητρίωση και το γλαύκωμα, οι οποίες υποστηρίζονται από επιδημιολογικά στοιχεία αλλά απαιτούν περαιτέρω διερεύνηση για την αιτιακή τους βάση. Ένα από τα πιο ενδιαφέροντα ευρήματα ήταν η ανίχνευση αρνητικών συσχετίσεων, οι οποίες μπορεί να υποδηλώνουν προστατευτική δράση. Χαρακτηριστικά παραδείγματα αποτελούν η αρνητική σχέση μεταξύ διαβήτη και ανευρύσματος αορτής, καθώς και μεταξύ υπέρτασης και άνοιας.

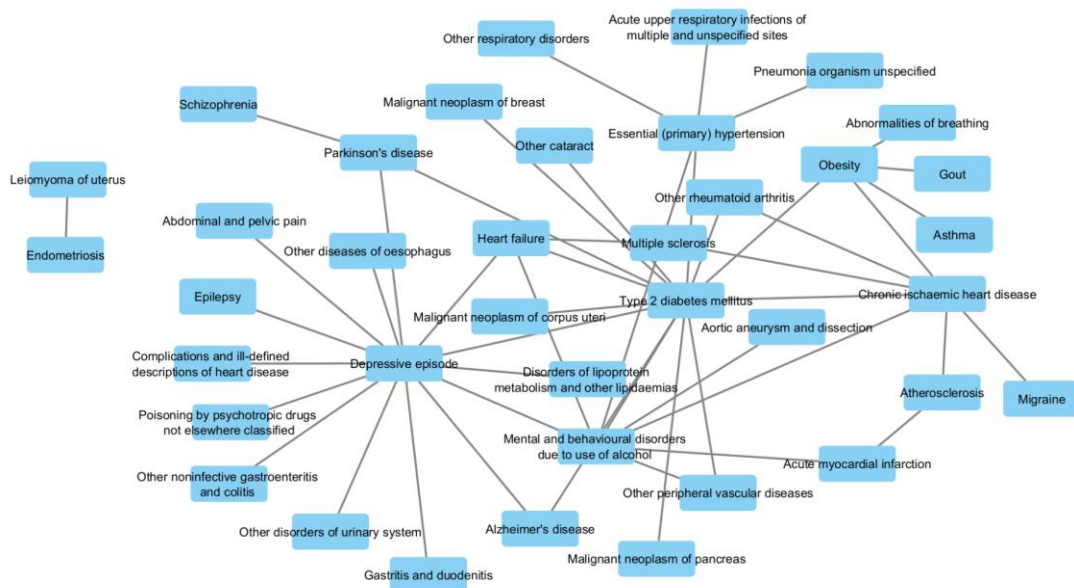
Η μελέτη αποδεικνύει ότι η ανάλυση μερικής συσχέτισης σε δεδομένα συνοπτικών στατιστικών (summary-level data) αποτελεί ένα ισχυρό εργαλείο για την κατασκευή αιτιακών δικτύων χωρίς την ανάγκη πρόσβασης σε ατομικά δεδομένα ασθενών. Επιπλέον, μειώνει τις ψευδείς συσχετίσεις και παρέχει μια πιο εκλεπτυσμένη εικόνα των αλληλεπιδράσεων μεταξύ των ασθενειών. Τα δεδομένα προέρχονται από ηλικιωμένο πληθυσμό των ΗΠΑ (Medicare), γεγονός που ίσως επηρεάζει τη γενίκευση των αποτελεσμάτων. Επίσης, η μέθοδος δεν προσδιορίζει την κατεύθυνση της αιτιότητας (ποια νόσος προκαλεί ποια). Μελλοντικά προτείνεται η ενσωμάτωση γενετικών δεδομένων (GWAS) και η χρήση πιο σύνθετων αλγορίθμων αιτιακής συμπερασματολογίας.

Στη συνέχεια, το αρχείο δεδομένων οπτικοποιήθηκε και ενσωματώθηκε σε μια **βάση δεδομένων**, η οποία περιέχει όλα τα αποτελέσματα της ανάλυσης, συμπεριλαμβανομένων των συντελεστών συσχέτισης Pearson, των συντελεστών μερικής συσχέτισης, καθώς και των αντίστοιχων τιμών p-value. Η βάση αυτή είναι πλέον διαθέσιμη στο ευρύ κοινό για περαιτέρω αναλύσεις από την επιστημονική κοινότητα και τους χρήστες.

Τα συνολικά αποτελέσματα της μελέτης έχουν παρουσιαστεί στο διεθνές επιστημονικό συνέδριο 12th International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO 2025, July 16th-18th, 2025, Gran Canaria, Spain) και έχουν δημοσιευτεί στον διεθνή συλλογικό τόμο (Kontou et al., 2026).

Αφού περιορίσαμε το δίκτυο συνοσηροτήτων σε ζεύγη ICD-10, προχωρήσαμε σε σύγκριση με το δίκτυο Μεντελιανής Τυχαιοποίησης. Μετά την ομογενοποίηση, το δίκτυο Hudine περιείχε 7.179 ακμές, ενώ το δίκτυο της Μεντελιανής Τυχαιοποίησης 581 ακμές. Η ένωση των κόμβων ανήλθε σε 688, ορίζοντας χώρο σύγκρισης  $M = n(n-1)/2 = 236.328$  δυνητικών ακμών. Η τομή των δικτύων περιείχε 48 κοινές ακμές, με 7.131 ακμές να είναι αποκλειστικές του Hudine και 531 αποκλειστικές του δικτύου Μεντελιανής Τυχαιοποίησης. ο

υπεργεωμετρικός έλεγχος κατέληξε στο συμπέρασμα ότι η παρατηρούμενη επικάλυψη είναι στατιστικά σημαντική με  $p\text{-value} = 6,95 \times 10^{-10}$ . Το γεγονός αυτό επιβεβαιώνει ότι η σχέση των δύο δικτύων δεν είναι τυχαία, αλλά υποδηλώνει ισχυρό κοινό βιολογικό υπόβαθρο. Στη συνέχεια εστιάζουμε στο υποδίκτυο των 48 κοινών ακμών για λεπτομερή οπτικοποίηση και λειτουργική ερμηνεία.



**Εικόνα 2:** Δίκτυο ασθενειών των 48 κοινών αλληλεπιδράσεων (Common Huidine-Mendelian Network).

Το κοινό υποδίκτυο (38 κόμβοι, 48 ακμές) χαρακτηρίζεται από χαμηλή πυκνότητα ( $\text{density} = 0,034$ ) και μέσο βαθμό  $\sim 2,53$ , στοιχείο που υποδηλώνει μια αραιή αλλά στοχευμένη συνδεσιμότητα. Παρά τη σπανιότητα των ακμών, η χαρακτηριστική απόσταση είναι μικρή (1,827) και η διάμετρος μόλις 4, άρα τα περισσότερα ζεύγη νοσημάτων απέχουν 1–2 βήματα. Ο πολύ χαμηλός συντελεστής συσσωμάτωσης ( $\text{clustering} = 0,060$ ) δείχνει ότι η δομή δεν είναι τυχαία, αλλά οργανωμένη γύρω από συγκεκριμένους κόμβους-hubs. Συγκεκριμένα, αναδεικνύονται ο Σακχαρώδης Διαβήτης Τύπου 2 (E11) και η Κατάθλιψη (F32) με 14 αλληλεπιδράσεις έκαστος, καθώς και οι Διαταραχές χρήσης Αλκοόλ (F10) με 8 αλληλεπιδράσεις. Η συγκέντρωση των ακμών γύρω από αυτούς τους φαινοτύπους υπογραμμίζει τον κεντρικό –και πιθανώς αιτιώδη– ρόλο της μεταβολικής απορρύθμισης και της ψυχικής υγείας, καθώς συνδέουν άμεσα τα συστήματα αυτά με σοβαρές συννοσηρότητες, όπως καρδιαγγειακές παθήσεις (I25, I50) και νεοπλασίες (C50, C25).

**Πίνακας 1:** Οι 48 κοινές ακμές μεταξύ των δικτύων Hudine και Μενδελιανής Τυχαιοποίησης. Συμπεριλαμβάνονται το μέγεθος επίδρασης (effect size) του δικτύου Μενδελιανής Τυχαιοποίησης και η μερική συσχέτιση (partial correlation) του δικτύου Hudine.

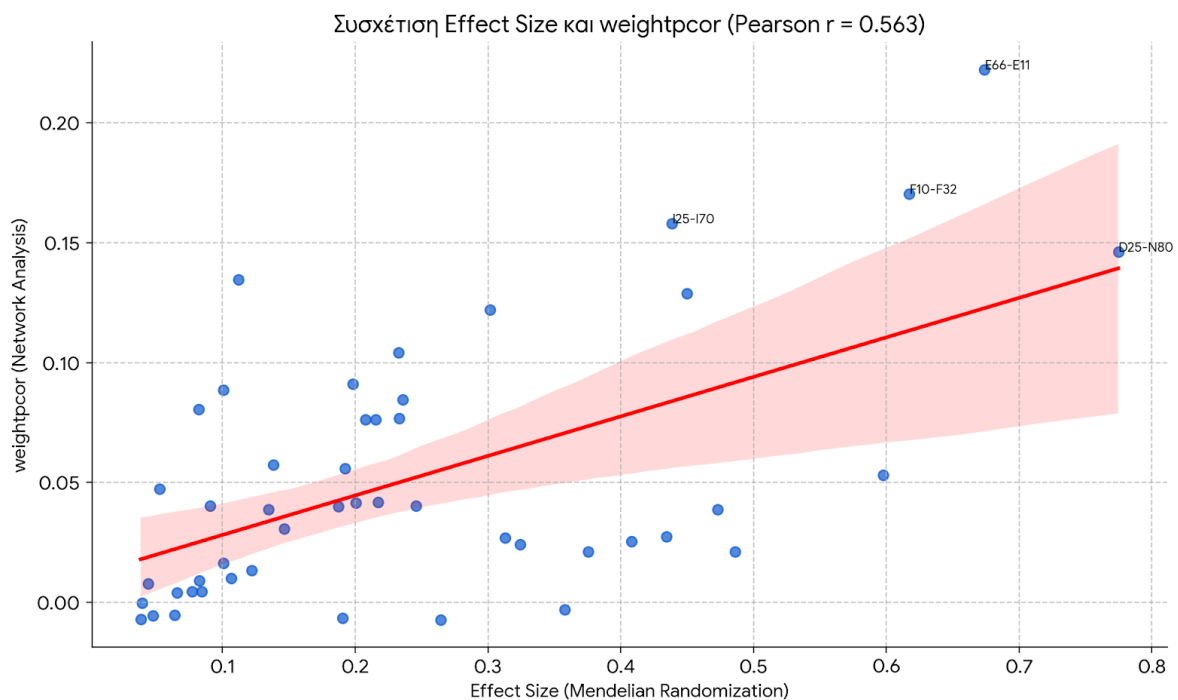
ICD10_1	Disease_Name_1	ICD10_2	Disease_Name_2	Effect Size	Partial Correlation
D25	Leiomyoma of uterus	N80	Endometriosis	0.7747	0.1462
E11	Type 2 diabetes mellitus	C25	Malignant neoplasm of pancreas	0.1222	0.0133
E11	Type 2 diabetes mellitus	C50	Malignant neoplasm of breast	0.3577	-0.0031
E11	Type 2 diabetes mellitus	C54	Malignant neoplasm of corpus uteri	0.0770	0.0045
E11	Type 2 diabetes mellitus	G20	Parkinson's disease	0.1906	-0.0067
E11	Type 2 diabetes mellitus	G30	Alzheimer's disease	0.0639	-0.0052
E11	Type 2 diabetes mellitus	H26	Other cataract	0.0906	0.0440
E11	Type 2 diabetes mellitus	I10	Essential (primary) hypertension	0.0823	0.0884
E11	Type 2 diabetes mellitus	I25	Chronic ischaemic heart disease	0.1008	0.0392
E11	Type 2 diabetes mellitus	I50	Heart failure	0.1120	0.0653
E11	Type 2 diabetes mellitus	I73	Other peripheral vascular diseases	0.1385	0.0359

E11	Type 2 diabetes mellitus	M06	Other rheumatoid arthritis	0.0478	-0.0122
E66	Obesity	E11	Type 2 diabetes mellitus	0.6739	0.0638
E66	Obesity	I25	Chronic ischaemic heart disease	0.0526	-0.0024
E66	Obesity	J45	Asthma	0.1873	0.0215
E66	Obesity	M10	Gout	0.2079	0.0151
E66	Obesity	R06	Abnormalities of breathing	0.2007	0.0180
E78	Disorders of lipoprotein metabolism and other lipidaemias	F32	Depressive episode	0.0440	-0.0054
F10	Mental and behavioural disorders due to use of alcohol	E11	Type 2 diabetes mellitus	0.5976	-0.0064
F10	Mental and behavioural disorders due to use of alcohol	F32	Depressive episode	0.6172	0.0106
F10	Mental and behavioural disorders due to use of alcohol	I10	Essential (primary) hypertension	0.4729	0.0073
F10	Mental and behavioural disorders due to use of alcohol	I21	Acute myocardial infarction	0.3756	0.0171
F10	Mental and behavioural	I25	Chronic ischaemic heart disease	0.4863	0.0123

	disorders due to use of alcohol				
F10	Mental and behavioural disorders due to use of alcohol	I50	Heart failure	0.4080	-0.0049
F10	Mental and behavioural disorders due to use of alcohol	I71	Aortic aneurysm and dissection	0.3243	0.0086
F10	Mental and behavioural disorders due to use of alcohol	I73	Other peripheral vascular diseases	0.3130	0.0161
F20	Schizophrenia	G20	Parkinson's disease	0.0846	0.0162
F32	Depressive episode	E11	Type 2 diabetes mellitus	0.4343	-0.0028
F32	Depressive episode	G20	Parkinson's disease	0.2642	0.0287
F32	Depressive episode	G30	Alzheimer's disease	0.1065	0.0226
F32	Depressive episode	G40	Epilepsy	0.2458	0.0073
F32	Depressive episode	I50	Heart failure	0.2174	0.0055
F32	Depressive episode	I51	Complications and ill-defined descriptions of heart disease	0.2331	0.0069
F32	Depressive episode	K22	Other diseases of oesophagus	0.1984	0.0123

F32	Depressive episode	K29	Gastritis and duodenitis	0.2359	0.0110
F32	Depressive episode	K52	Other noninfective gastroenteritis and colitis	0.2156	0.0086
F32	Depressive episode	N39	Other disorders of urinary system	0.1923	0.0215
F32	Depressive episode	R10	Abdominal and pelvic pain	0.2327	0.0106
F32	Depressive episode	T43	Poisoning by psychotropic drugs not elsewhere classified	0.3014	0.0204
G35	Multiple sclerosis	I25	Chronic ischaemic heart disease	0.0385	-0.0032
G35	Multiple sclerosis	I50	Heart failure	0.0396	-0.0037
G43	Migraine	I25	Chronic ischaemic heart disease	0.0659	-0.0029
I10	Essential (primary) hypertension	J06	Acute upper respiratory infections of multiple and unspecified sites	0.1009	0.0039
I10	Essential (primary) hypertension	J18	Pneumonia organism unspecified	0.1346	-0.0064
I10	Essential (primary) hypertension	J98	Other respiratory disorders	0.1465	-0.0048

121	Acute myocardial infarction	170	Atherosclerosis	0.4498	-0.0054
125	Chronic ischaemic heart disease	170	Atherosclerosis	0.4385	0.0185
125	Chronic ischaemic heart disease	M06	Other rheumatoid arthritis	0.0827	-0.0037



**Εικόνα 3:** Διάγραμμα διασποράς (scatter plot) που απεικονίζει τη συσχέτιση μεταξύ του μεγέθους επίδρασης (effect size) της Μενδελιανής Τυχαιοποίησης και της μερικής συσχέτισης (weightpcor) της Ανάλυσης του Δικτύου Συννοσηρότητας.

Το διάγραμμα διασποράς (Scatter Plot) αναδεικνύει μια θετική συσχέτιση (Pearson  $r \approx 0.56$ ) μεταξύ του γενετικού μεγέθους επίδρασης (Effect Size) και της σταθμισμένης φαινοτυπικής συσχέτισης (weightpcor). Η ανοδική τάση της γραμμής παλινδρόμησης υποδεικνύει ότι ζεύγη νοσημάτων με ισχυρότερη γενετική αιτιώδη σχέση τείνουν να εμφανίζουν και υψηλότερη συννοσηρότητα στα κλινικά δεδομένα. Ζεύγη που ξεχωρίζουν στο άνω δεξί τεταρτημόριο, όπως η Παχυσαρκία με τον Διαβήτη (E66-E11) και οι Διαταραχές Αλκοόλ με την Κατάθλιψη (F10-F32), επιβεβαιώνουν την ισχυρή βιολογική και κλινική τους σύνδεση,

καθώς παρουσιάζουν ταυτόχρονα υψηλό γενετικό ρίσκο και υψηλή συχνότητα συνύπαρξης στο δίκτυο.

#### 4. Ανάπτυξη και Αξιολόγηση Μοντέλων Πρόβλεψης Κινδύνου Βιοχημικών Δεικτών

Σε αυτό το κομμάτι της μελέτης επικεντρωθήκαμε σε τρεις βασικούς άξονες: τη συλλογή/δημιουργία βιοϊατρικών δεδομένων, την εκπαίδευση των προγνωστικών μοντέλων και την αξιολόγησή τους. Αρχικά, έγινε κωδικοποίηση των εξετάσεων βάσει του διεθνούς προτύπου Logical Observation Identifiers Names and Codes (LOINC). Για την εκπαίδευση των μοντέλων, χρησιμοποιήθηκαν οι βιοχημικές εξετάσεις που παρουσιάζονται στον Πίνακα 2.

**Πίνακας 2:** Κωδικοί εξετάσεων προς ανάλυση κατά πρότυπο LOINC.

Κωδικός LOINC	Όνομα εξέτασης
2085-9	Cholesterol in HDL
2093-3	Cholesterol
2160-0	Creatinine
2339-0	Glucose
2571-8	Triglyceride
3016-3	Thyrotropin
4548-4	Hemoglobin A1c
18262-6	Cholesterol in LDL

Για την εκπαίδευση των μοντέλων, χρησιμοποιήθηκε ένα σύνολο συνθετικών δεδομένων. Η παραγωγή τους βασίστηκε στα φυσιολογικά εύρη αναφοράς, όπως ορίζονται από τη διεθνή βιβλιογραφία και τον Παγκόσμιο Οργανισμό Υγείας (ΠΟΥ). Προκειμένου να διασφαλιστεί ο ρεαλισμός και η φυσική διακύμανση των δεδομένων, υπολογίστηκε η τυπική απόκλιση, λαμβάνοντας υπόψη παγκόσμιες πληθυσμιακές μελέτες του ΠΟΥ, καθώς και, όπου ήταν διαθέσιμες, σχετικές πληροφορίες για την Ελλάδα. Για παράδειγμα, κατά τη δημιουργία του

συνόλου δεδομένων για την HDL, η κατανομή των τιμών καθορίστηκε με βάση τις επίσημες αναφορές του ΠΟΥ για την Ελλάδα (περίοδος 1980-2018) και τις αντίστοιχες παγκόσμιες αναφορές.

Στη συνέχεια, σχεδιάστηκε η αρχιτεκτονική του συστήματος πρόβλεψης ρίσκου. Στόχος του συστήματος είναι να προβλέψει το ρίσκο του ατόμου να είναι εκτός φυσιολογικών ορίων στην επόμενη βιοχημική εξέταση. Η προτεινόμενη προσέγγιση βασίζεται σε καθιερωμένα μοντέλα κινδύνου και πρακτικές, όπως περιγράφονται στη σχετική διεθνή βιβλιογραφία. Για το σύστημα αυτό βασιστήκαμε στον αλγόριθμο Μηχανικής Μάθησης Random Forest (RF) καθώς και στους αλγορίθμους Βαθιάς Μάθησης όπως Recurrent Neural Networks (RNNs) και Long Short-Term Memory (LSTMs).

Το τελικό μοντέλο πρόβλεψης ρίσκου βασίστηκε στη μέθοδο συνδυαστικών μοντέλων (ensemble models) και πιο συγκεκριμένα σε ένα Weighted Ensemble Model. Συνδυάζουμε τις προβλέψεις και των τριών μοντέλων (RF, RNN, LSTM), χρησιμοποιώντας βάρη (weights) που καθορίστηκαν εμπειρικά, βασισμένα στην απόδοση του κάθε μοντέλου μεμονωμένα σε προηγούμενο στάδιο (Theocharopoulos et al., 2025). Ο στόχος είναι να πάρουμε τα δυνατά στοιχεία από κάθε αρχιτεκτονική.

Για την αξιολόγηση του συστήματος χρησιμοποιούμε δεδομένα που το μοντέλο δεν είχε δει κατά τη διάρκεια της εκπαίδευσης. Για τη δυνατότητα χρήσης του συστήματος έχουν εισαχθεί 15 εγγραφές στη βάση δεδομένων ώστε να μπορεί να τρέξει το προτεινόμενο σύστημα.

Το μοντέλο αυτό και τα δεδομένα ενσωματώθηκαν στον ιατρικό φάκελο και είναι εκεί διαθέσιμα στην Ενότητα Εργασίας 12.

## 5. Συνθετικά γενετικά δεδομένα ασθενών

Για τη διασφάλιση της προστασίας των προσωπικών δεδομένων, πραγματοποιήθηκε προσομοίωση βιοϊατρικών δεδομένων αναφορικά με τον γονότυπο “πρότυπων” ασθενών που θα ενταχθούν στον ιατρικό φάκελο. Η διαδικασία ξεκίνησε με την επιλογή αρχικού αρχείου από την πλατφόρμα Michigan Imputation Server 2 (<https://imputationserver.sph.umich.edu/#!>), το οποίο περιλάμβανε ατομικά δεδομένα 50 ατόμων. Στο πλαίσιο της ίδιας πλατφόρμας, διενεργήθηκε ποιοτικός έλεγχος, imputation πολυμορφισμών μέσω του εργαλείου Minimac4 και phasing με τη χρήση του λογισμικού Eagle, ώστε να εξαχθούν οι γονότυποι με βάση τον υπό μελέτη πληθυσμό.

Στη συνέχεια, επιλέχθηκε τυχαία το δείγμα με κωδική ονομασία “HG00096” ως αντιπροσωπευτικό αρχείο ατομικών δεδομένων. Ακολούθησε μετατροπή των συντεταγμένων του γονιδιώματος από την έκδοση hg19 στην έκδοση hg38 με τη χρήση του

εργαλείου CrossMap (<https://github.com/liguowang/CrossMap>). Το τελικό αρχείο χρησιμοποιήθηκε ως αναφορά (reference) για τη δημιουργία 18 επιπλέον αρχείων, στα οποία τροποποιήθηκαν στοχευμένα συγκεκριμένοι πολυμορφισμοί που σχετίζονται με επιλεγμένες ασθένειες και φαρμακογονίδια.

Ειδικότερα, για κάθε ένα από τα παραχθέντα αρχεία επιλέχθηκε μια συγκεκριμένη ασθένεια και εντοπίστηκαν οι σχετικοί γενετικοί συσχετισμοί μέσω του GWAS Catalog (<https://www.ebi.ac.uk/gwas/>). Η τροποποίηση των δεδομένων περιλάμβανε τη μετατροπή των γονοτύπων σε ομοζυγωτία ως προς το εναλλακτικό αλληλόμορφο. Με την ίδια μεθοδολογία πραγματοποιήθηκαν τροποποιήσεις και σε πολυμορφισμούς που αφορούν φαρμακογονίδια, προκειμένου να εξυπηρετηθούν οι ανάγκες των μεταγενέστερων αναλύσεων. Τα τελικά αρχεία αυτών των ασθενών χρησιμοποιήθηκαν με τις μεθόδους που αναπτύσσονται στο Π7.2 και προέκυψαν έτσι δεδομένα για ασθένειες μονογονιδιακού χαρακτήρα, PRS και αποτελέσματα φαρμακογονιδιωματικής ανάλυσης που ενσωματώθηκαν στον ιατρικό φάκελο.

Τα ευρήματα των αναλύσεων PRS όπως περιγράφονται στην Π7.2 οι οποίες εφαρμόστηκαν στα αρχεία των ασθενών, η κατασκευή των οποίων προηγουμένως ανεφέρθηκε, θα ενσωματωθούν στον ατομικό ιατρικό φάκελο όπου για κάθε ασθενή θα καταχωρούνται τα ακόλουθα πεδία, όπως φαίνονται στον Πίνακα 3.

**Πίνακας 3:** Πεδία αναλύσεων PRS στον ιατρικό φάκελο.

PRS ID	Το μοναδικό αναγνωριστικό του polygenic score για κάθε trait, όπως είναι καταχωρημένο στο PGS Catalog
PRS ID link	Σύνδεσμος προς την αντίστοιχη εγγραφή στο PGS Catalog
PRS Publication	Αναγνωριστικό της επιστημονικής δημοσίευσης από την οποία προέρχεται το PRS
PRS Publication link	Σύνδεσμος προς την επιστημονική δημοσίευση στο PGS Catalog
PRS Trait	Ο φαινότυπος, νόσημα ή χαρακτηριστικό για το οποίο υπολογίζεται το PRS
PRS Number of Variants	Ο συνολικός αριθμός SNPs που χρησιμοποιεί το συγκεκριμένο PRS
PRS Genome Build	Γονιδίωμα αναφοράς που χρησιμοποιήθηκε για την ανάλυση

PRS Match %	Το ποσοστό των variants του PRS που βρέθηκαν στο αρχείο του ασθενούς
PRS Total Variants Matched	Αριθμός variants που αντιστοιχίστηκαν επιτυχώς στο δείγμα
PRS Total Variants Unmatched	Αριθμός variants του PRS που δεν βρέθηκαν στο αρχείο του ασθενούς
PRS FID	Family ID του αρχείου
PRS IID	Individual ID του αρχείου
PRS Patient score	Ακατέργαστο polygenic risk score του ασθενούς
PRS Mean	Μέση τιμή PRS στον πληθυσμό αναφοράς
PRS Variance	Διακύμανση PRS στον πληθυσμό αναφοράς
PRS SD	Τυπική απόκλιση PRS
PRS Patient z-score	Z-score του ασθενούς (απόκλιση από τον μέσο πληθυσμό σε μονάδες SD)
PRS Percentile of patient	Ποσοστημόριο του ασθενούς στον πληθυσμό αναφοράς (σχετική θέση κινδύνου)

## 6. Ανάπτυξη του διαδικτυακού εργαλείου Flame (v2.0) για τη λειτουργική ανάλυση και ενοποίηση βιολογικών δεδομένων»

Στο πλαίσιο της ενότητας εργασίας, αναπτύχθηκε το Flame (v2.0), ένα προηγμένο και διαδραστικό διαδικτυακό εργαλείο ανάλυσης, το οποίο ενσωματώνει δεδομένα και μεθόδους από πολλαπλές πλατφόρμες λειτουργικής εμπλουτισμένης ανάλυσης (g:Profiler, aGTool, Enrichr και WebGestalt), επιτρέποντας συνδυαστική διερεύνηση και οπτικοποίηση των αποτελεσμάτων. Το Flame υποστηρίζει: Επεξεργασία πολλαπλών λιστών γονιδίων/πρωτεϊνών και απεικόνιση συνόλων ένωσης ή τομής μέσω διαδραστικών UpSet plots. Αυτόματη εξαγωγή βιολογικών οντοτήτων από ελεύθερο κείμενο με τεχνικές text mining και Named Entity Recognition (NER). Ανάλυση μονοκλωνικών πολυμορφισμών (SNPs) και εξαγωγή των σχετικών γονιδίων. Διαδραστική επιλογή και ανάλυση διαφορικά εκφρασμένων γονιδίων μέσω volcano plots. Η τρέχουσα έκδοση υποστηρίζει πλέον 14.436 οργανισμούς, σημαντική αύξηση σε σχέση με τους 197 της αρχικής έκδοσης. Στο επόμενο στάδιο, σχεδιάζουμε την ενσωμάτωση νέων δεδομένων και λειτουργιών για την υποστήριξη

αιτιολογικής διερεύνησης βάσει μεγάλης κλίμακας πληθυσμιακών και γονιδιωματικών δεδομένων. Επιπλέον, βρίσκεται σε εξέλιξη μια εκτενής εμπειρική συστηματική ανασκόπηση της διεθνούς βιβλιογραφίας, με σκοπό την καταγραφή και κατηγοριοποίηση όλων των αξιόπιστων συσχετίσεων μεταξύ φαινοτύπων που έχουν προκύψει μέσω Μεντελικής Τυχαιοποίησης (Mendelian Randomization). Τα δεδομένα αυτά σε συνεργασία με την ομάδα του Πανεπιστημίου Θεσσαλίας επεξεργάζονται και οργανώνονται σε μια νέα βάση δεδομένων, η οποία θα ενσωματωθεί στο Flame, επιτρέποντας στοχευμένη ανάλυση και διαλειτουργικότητα με τα αποτελέσματα λειτουργικού εμπλουτισμού. Η ενσωμάτωση αυτών των δεδομένων και μεθόδων στο Flame v2.0 ανοίγει τον δρόμο για ένα ευρύτερα χρήσιμο εργαλείο συστημικής βιοϊατρικής, που υποστηρίζει τη λειτουργική ερμηνεία, τη βιολογική τεκμηρίωση και τη χαρτογράφηση αιτιολογικών μηχανισμών μεταξύ φαινοτύπων και ασθενειών. Η παραπάνω εργασία περιγράφεται αναλυτικά στη δημοσίευση: Evangelos Karatzas, Fotis A Baltoumas, Eleni Aplakidou, Panagiota I Kontou, Panos Stathopoulos, Leonidas Stefanis, Pantelis G Bagos, Georgios A Pavlopoulos, Flame (v2.0): advanced integration and interpretation of functional enrichment results from multiple sources, *Bioinformatics*, Volume 39, Issue 8, August 2023, btad490, <https://doi.org/10.1093/bioinformatics/btad490>.

## 7. Ανάπτυξη της βάσης δεδομένων neomerDB για την ταυτοποίηση και αξιοποίηση καρκινικών βιοδεικτών

Η ανάπτυξη βιοδεικτών για πληθυσμιακό έλεγχο, έγκαιρη ανίχνευση καρκίνου, παρακολούθηση και επιτήρηση υποτροπών προσφέρει σημαντικές δυνατότητες βελτίωσης των εκβάσεων των ασθενών και σωτηρίας ζωών. Τα nullomers είναι μικρά k-μερή (k-mers) που απουσιάζουν από το ανθρώπινο γονιδίωμα και τα neomers αποτελούν το υποσύνολο των nullomers που εμφανίζονται επανειλημμένα λόγω σωματικών μεταλλάξεων κατά τη διάρκεια της ανάπτυξης του καρκίνου. Εδώ, αναπτύξαμε τη neomerDB, μια βάση δεδομένων που περιλαμβάνει έναν κατάλογο των neomers σε διάφορους τύπους καρκίνου και όργανα. Εξετάσαμε 10.000 δείγματα αλληλούχισης ολόκληρου εξώματος (whole exome sequencing) και 2.658 δείγματα αλληλούχισης ολόκληρου γονιδιώματος (whole genome sequencing) που αντιστοιχούσαν σε όγκους, και εντοπίσαμε το σύνολο των neomers που σχετίζονται με κάθε τύπο καρκίνου και όργανο. Αναλύσαμε επίσης 76.215 ολόκληρα γονιδιώματα και 730.947 ολόκληρα εξώματα ατόμων από διαφορετικές καταγωγές, από τα οποία αφαιρέσαμε τα nullomers και τα neomers που μπορεί να προκύψουν λόγω βλαστικών παραλλαγών (germline variants) στον πληθυσμό. Τέλος, διεξαγάγαμε μια μελέτη περίπτωσης που αποδεικνύει ότι τα neomers μπορούν να χρησιμοποιηθούν για την ανίχνευση του γλοιοβλαστώματος από δείγματα υγρής βιοψίας (n = 38), χρησιμοποιώντας ελεύθερο κυττάρων DNA (cell-free DNA) και ελεύθερο κυττάρων RNA (cell-free RNA), επιτυγχάνοντας βαθμολογία 0,98 στην Καμπύλη Λειτουργικού Χαρακτηριστικού Δέκτη - Εμβαδόν Κάτω από την Καμπύλη (ROC-AUC) και βαθμολογία ακρίβειας-ανάκλησης (precision-recall) 0,99. Η neomerDB είναι μια εύχρηστη βάση δεδομένων που επιτρέπει προηγμένες αναζητήσεις, παρέχει διαδραστικές απεικονίσεις και επιλογές λήψης (download) για τους βιοδείκτες

neomer. Η neomerDB είναι διαθέσιμη στο κοινό στη διεύθυνση <https://neomerDB.com/>.» Η παραπάνω εργασία περιγράφεται αναλυτικά στη δημοσίευση: Kimonas Provatias, Candace S Y Chan, Ioannis Kerasiotis, Eleftherios Bochalis, Akshatha Nayak, Brad E Zacharia, Georgios A Pavlopoulos, Wei Li, Ilias Georgakopoulos-Soares, neomerDB: a comprehensive database of neomer biomarkers in cancer, Database, Volume 2026, 2026, baag006, <https://doi.org/10.1093/database/baag006>.

## 8. Βιβλιογραφία

Bowden, J., Davey Smith, G., & Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International journal of epidemiology*, 44(2), 512-525.

Burgess, S., & Thompson, S. G. (2015). *Mendelian randomization: methods for using genetic variants in causal estimation*. CRC Press.

Canese, K., & Weis, S. (2013). PubMed: the bibliographic database. *The NCBI handbook*, 2(1), 2013.

Chatterjee, N., Shi, J., & García-Closas, M. (2016). Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics*, 17(7), 392-406.

Davies, N. M., Holmes, M. V., & Smith, G. D. (2018). Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *bmj*, 362.

Hedges, L. V. (1992). Meta-analysis. *Journal of Educational Statistics*, 17(4), 279-296.

Kontou, P. I., Sasilioglou, I. V., & Bagos, P. G. (2026). A Partial Correlation Network from Summary Data Can Identify Causally Related Diseases. In I. Rojas et al. (Eds.), *Bioinformatics and Biomedical Engineering* (pp. 1–11). Springer Nature Switzerland. [https://doi.org/10.1007/978-3-032-08455-2\\_13](https://doi.org/10.1007/978-3-032-08455-2_13)

Sanderson, E., Davey Smith, G., Windmeijer, F., & Bowden, J. (2019). An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings. *International journal of epidemiology*, 48(3), 713-727.

Spain, S. L., & Barrett, J. C. (2015). Strategies for fine-mapping complex traits. *Human molecular genetics*, 24(R1), R111-R119.

Theocharopoulos, P. C., Bersimis, S., Georgakopoulos, S. V., Plagianakos, V. P., & Tasoulis, S. K. (2025, March). Transforming Medical Practice: Harnessing the Power of Big Data and

Machine Learning for Predictive Precision Medicine. In 2025 IEEE Symposium on Computational Intelligence in Health and Medicine (CIHM) (pp. 1-7). IEEE.

Verbanck, M., Chen, C.-Y., Neale, B., & Do, R. (2018). Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature genetics*, 50(5), 693-698.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., & Hindorff, L. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research*, 42(D1), D1001-D1006.

Werme, J., van der Sluis, S., Posthuma, D., & de Leeuw, C. A. (2022). An integrated framework for local genetic correlation analysis. *Nature genetics*, 54(3), 274-282.