ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ

ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

Η ΧΡΗΣΗ ΤΩΝ ΜΕΘΟΔΩΝ ΤΗΣ ΑΝΑΛΥΤΙΚΗΣ ΤΩΝ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΤΗΣ ΣΤΑΤΙΣΤΙΚΗΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΣΤΗΝ ΠΡΟΑΓΩΓΗ ΤΗΣ ΥΓΕΙΑΣ

Δημήτριος Π. Ζαρογιάννης

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς

Μάρτιος 2023

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ

ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

Η ΧΡΗΣΗ ΤΩΝ ΜΕΘΟΔΩΝ ΤΗΣ ΑΝΑΛΥΤΙΚΗΣ ΤΩΝ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΤΗΣ ΣΤΑΤΙΣΤΙΚΗΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΣΤΗΝ ΠΡΟΑΓΩΓΗ ΤΗΣ ΥΓΕΙΑΣ

Δημήτριος Π. Ζαρογιάννης

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς

Μάρτιος 2023

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. ……. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Αναπληρωτής Καθηγητής Μπερσίμης Σωτήριος (Επιβλέπων)

- Καθηγητής Πλαγιανάκος Βασίλειος

- Επίκουρος Καθηγητής Τασουλής Σωτήριος

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN

APPLIED STATISTICS

THE USE OF DATA ANALYTICS AND MACHINE LEARNING METHODS FOR PROMOTING HEALTH

By

Dimitris P. Zarogiannis

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece

March 2023

Στους γονείς μου

Παύλο και Φωτεινή

Ευχαριστίες

Φτάνοντας στο τέλος αυτής της εργασίας με την οποία ολοκληρώνεται ένας μακρύς ομολογουμένως κύκλος σπουδών στο ΠΑΠΕΙ, νιώθω την ανάγκη να ευχαριστήσω τους ανθρώπους που συνέβαλαν στην εκπονησή της. Πρώτα απ'όλα θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Μπερσίμη Σωτήριο για την ανάθεση ενός τόσο ενδιαφέροντος θέματος και την υποστήριξή του καθ'όλη την διαρκεια εκπόνησης της εργασίας. Ακολούθως θα ήθελα να ευχαριστήσω τα μέλη της Επιτροπής τον κ. Πλαγιανάκο Βασίλειο και τον κ. Τασουλή Σωτήριο. Επιπλέον, θα ήθελα να ευχαριστήσω τους καθηγητές μου στο Πανεπιστήμιο Ιωαννίνων, όπου αποφοίτησα από το τμήμα Μαθηματικών, τον κ.Ζωγράφο Κωνσταντίνο και τον κ.Μπατσίδη Απόστολο για τις συμβουλές που μου έδωσαν και την υποστηριξη τους πριν έρθω να σπουδάσω σε αυτό το μεταπτυχιακό πρόγραμμα. Τέλος το μεγαλύτερο ευχαριστώ το οφείλω στους ανθρώπους που αποτελούν φάρους στην πορεία της ζωής μου. Τους γονείς μου Παύλο και Φωτεινή και τον αδερφό μου Παναγιώτη, που είναι πάντα δίπλα μου και δεν σταμάτησαν ποτέ να πιστεύουν σε μένα και τους ευχαριστώ για όλες τις θυσίες που έκαναν για να φτάσουμε εγώ και ο αδερφός μου ως εδώ. Εύχομαι μια μέρα να καταφέρω να τους κάνω περήφανους.

Περίληψη

Ζαρογιάννης Δημήτριος

Η ΧΡΗΣΗ ΤΩΝ ΜΕΘΟΔΩΝ ΤΗΣ ΑΝΑΛΥΤΙΚΗΣ ΤΩΝ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΤΗΣ ΣΤΑΤΙΣΤΙΚΗΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΣΤΗΝ ΠΡΟΑΓΩΓΗ ΤΗΣ ΥΓΕΙΑΣ

Μάρτιος 2023

Στην σημερινή εποχή ο όγκος και η ταχύτητα των δεδομένων που αποκτώνται από διάφορες πηγές που σχετίζονται με το περιβάλλον, τον πληθυσμό και την δημόσια υγεία αυξάνονται ραγδαία. Συγκεκριμένα στον χώρο της υγείας, ο όγκος των δεδομένων είναι μεγάλος κάτι που έχει ως αποτέλεσμα να υπάρξει η ανάγκη της χρήσης των μεθόδων της αναλυτικής των δεδομένων και της στατιστικής μηχανικής μάθησης. Ένα σημαντικό ζήτημα στον χώρο της υγείας είναι η πρόληψη των ασθενειών, καθώς πολλά προβλήματα υγείας που εμφανίζονται στους ανθρώπους έχουν σχέση με τον σύγχρονο τρόπο ζωής και με το φυσικό και κοινωνικό περιβάλλον. Συνεπώς, οι αλλαγές στη συμπεριφορά και στη στάση σε τομείς όπως η διατροφή, η άσκηση και η υγιεινή θα μπορούσαν να βοηθήσουν στην καταπολέμηση τους. Έτσι η λήψη αποφάσεων στην δημόσια υγεία είναι κρίσιμη και σε αυτό μπορεί να βοηθήσει η στατιστική μηχανική μάθηση. Σε αυτήν την εργασία θα γίνει μία εκτενής περιγραφή των προβλημάτων που εφάπτονται του τομέα της προαγωγής της υγείας αλλά και το πως συνέβαλε στην λύση αυτών των προβλημάτων η στατιστική μηχανική μάθηση με τη βοήθεια διάφορων εφαρμογών της. Τέλος, θα παρουσιαστεί μία εφαρμογή που οδηγεί στην λήψη αποφάσεων με τη χρήση μεθόδων αναλυτικής των δεδομένων και στατιστικής μηχανικής μάθησης που σχετίζεται με την πανδημία του κορονοϊού (COVID – 19) και συγκεκριμένα το πότε μπορούμε να περιμένουμε ότι μία χώρα θα χορηγήσει αρκετά εμβόλια για να επιτύχει την ανοσία της αγέλης.

Abstract

Zarogiannis Dimitris

THE USE OF DATA ANALYTICS AND MACHINE LEARNING METHODS FOR PROMOTING HEALTH

March 2023

In today's world, the volume and speed of data acquired from various sources related to the environment, population and public health are increasing rapidly. Particularly in the field of health, the volume of data is large which has resulted in the need for the use of data analytics and statistical machine learning methods. An important issue in the field of health is the prevention of diseases, as many health problems that occur in people are related to modern lifestyles and the physical and social environment. Therefore, changes in behaviour and attitudes in areas such as diet, exercise and hygiene could help to combat them. Thus public health decision making is critical and this is where statistical machine learning can help. In this thesis we will give an extensive description of the problems that are tangential to the field of health promotion and how statistical machine learning has helped in solving these problems with the help of various applications of statistical machine learning. Finally, an application leading to decision making using data analytics and statistical machine learning methods will be presented related to the coronavirus pandemic (COVID - 19) and in particular when we can expect a country to provide enough vaccines to achieve herd immunity.

.

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

Κ Ε Φ Α Λ Α Ι Ο 1

Εισαγωγή

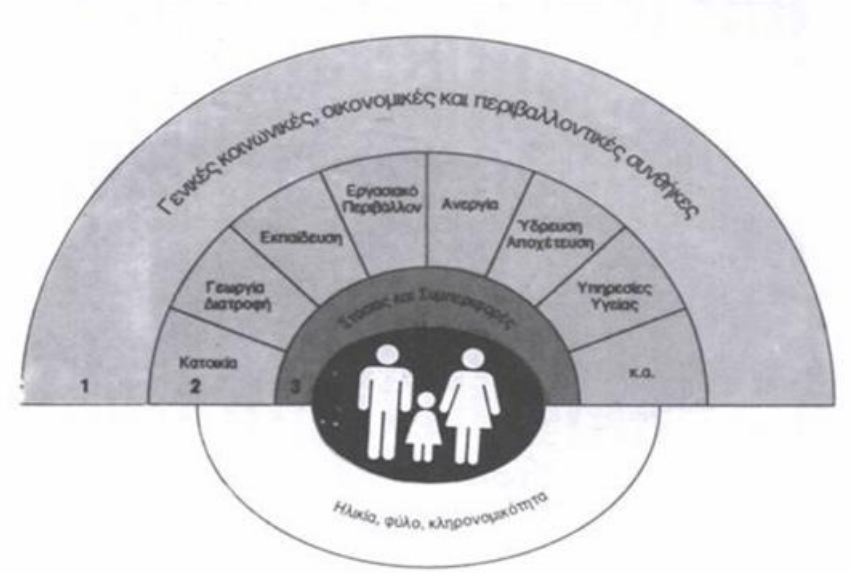
1.1 Προαγωγή της Υγείας

Στην ψηφιακή εποχή, ο όγκος και η ταχύτητα των δεδομένων, τα οποία προκύπτουν από ένα ευρύ φάσμα πηγών, που σχετίζονται με το περιβάλλον, τον πληθυσμό και την δημόσια υγεία αυξάνονται ραγδαία. Οι τεχνικές ανάλυσης των Big Data, όπως η Στατιστική Ανάλυση (Statistical Analysis), η Εξόρυξη Δεδομένων (Data Mining), η Στατιστική Μηχανική Μάθηση (Statistical Machine Learning) και η Βαθιά Μάθηση (Deep Learning), προσέλκυσαν το αυξανόμενο ενδιαφέρον των ερευνητών και των επιστημόνων τις πρόσφατες δεκαετίες. Μάλιστα η λήψη αποφάσεων με βάση συγκεκριμένα στοιχεία είναι κρίσιμη και έχει σημαντικό αντίκτυπο στην δημόσια υγεία, στην ατομική υγεία των ανθρώπων και στην εφαρμογή διαφόρων προγραμμάτων (Chan & Chang, 2020).

Ο Γ.Κ. Τούντας (2005) αναφέρει ότι οι υπηρεσίες της υγείας και η ιατρική φροντίδα υπάρχουν για να καλύπτουν τις ανάγκες υγείας κάθε κοινωνίας. Όμως, οι κοινωνίες αλλάζουν, όπως και οι ανάγκες υγείας και οι προσδοκίες του πληθυσμού για φροντίδα και περίθαλψη. Έτσι αυτές οι αλλαγές, είναι λογικό να κατευθύνουν τα συστήματα υγείας σε νέους προσανατολισμούς, τα οποία δεν παραμένουν στάσιμα αλλά συνεχώς εξελίσσονται, λόγω διαφόρων διεργασιών, όπως είναι η ανάπτυξη της οικονμίας και της τεχολογίας ή από εσωτερικές δυνάμεις όπως είναι οι εμπνεύσεις και η άοκνη προσφορά ορισμένων λειτουργών τους. Μία πολύ σημαντική αλλαγή στο χώρο της υγείας διεθνώς, ήταν το αυξανόμενο ενδιαφέρον για παροχή πρωτοβάθμιας φροντίδας και πρόληψης. Οι άνθρωποι άρχισαν να συνειδητοποιούν τα όρια της θεραπευτικής-νοσοκομειακής ιατρικής ενάντια στα μείζονα προβλήματα που υπάρχουν στην εποχή μας, όπως είναι ο καρκίνος, οι καρδιαγγειακές παθήσεις, τα ατυχήματα, οι ψυχικές διαταραχές, τα χρόνια εκφυλιστικά νοσήματα, το AIDS κτλ. Ταυτόχρονα με την ύπαρξη αυτών των προβλημάτων, οι κρατικοί προϋπολογισμοί δυσκολεύονται ή αδυνατούν να τα ανταποκριθούν στις μεγάλες δαπάνες για την υγεία, που στην μεγαλύτερη τους πλειοψηφία καταναλώνονται από το θεραπευτικό-νοσοκομειακό τομέα. Τα δύο αυτά φαινόμενα, των μεγάλων δαπανών και της περιορισμένης αποτελεσματικότητας, που συμβάλλουν στην πολυσυζητημένη κρίση της σύγχρονης Ιατρικής, έχουν οδηγήσει σε μία προσπάθεια αναζήτησης νέων αλλά και αναβάθμισης πολιτικών υγείας, με σκοπό να ανταποκριθούν αποτελεσματικά αλλά και οικονομικά στα σύγχρονα προβλήματα υγείας. Συγκεκριμένα έχουν οδηγήσει στην ανάπτυξη της Πρόληψης, καθώς όλα σχεδόν τα

προβλήματα υγείας της εποχής μας σχετίζονται με τον σύγχρονο τρόπο ζωής και με το σύγχρονο φυσικό και κοινωνικό περιβάλλον. Η 34η γενική συνέλευση της Παγκόσμιας Οργάνωσης Υγείας (Π.Ο.Υ) που διεξήχθη το 1981, έθεσε σαν στρατηγική επιδίωξη το "Υγεία για όλους το έτος 2000" (Health For All by the year 2000-HFA 2000). Η ανάπτυξη της Πρωτοβάθμιας Φροντίδας Υγείας (Π.Φ.Υ), ήταν μοχλός αυτής της προσπάθειας, καθώς ο αναπροσανατολισμός των υπηρεσιών υγείας θεωρήθηκε ένας από τους βασικούς στόχους. Οι υπόλοιποι βασικοί στόχοι ήταν η πρόληψη των προλήψιμων νοσημάτων και των προβλημάτων υγείας, η αναβάθμιση του φυσικού και κοινωνικού περιβάλλοντος και η διαμόρφωση υγιεινών στάσεων και συμπεριφορών. Για την πραγματοποίηση αυτών των στόχων, το 1986 ο Π.Ο.Υ θεσμοθέτησε, με την διακύρηξη της Οττάβας, την πολιτική της Προαγωγής Υγείας (Health Promotion), που έχει ως σκοπό την αναβάθμιση του φυσικού και κοινωνικού περιβάλλοντος, την ενίσχυση των ευρύτερων παραγόντων που επιδρούν θετικά στην ανθρώπινη υγεία και την διαμόρφωση υγιεινών στάσεων και συμπεριφορών (Σχήμα 1.1 - Γ.Κ. Τούντας, 2005).

Σχήμα 1.1
Οι παράγοντες που επιδρούν στην υγεία



Επομένως, η Προαγωγή της Υγείας, η οποία προσανατολίζεται στην κοινότητα και στις ανάγκες του κάθε τοπικού πληθυσμού και είναι διαποτισμένη από μια ουμανιστική αντίληψη για την υγεία και την αρρώστια, κοινωνικά προσανατολισμένη, έρχεται να καλύψει ένα κενό που γινόταν όλο και μεγαλύτερο τα τελευταία χρόνια με την αύξηση του ειδικού βάρους της σύγχρονης νοσοκομειακής-θεραπευτικής ιατρικής.Η Προαγωγή Υγείας σύμφωνα με τη διακήρυξη της Π.Ο.Υ είναι εκείνη η διαδικασία μέσα στην οποία τα άτομα έχουν την δυνατότητα να γίνουν ικανά να αναπτύξουν τον έλεγχο πάνω στην υγεία τους και να τη βελτιώσουν. Η άποψη αυτή πηγάζει από μια ευρύτερη αντίληψη που ταυτίζει υγεία με τη δυνατότητα που έχει ένα άτομο ή μια ομάδα, να μπορεί από την μία μεριά να πραγματοποιεί τις επιθυμίες του και να ικανοποιεί τις ανάγκες του και από την άλλη μεριά να μεταβάλλει το περιβάλλον ή να προσαρμόζεται σε αυτό. Επίσης, η υγεία θα πρέπει να θεωρείται ως ένας

συντελεστής της καθημερινής ζωής και όχι ως αυτοσκοπός. Δηλαδή είναι μία θετική έννοια η οποία στοχεύει στις κοινωνικές και προσωπικές δυνατότητες και ικανότητες.

Ακόμη, σύμφωνα με την ίδια διακύρηξη, η Προαγωγή Υγείας χαρακτηρίζεται από τις ακόλουθες αρχές, δραστηριότητες, προτεραιότητες και διλήμματα (Γ.Κ. Τούντας, 2005).

1.1.1 Αρχές

Η Προαγωγή Υγείας επικεντρώνεται περισσότερο στο σύνολο του πληθυσμού μέσα στα πλαίσια της καθημερινής του ζωής, παρά στα μεμονωμένα άτομα που διατρέχουν κάποιο συγκεκριμένο κίνδυνο να νοσήσουν. Καθιστά τα άτομα ικανά να αναλαμβάνουν αυθόρμητα ή οργανωμένα υπεύθυνη δράση για την υγεία τους,που όπως αναφέρθηκε παραπάνω, αποτελεί σημαντικό συντελεστή της καθημερινής τους ζωής. Για να συμβεί αυτό, απαιτείται πλήρη και συνεχή πρόσβαση στην πληροφόρηση για θέματα υγείας,καθώς και την μέγιστη δυνατή κοινοποιήση όλων των σχετικών πληροφοριών στο σύνολο του πληθυσμού. Ακόμη, η Προαγωγή Υγείας, εκτός από τη στενή συνεργασία των υπηρεσιών υγείας, προϋποθέτει και την συνεργασία όλων των φορέων που είναι σχετικοί με τους παράγοντες που επιδρούν στην υγεία. Ακόμη, η Προαγωγή Υγείας συνδυάζει διάφορες συμπληρωματικές μεθόδους ή τρόπους προσέγγισης, όπως την επικοινωνία, την εκπαίδευση, τη νομοθεσία, τα οικονομικά μέτρα, τις οργανωτικές αλλαγές, την ανάπτυξη της κοινότητας, αλλά και αυτογενείς τοπικές δραστηριότες ενάντια στους κινδύνους της υγείας. Επιπροσθέτως η Προαγωγή της Υγείας έχει ως στόχο την αποτελεσματική και ουσιαστική συμμετοχή του κοινού, κάτι το οποίο απαιτεί την ανάπτυξη επιδεξιοτήτων σε ατομικό και συλλογικό επίπεδο οι οποίες διευκολύνουν την κατανόηση των προβλημάτων και την λήψη σωστών αποφάσεων. Τέλος,η Προαγωγή Υγείας είναι μια ευρύτερη δραστηριότητα στο χώρο της υγείας και της κοινωνίας και όχι μία ιατρική υπηρεσία. Επομένως, όλοι οι επαγγελματίες της υγείας και συγκεκριμένα όσοι ασχολούνται με την Πρωτοβάθμια Φροντίδα Υγείας, έχουν ένα σημαντικό ρόλο στην ανάπτυξη της Προαγωγής της Υγείας (Γ.Κ. Τούντας, 2005).

1.1.2 Δραστηριότητες-Προτεραιότητες

Η Προαγωγή της Υγείας έχει ως σκοπό την αναβάθμιση των παραγόντων που επηρεάζουν την υγεία, όπως είναι οι οικονομικοί, οι περιβαλλοντολογικοί, οι κοινωνικοί κ.α. Βέβαια λόγω της πληθώρας αυτών των παραγόντων, θα μπορούσε να καταγραφεί ένας ατελείωτος κατάλογος δραστηριοτήτων, όπως η διατροφική πολιτική, η στέγαση, το κάπνισμα, επιδεξιότητες προσαρμογής, κοινωνική υποστήριξη κ.α. Όμως, για να γίνει πιο εύκολο το έργο της Προαγωγής Υγείας, η διακήρυξη της Οττάβας αναφέρει ότι:

- Στο επίκεντρο της Προαγωγής Υγείας είναι η εξασφάλιση της πρόσβασης στην υγεία, δηλαδή η μείωση των ανισοτήτων στην υγεία και η αύξηση των ευκαιριών για την βελτίωση της υγείας. Αυτή η επιδίωξη συνεπάγεται αλλαγές στις σχετικές

πολιτικές του κράτους και των υπεύθυνων φορέων, καθώς και τον αναπροσανατολισμό των υπηρεσιών της υγείας.

- Η βελτίωση της υγείας εξαρτάται από τη διαμόρφωση ενός υγιεινού περιβάλλοντος, συγκεκριμένα τις συνθήκες εργασίες και κατοικίας. Επειδή, και οι δύο περιπτώσεις σχετίζονται με ένα δυναμικό περιβάλλον, η Προαγωγή Υγείας αποτελείται από μεθόδους για την εκτίμηση της δυναμικής αυτής μέσα από την αξιολόγηση των τεχνολογικών, πολιτιστικών και οικονομικών τάσεων και προοπτικών.
- Η Προαγωγή Υγείας έχει ως στόχο την ενίσχυση των κοινωνικών δομών και της κοινωνικής υποστήριξης. Ο στόχος έχει ως βάση την αποδοχή του σημαντικού ρόλου που διαδραματίζουν οι κοινωνικές δυνάμεις και οι κοινωνικές σχέσεις στον καθορισμό των αξιών και των συμπεριφορών που σχετίζονται με την υγεία.
- Ο κυρίαρχος τρόπος ζωής σε μια κοινωνία είναι καθοριστικής σημασίας διότι διαμορφώνει τις προσωπικές συμπεριφορές που είτε είναι ωφέλιμες είτε βλαβερές. Η υιοθέτηση συμπεριφορών που συμβάλλουν στην υγεία προϋποθέτει την αναζήτηση των σχετικών πεποιθήσεων και αξιών. Η διαμόρφωση μιας υγιεινούς συμπεριφοράς και η ανάπτυξη των κατάλληλων μηχανισμών προσαρμογής στο περιβάλλον και συναλλαγής με αυτό, αποτελούν θεμελιώδους στόχους της Προαγωγής Υγείας.
- Η Προαγωγή Υγείας βασίζεται στην συλλογική προσπάθεια για την προάσπιση της υγείας. Οι αρχές, με την άσκηση κοινωνικής πολιτικής, ευθύνονται στο να καταστήσουν προσιτές τις ενδεικνυόμενες επιλογές, καθώς και να εξασφαλίσουν τις βασικές προϋποθέσεις για μία υγιή ζωή. Οι υποστηρικτές της Προαγωγής της Υγείας μέσα στα κέντρα αποφάσεων, πρέπει να έχουν συνείδηση της σημασίας της αυθόρμητης δραστηριότητας για την υγεία, όπως είναι τα κοινωνικά κινήματα ή τα φαινόμενα της αυτοβοήθειας και της αυτοφροντίδας, καθώς και να αναγνωρίζουν την ανάγκη ύπαρξης συνεχούς συνεργασίας με το κοινό πάνω σε όλα τα θέματα της Προαγωγής Υγείας.
- Η έννοια της Προαγωγής Υγείας πρέπει να αποσαφηνίζεται σε κάθε στάδιο του σχεδιασμού δίνοντας έμφαση στις κοινωνικές, οικονομικές και οικολογικές διαστάσεις της υγείας. Οι πολιτικές της Προαγωγής Υγείας μπορεί να σχετιστούν με άλλες πολιτικές που αφορούν την εργασία, την στέγαση, τις κοινωνικές υπηρεσίες, την Πρωτοβάθμια Φροντίδα Υγείας κ.α.
- Η πολιτική βούληση για την ανάπτυξη της Προαγωγής Υγείας πρέπει να οδηγεί στη δημιουργία κέντρων αναφοράς σε όλα τα επίπεδα (τοπικό, περιφερειακό και εθνικό). Τα κέντρα αυτά θα λειτουργούν ως οργανωτικοί μηχανισμοί για το διατομεακό και συντονισμένο σχεδιασμό της Προαγωγής Υγείας. Θα παρέχουν επίσης ηγεσία και υπευθυνότητα με σκοπό την εξασφάλιση και την ανάπτυξη των σχετικών δραστηριοτήτων. Για να αναπτυχθούν μακροπρόθεσμα προγράμματα είναι σημαντική η ύπαρξη επαρκών κονδυλίων και ειδικευμένου προσωπικού.
- Κατά την ανάπτυξη των δραστηριοτήτων της Προαγωγής Υγείας θα πρέπει να υπάρχει συνεχής διάλογος και ανταλλαγή απόψεων μεταξύ ατόμων ή ομάδων και

ειδικών επαγγελματιών. Επιπλέον, απαιτείται η καθιέρωση μηχανισμών που θα εξασφαλίσουν ευκαιρίες με σκοπό το κοινό να εκφράζει τις απόψεις του και να καλλιεργείται το δημόσιο ενδιαφέρον για την υγεία.

- Για την επιλογή των τομέων προτεραιότητας, θα πρέπει να έχει προηγηθεί μια αξιολόγηση των δεικτών υγείας, των γνώσεων, των ικανοτήτων και πρακτικών του πληθυσμού σε θέματα υγείας καθώς και της ισχύουσας εθνικής ή τοπικής νομοθεσίας και πολιτικής.
- Ακόμη, θα πρέπει να έχει εκτιμηθεί η προσδοκώμενη επίδραση στην υγεία των σχεδιαζόμενων μέτρων και προγραμμάτων, οι οικονομικοί περιορισμοί και τα οικονομικά οφέλη, η κοινωνική και η πολιτιστική αποδοχή τους και η δυνατότητα ολοκληρωμένης πραγμάτωσής τους.
- Η ερευνητική υποστήριξη είναι καθοριστική για την ανάπτυξη των δραστηριοτήτων και για την αξιολόγησή τους. Είναι αναγκαίο να αναπτυχθούν νέες μεθοδολογίες έρευνας και να επινοηθούν κατάλληλοι τρόποι αξιολόγησης. Τα αποτελέσματα της έρευνας θα πρέπει να γίνονται γνωστά σε όλους και να εκτελούνται συγκρίσεις ενδοκρατικές και διεθνείς (Γ.Κ. Τούντας, 2005).

1.1.3 Διλήμματα

Τα βασικά πολιτικά και ηθικά διλήμματα πάντα θα χρείαζεται να αντιμετωπίζονται από την κοινωνική πολιτική της υγείας. Όσοι ασχολούνται με την Προαγωγή Υγείας θα πρέπει να γνωρίζουν τις πιθανές συγκρούσεις συμφερόντων στο κοινωνικό και στο ατομικό επίπεδο. Αρχικά υπάρχει ο κίνδυνος η υγεία να θεωρηθεί ως ο απόλυτος σκοπός που αγκαλιάζει όλες τις πτυχές της ζωής. Η ιδεολογία αυτή, που συχνά αποκαλείται υγιεινισμός (Healthism) θα μπορούσε να οδηγήσει στην κηδεμόνευση των ατόμων και στον έλεγχο της συμπεριφοράς, γεγονός που αντιστρατεύεται τις βασικές αρχές της Προαγωγής Υγείας. Επίσης ενδέχεται, προγράμματα Προαγωγής Υγείας να προσανατολίζονται κυρίως σε ατομικές λύσεις, αντί να στοχεύουν στην επίλυση των γενικότερων προβλημάτων του πληθυσμού. Εκείνοι που αποφασίζουν, ορισμένες φορές θεωρούν τα άτομα αποκλειστικά υπεύθυνα για την υγεία τους. Μάλιστα πολλές φορές υπάρχει η εσφαλμένη εντύπωση ότι οι άνθρωποι έχουν τη δύναμη να σχεδιάζουν εξ' ολοκλήρου τη ζωή τους με σκοπό να μπορούν να αντιμετωπίζουν τους προλήψιμους κινδύνους. Κατά συνέπεια, όταν αρρωσταίνουν, θεωρούνται υπεύθυνοι και στιγματίζονται ανάλογα (victim blaming). Ορισμένες φορές οι μέθοδοι Προαγωγής της Υγείας, δεν ανταποκρίνονται στις προσδοκίες, πεποιθήσεις, προτιμήσεις ή ικανότητες του κοινού, κάτι το οποίο μπορεί να αυξήσει τις κοινωνικές ανισότητες. Η παροχή πληροφοριών από μόνη της π.χ. είναι ανεπαρκής σαν μέτρο. Η ευαισθητοποίηση των ατόμων γύρω από ένα πρόβλημα υγείας χωρίς να παρέχονται οι δυνατότητες για την αντιμετώπισή του, θα μπορούσε να έχει σαν μοναδικό αποτέλεσμα τη δημιουργία ανησυχίας και αισθήματος ανασφάλειας.

Τέλος, υπάρχει κίνδυνος για την οικειοποίηση της Προαγωγής Υγείας από μια επαγγελματική ομάδα που μπορεί να την μετατρέψει σε ένα εξειδικευμένο κλάδο,

αποκλείοντας έτσι τα άλλα ενδιαφερόμενα μέρη. Το κοινό για να μπορέσει να αυξήσει τον έλεγχό του στα θέματα που αφορούν την υγεία του, θα πρέπει να απαιτήσει και να διεκδικήσει από τους επαγγελματίες και τις αρχές μεγαλύτερη συμμετοχή στη διαχείριση των πόρων και στην χάραξη της πολιτικής υγείας και ειδικότερα της Προαγωγής Υγείας (Γ.Κ. Τούντας, 2005).

1.2 Σκοπός

Σκοπός της παρούσας διπλωματικής εργασίας είναι να μπορέσουμε να αναδείξουμε την σημασία και την συνεισφορά της χρήσης των μεθόδων της Αναλυτικής των Δεδομένων και της Στατιστικής Μηχανικής Μάθησης στην Προαγωγή της Υγείας. Συγκεριμένα να παρουσιάσουμε τις μεθόδους της Στατιστικής Μηχανικής Μάθησης που χρησιμοποιούνται για την ανίχνευση των παραγόντων που επηρεάζουν την Προαγωγή της Υγείας, καθώς και να εξατάσουμε το πως μπορούν αυτοί οι μέθοδοι να αναβαθμίσουν τους τρόπους με τους οποίους θα διαμορφωθεί ένα υγιεινό περιβάλλον ώστε να υπάρχει βελτίωση της υγείας των ανθρώπων. Τέλος, με την μελέτη την οποία διενεργήσαμε στο τελευταίο κεφάλαιο, ασχοληθήκαμε με ένα θέμα που αφορά την Δημόσια Υγεία και συγκεκριμένα την πανδημία του κορονοϊού (COVID-19) που προκάλεσε πολλές αναταραχές παγκοσμίως. Στόχος αυτής της μελέτης ήταν με την βοήθεια στατιστικών μοντέλων να μπορέσουμε να προβλέψουμε το πότε σε μία χώρα με την χορήγηση των εμβολίων στους πολίτες της θα επιτευχθεί η ανοσία της αγέλης.

Κ Ε Φ Α Λ Α Ι Ο 2

Ο ρόλος των Big Data στην Δημόσια Υγεία

2.1 Big Data και Στατιστική Μηχανική Μάθηση

Η συνεχής πρόοδος της τεχνολογίας οδήγησε σε μία έκρηξη των μεγάλου όγκου, υψηλής ταχύτητας και μεγάλης ποικιλίας δεδομένων στον χώρο της υγείας. Αυτά είναι τα χαρακτηριστικά των δεδομένων που ορίζουν τα «Big Data». Έτσι, πολλές φορές τα δεδομένα αυτά χαρακτηρίζονται με την φράση "3 V", όπου αυτή η φράση προκύπτει από το αρχικό γράμμα των λέξεων volume (όγκος), velocity (ταχύτητα) και variety (ποικιλία). Επίσης, πολλοί συγγραφείς έχουν επεκτείνει τον ορισμό των Big Data, έχοντας συμπεριλάβει πολλά περισσότερα "V". Για παράδειγμα, σε μία βιβλιογραφική ανασκόπηση των μεθόδων των Big Data, έχουν προστεθεί και άλλα χαρακτηριστικά, όπως variability (μεταβλητότητα), veracity (εγκυρότητα), value (αξία) και visualization (απεικόνιση) (Velmovitsky *et al.*, 2021).

Ακόμη, τα Big Data αναφέρονται σε σύνολα δεδομένων που συνήθως περιλαμβάνουν πολλές παρατηρήσεις και πολλές μεταβλητές, με αποτέλεσμα η χρήση των παραδοσιακών στατιστικών μεθόδων να γίνεται δύσκολη. Έτσι, υπάρχει η ανάγκη για πιο ευέλικτη μοντελοποίηση από αυτή που παρέχεται στην κλασική Στατιστική Ανάλυση. Επίσης, τα σύνολα αυτών των δεδομένων πολλές φορές είναι λιγότερο δομημένα σε σχέση με τα παραδοσιακά συλλεχθέντα δεδομένα και μάλιστα μπορεί να είναι υποσύνολα από άλλα δεδομένα, παρά ένα σκόπιμα συλλεγμένο δείγμα. Τα Big Data αυξήθηκαν παράλληλα με την εκθετική βελτίωση και την επέκταση της χωρητικότητας αποθήκευσης δεδομένων των υπολογιστικών συσκευών (Morgenstern *et al.*, 2021).

Έτσι, για να μπορέσουν οι ερευνητές να αξιοποιήσουν και να κατανοήσουν την πληθώρα των πλούσιων και ποικίλων δεδομένων που παράγονται σήμερα στον χώρο της υγείας, είναι αναγκαίο να χρησιμοποιηθούν οι τεχνολογίες και οι τεχνικές των Big Data. Για να πραγματοποιηθεί η ανάλυση των δεδομένων, η χρήση ισχυρών υπολογιστικών και αποθηκευτικών τεχνολογιών, σε συνδυασμό με τους αλγορίθμους για την κατανόηση και την απόκτηση γνώσης σχετικά με τα Big Data, είναι απαραίτητη. Έτσι, η τεράστια πρόοδος στην υπολογιστική ισχύ, τις τεχνολογίες συλλογής δεδομένων και αποθήκευσης, παρέχει τους απαραίτητους πόρους στους ερευνητές για να μελετήσουν την αλληλεπίδραση σε μεγάλα σύνολα δεδομένων, συμπεριλαμβανομένου των genomics (γονιδιωματική), του περιβάλλοντος,των κοινωνικών δικτύων, καθώς και άλλων ειδών υγείας και προσωπικών δεδομένων. Για να καταστούν, λοιπόν, δυνατές αυτές οι αναλύσεις, εφαρμόζονται μέθοδοι της Στατιστικής Μηχανικής Μάθησης (Velmovitsky *et al.*, 2021).

Η Μηχανική Μάθηση αποτελεί έναν γενικό όρο που χρησιμοποιείται για να περιγράψει μια ευρεία ποικιλία μοντέλων και στρατηγικών που εστιάζουν στην αλγοριθμική μοντελοποίηση. Η ύπαρξη της έννοιας της Μηχανικής Μάθησης χρονολογείται περίπου από τις αρχές της δεκαετίας του 1950 και ο σκοπός της ήταν να αντιμετωπίσει την δυνατότητα οι υπολογιστές να προσεγγίζουν την ανθρώπινη διαδικασία σκέψης μέσω της αντιστοίχισης προτύπων, της αναγνώρισης και της λήψης αποφάσεων. Η έρευνα για την Μηχανική Μάθηση συνεχίστηκε από τον Arthur Samuel και τον Frank Rosenblatt. Ο Arthur Samuel έγραψε ένα πρόγραμμα για να μάθει να παίζει το επιτραπέζιο παιχνίδι ντάμα, ενώ ο Frank Rosenblatt σχεδίασε το πρώτο τεχνητό νευρωνικό δίκτυο, το οποίο χρησιμοποίησε τις αρχές της νευροβιολογίας για την εκτέλεση υπολογισμών. Έκτοτε, έχει αναπτυχθεί ένας μεγάλος αριθμός αλγορίθμων της Μηχανικής Μάθησης για της επίλυση πολλών προβλημάτων. Αυτοί οι αλγόριθμοι κατηγοριοποιούνται γενικά σε μοντέλα με επίβλεψη (Supervised Models) και σε μοντέλα χωρίς επίβλεψη (Unsupervised Models). Τα μοντέλα με επίβλεψη χρησιμοποιούνται συνήθως για την πρόβλεψη ενός αποτελέσματος, παρόμοια με την μοντελοποίηση με χρήση παλινδρόμησης. Τα μοντέλα χωρίς επίβλεψη χρησιμοποιούνται συνήθως για την ανίχνευση άγνωστων μοτίβων που προκύπτουν από τα δεδομένα, χωρίς να χρειάζεται να προβλέψουμε κάποιο αποτέλεσμα. (Wiemken & Kelley, 2019).

2.2 Big Data και Δημόσια Υγεία

Μία από τις βασικές αρχές της Προαγωγής της Υγείας είναι πως έχει ως στόχο το σύνολο του πληθυσμού μέσα στα πλαίσια της καθημερινής του ζωής, παρά τα μεμονωμένα άτομα που διατρέχουν κάποιο συγκεκριμένο κίνδυνο να νοσήσουν. Έτσι, αφού αναφερόμαστε στο σύνολο του πληθυσμού τότε είναι φανερό ότι μιλάμε για την δημόσια υγεία. Ο διευθυντής του γραφείου της Γονιδιωματικής Δημόσιας Υγείας στα Κέντρα Ελέγχου και Πρόληψης Νοσημάτων (Centers for Diseases Control and Prevention - CDC) όρισε την «ακρίβεια» στο πλαίσιο της Δημόσιας Υγείας ως «τη βελτίωση της ικανότητας πρόληψης των ασθενειών, την προαγωγή της υγείας και τη μείωση των ανισοτήτων στην υγεία των πληθυσμών με: 1) την εφαρμογή αναδυόμενων μεθόδων και τεχνολογιών για τη μέτρηση νόσων, παθογόνων παραγόντων, εκθέσεων, συμπεριφορών και ευαισθησίας σε πληθυσμούς και 2) την ανάπτυξη πολιτικών και στοχευμένων προγραμμάτων εφαρμογής για τη βελτίωση της υγείας» (Khoury, 2015). Μάλιστα, οι πρώτες προτεραιότητες περιελάμβαναν τα εξής: την έγκαιρη ανίχνευση κρουσμάτων, τον εκσυγχρονισμό επιτήρησης και τις στοχευμένες παρεμβάσεις στον τομέα της υγείας.

Για να γίνει εφικτό να επιτευχθούν αυτές οι βελτιώσεις, είναι απαραίτητη η παρακολούθηση και η ανάλυση των δεδομένων σε πραγματικό χρόνο. Ακόμη, η επιστήμη της Επιδημιολογίας πρέπει να επεκτείνει την παρακολούθηση της σε πολλαπλούς και διαφορετικούς τομείς, όπως το Διαδίκτυο και τα μέσα κοινωνικής δικτύωσης, π.χ. infodemiology (Prosperi *et al.,* 2018). Συγκεκριμένα, οι πληροφορίες για την επιδημιολογία ή infodemiology, προσδιορίζει τους τομείς όπου υπάρχει κενό μετάφρασης της γνώσης μεταξύ

των βέλτιστων αποδείξεων (τι γνωρίζουν ορισμένοι ειδικοί) και της πρακτικής (τι κάνουν ή τι πιστεύουν οι περισσότεροι άνθρωποι), καθώς και τους δείκτες για την "υψηλής ποιότητας" πληροφορία (Eysenbach, 2002).

Οι Prosperi *et al.* (2018) αναφέρουν πως τα Big Data δεν σημαίνει μόνο ότι πρόκειται για ένα μεγάλο μέγεθος δείγματος ή μία λεπτομερής δειγματοληψία, αλλά και για την ύπαρξη μίας μεγάλης ποικιλίας μεταβλητών. Έως τώρα, η έμφαση των Big Data δίνεται στην αλληλουχία των γονιδιωμάτων σε επίπεδο πληθυσμού, αλλά πλέον υπάρχουν έρευνες που εξετάζουν και άλλους τομείς, όπως η ενσωμάτωση της κλασικής επιτήρησης με την γεωχωρική μοντελοποίηση. Επιπλέον, σε μία έρευνα της Δημόσιας Υγείας, που έχει ως επίκεντρο τις αναπτυσσόμενες χώρες, καλύτερα δεδομένα επιτήρησης, καλύτερες αναλύσεις δεδομένων και ταχείες δράσεις, φαίνεται και πάλι ότι τα Big Data είναι το κλειδί, με έμφαση στην κοινή χρήση των δημόσιων δεδομένων και στα χαρακτηριστικά των δεδομένων, δηλαδή ταχύτητα, ακρίβεια και ισότητα. Ο Ουίνστον Τσόρτσιλ κάποτε είχε δηλώσει ότι «οι υγιείς πολίτες είναι το μεγαλύτερο περουσιακό στοιχείο που μπορεί να έχει μία χώρα» και για να μπορέσει να επιτευχθεί η υγεία για όλους τους πολίτες, θα πρέπει να υπάρξει μία μετάβαση από την ιατρική ακρίβειας (precision medicine), η οποία είναι εξατομικευμένη, στην δημόσια υγεία.

Στην πραγματικότητα, η ιατρική ακρίβειας μπορεί να χρησιμοποιηθεί για να βελτιωθεί η υγεία ενός ατόμου, όμως αυτό δεν σημαίνει απαραίτητα ότι θα υπάρχει ομοιόμορφο όφελος για τον πληθυσμό. Για παράδειγμα, ένα μοντέλο της ιατρικής ακρίβειας το οποίο είναι συντονισμένο για την πλειονότητα ενός πληθυσμού, μπορεί να βελτιώσει τον μέσο όρο των αποτελεσμάτων υγείας συνολικά, αλλά θα παραμελήσει τις μειονότητες. Είναι αξιοσημείωτο να αναφέρουμε ότι όρος ακρίβεια που τοποθετείται δίπλα στις πληθυσμιακές προτεραιότητες ορισμένες φορές φαίνεται να είναι συγκρουσιακός και αυτό μπορεί να οφείλεται στην εφαρμογή ενός μόνο μοντέλου της δημόσιας υγείας σε έναν ολόκληρο πληθυσμό, αντί για την χρήση πολλαπλών τμηματοποιημένων/συγκεντρωτικών μοντέλων. Έτσι, πολλές φορές τίθονται δάφορα ερωτήματα, όπως: Είναι το άτομο; Είναι μία κοινή γεωγραφική περιοχή; Είναι ένας συγκεκριμένος υποπληθυσμός; Η δημόσια υγεία έχει να αντιμετωπίσει κοινωνικές προκλήσεις, συμπεριλαμβανομένων των φυλετικών ανισοτήτων (όσον αφορά την ευημέρια και το γενετικό υπόβαθρο), περιβαλλοντικά θέματα (π.χ. τροπικό κλίμα με υψηλότερα ποσοστά αρμποϊκών ασθενειών, βιομηχανικές περιοχές με υψηλή ρύπανση) και γενικές ηθικές ανησυχίες (π.χ. θρησκευτικές πεποιθήσεις, πολιτικές απόψεις).

Ένα ατομοκεντρικό μοντέλο, όπως το avatar υγείας, θέτει μία σειρά από διάφορους περιορισμούς, διότι μπορεί να μην υπάρχει δυναμική υψηλότερου επιπέδου σε κοινωνικό-περιβαλλοντικό επίπεδο (Σχήμα 2.1 - Prosperi *et al.,* 2018). Ενδιαφέρον είναι ότι τέτοιες δυναμικές, πολλές φορές μπορούν να επηρεάσουν το ίδιο το άτομο, με αποτέλεσμα να είναι αναγκαίο να λαμβάνονται υπόψη και να προβάλλονται στα προσωποκεντρικά μοντέλα.

Σχήμα 2.1
Public health

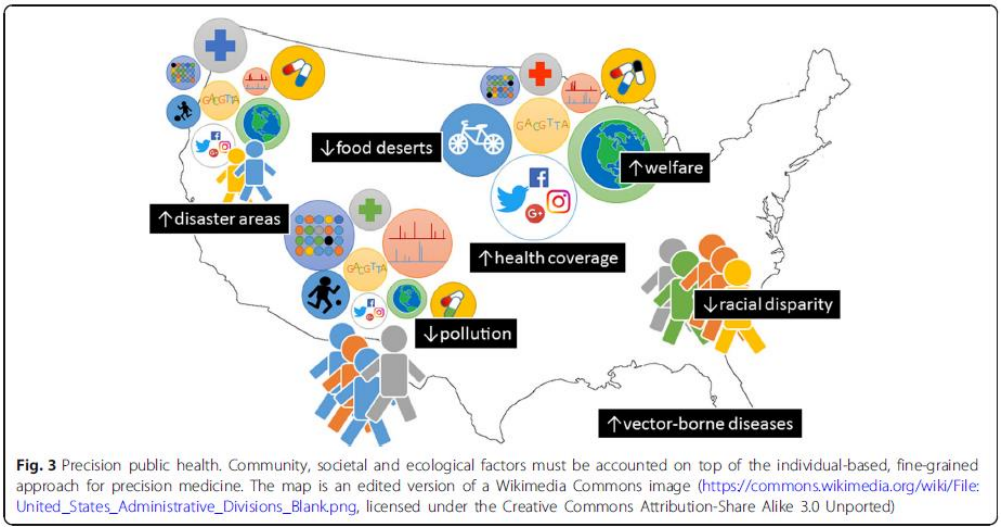


Fig. 3 Precision public health. Community, societal and ecological factors must be accounted on top of the individual-based, fine-grained approach for precision medicine. The map is an edited version of a Wikimedia Commons image (https://commons.wikimedia.org/wiki/File:United_States_Administrative_Divisions_Blank.png, licensed under the Creative Commons Attribution-Share Alike 3.0 Unported)

2.2.1 Μοντελοποίηση των Big Data για την δημόσια υγεία

Οι Prosperi *et al.* (2018) αναφέρουν πως η ύπαρξη εμποδίων στη σύνδεση και την αποτελεσματική αξιοποιήση των πληροφοριών που αποκτώνται στον χώρο της υγείας σε διαφορετικούς δικτυακούς τόπους, επιβραδύνουν την έρευνα στην υγειονομική περίθαλψη και την ανάπτυξη της εξατομικευμένης φροντίδας. Διάφορα συστήματα των EHR (Electronic Health Records) ορίζουν ανεξάρτητα τις δικές τους δομικές μορφές δεδομένων, με αποτέλεσμα αυτή η ανεξάρτητη και ετερογενής διαχείρηση να θέτει προκλήσεις στην απεικόνιση και στην κωδικοποίηση των πληροφοριών. Για παράδειγμα, η συγχώνευση δεδομένων από πολλαπλά συστήματα των EHR ή από διαφορετικές τυποποιημένες διαδικασίες, χωρίς να υπάρχει πρόσβαση στα αρχικά δεδομένα. Ακόμη, η ενσωμάτωση των δεδομένων σε πολλαπλούς τομείς και πολλαπλές πηγές αποτελεί ένα δύσκολο έργο που οφείλεται σε τουλάχιστον τρείς παράγοντες: 1) την ετερογένεια στη σύνταξη των δεδομένων, όπως οι διαφορετικές μορφές αρχείων και τα πρωτόκολλα πρόσβασης που χρησιμοποιούνται, 2) τις πολλαπλές δομές των δεδομένων και κυρίως 3) την διαφορετική ή διφορούμενη σημασιολογιά (π.χ. έννοιες ή ερμηνείες).

Είναι αναγκαίο να υπάρξει σημαντική προσπάθεια με σκοπό τη σύνδεση διαφορετικών πηγών λόγω της έλλειψης σαφών σημασιολογικών ορισμών των μεταβλητών, μέτρων και κατασκευών, όμως αυτό το ζήτημα μπορεί να διευκολυνθεί με τη σημασιολογική διαλειτουργικότητα, η οποία επιτρέπει την ανταλλαγή δεδομένων με κοινό νόημα. Μάλιστα, μία κοινή προσέγγιση στη σημασιολογική ολοκλήρωση των δεδομένων είναι μέσω της χρήσης οντολογιών, δηλαδή βασιζόμενοι σε ένα τυποποιημένο και ελεγχόμενο λεξιολόγιο για την περιγραφή των δεδομένων και των σχέσεων μεταξύ τους, όπου μία οντολογία μπορεί να αναπαραστήσει τυπικά και υπολογιστικά την γνώση για έναν τομέα. Συνεπώς, με μία καθολική εννοιολογική αναπαράσταση όλων των πληροφοριών, μία προσέγγιση της σημασιολογικής

ολοκλήρωσης μας δίνει την δυνατότητα να γεφυρώσουμε την ετερογένεια των δεδομένων σε πολλαπλές πηγές και πολλαπλούς τομείς.

Πολλές βιοϊατρικές οντολογίες είναι ήδη διαθέσιμες και χρησιμοποιούνται ευρέως στην Ιατρική, π.χ. The International Classification of Diseases (ICD) ή The Systematized Nomenclature of Medicine and Clinical Terms (SNOMED CT). Παρ' όλα αυτά ένα ενοποιημένο πλαίσιο δεδομένων, με βάση την οντολογία,απαιτείται για την ικανοποίηση των αυξανόμενων αναγκών σύνδεσης και ολοκλήρωσης των δεδομένων από πολλαπλούς τομείς. Επιπροσθέτως, πέρα από τις παραδοσιακές προσεγγίσεις της χρήσης κοινών στοιχείων των δεδομένων και κοινών μοντέλων των δεδομένων, όπως οι διεθνείς προσπάθειες που πραγματοποιούνται για την δημιουργία του Observational Medical Outcomes Partnership (OMOP) CDM, ένα πλαίσιο ενοποίησης των δεδομένων βασισμένο στην οντολογία, μπορεί να χρησιμοποιηθεί για την αναπαράσταση των μεταδεδομένων, την δημιουργία παγκόσμιων εννοιολογικών χαρτών, την αυτοματοποιήση του ελέγχου ποιότητας των δεδομένων, καθώς και την υποστήριξη υψηλού επιπέδου σημασιολογικών ερωτημάτων. Επιπλέον, η έρευνα που σχετίζεται με την σημασιολογία των EHR βελτιώνει όχι μόνο την ολοκλήρωση και την διαλειτουργικότητα των δεδομένων, αλλά μπορεί ακόμη να προωθήσει την επιστήμη για την φαινοτυποποίηση των ασθενειών. Είναι πολύ σημαντικό να αναφέρουμε ότι οι οντολογίες μπορούν να χρησιμοποιηθούν για την διευκόλυνση μίας τυπικής τεκμηρίωσης των διαδικασιών ολοκλήρωσης των δεδομένων. Για παράδειγμα, μέσω της κωδικοποίησης των σχέσεων μεταξύ των μεταβλητών που πρέπει να ενσωματωθούν σε διαφορετικές πηγές. Αυτό μπορεί να έχει σημαντικό αντίκτυπο στην ερευνητική αυστηρότητα, τη διαφάνεια και την αναπαραγωγικότητα μεταξύ των επιστημόνων, καθώς και στην επαναχρησιμοποίηση και την ευελιξία των δεδομένων.

Η σημασιολογική ολοκλήρωση μπορεί να συμβεί σε διαφορετικά επίπεδα έρευνας στον τομέα της υγειονομικής περίθαλψης και όχι μόνο στο επίπεδο των δεδομένων με τα EHR. Επιπλέον, τα σχέδια των μελετών σε ολοκληρωμένες πηγές δεδομένων πρέπει να υποστηρίζονται από την κατάλληλη σημασιολογία. Στο Σχήμα 2.2 (Prosperi *et al.,* 2018) συνοψίζεται το παράδειγμα σημασιολογικής ολοκλήρωσης σε διάφορα επίπεδα: 1) το επίπεδο των δεδομένων που ενσωματώνει τόσο τα EHR όσο και οι PHR(Personal Health Records) πηγές των δεδομένων (inter-domain), 2) το επίπεδο έννοιας, αντιστοίχηση ορολογιών και οντολογιών (domain--contextual), 3) το επίπεδο σχεδιασμού μελέτης, επιτρέποντας την χρήση τυποποιημένων λειτουργικών διαδικασιών και αναπαραγωγικότητα σε άλλες πηγές (domain-contextual), 4) το επίπεδο εξαγωγής συμπερασμάτων, όπου προσδιορίζονται οι κατάλληλες μέθοδοι της Στατιστικής Μάθησης όσον αφορά τον σχεδιασμό της μελέτης, την κλιμάκωση των αναλύσεων σε υπολογιστές υψηλής απόδοσης, καθώς και την δημιουργία μοντέλων και εφαρμογών προς όφελος της δημόσιας υγείας (trans-domain).

Η σημασιολογική ολοκλήρωση επιτρέπει την προσθήκη νέων δεδομένων ή στοιχείων της οντολογίας, την ευελιξία (π.χ. τροποποίηση των υπάρχοντων σχέδιων μελέτης ή εκτέλεση σε διαφορετικούς τομείς) και την διαφάνεια (π.χ. αναπαραγωγιμότητα των αποτελεσμάτων, επικύρωση, βελτίωση των μοντέλων). Για παράδειγμα, διαλειτουργικές σημασιολογίες και ερευνητικά αντικείμενα αποτέλεσαν την κινητήρια δύναμη για το έργο «asthma e-lab». Ως ένα ασφαλές διαδικτυακό περιβάλλον για την υποστήριξη ενσωμάτωσης, περιγραφής και κοινής

χρήσης δεδομένων, το e-lab συνδυαζόμενο με υπολογιστικούς πόρους και ένα επιστημονικό κοινωνικό δίκτυο, υποστηρίζει τη συνεργατική έρευνα και την μεταφορά γνώσεων.

Ένα άλλο σχετικό παράδειγμα είναι το πρόγραμμα Observational Health Data Sciences and Informatics (OHDSI), στόχος του οποίου είναι η δημιουργία και η εφαρμογή αναλυτικών δεδομένων ανοικτού κώδικα σε ένα μεγάλο δίκτυο βάσεων δεδομένων υγείας με σκοπό την βελτίωση της ανθρώπινης υγείας και ευημέριας.

Σχήμα 2.2
Σημασιολογική ολοκλήρωση των δεδομένων, σχεδιασμός μελέτης και εξαγωγή συμπερασμάτων

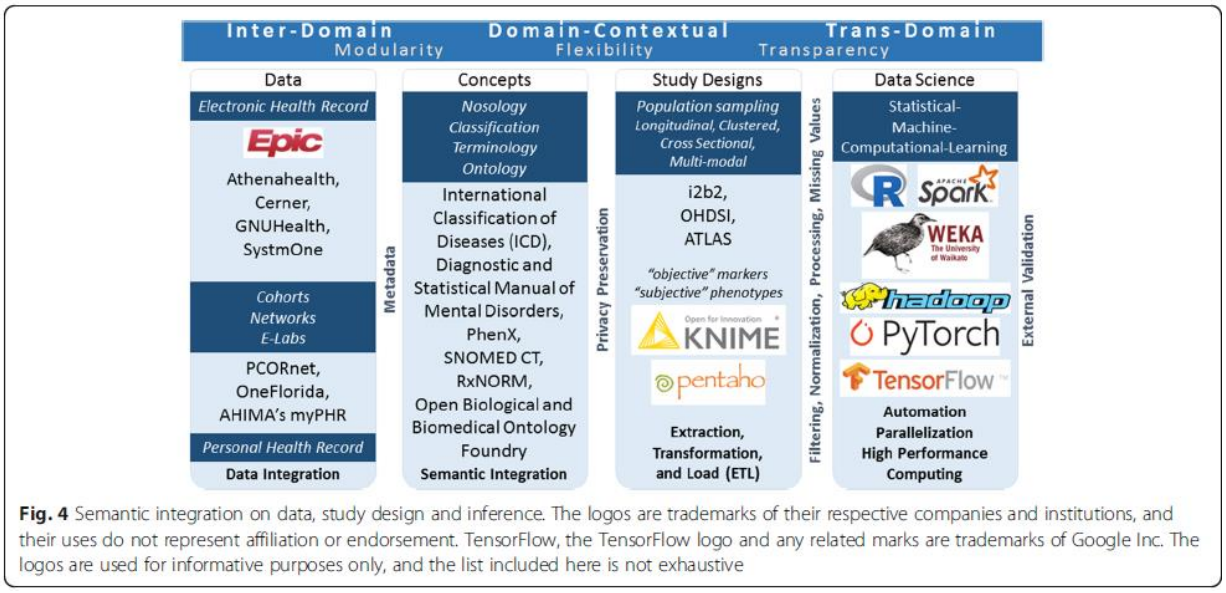


Fig. 4 Semantic integration on data, study design and inference. The logos are trademarks of their respective companies and institutions, and their uses do not represent affiliation or endorsement. TensorFlow, the TensorFlow logo and any related marks are trademarks of Google Inc. The logos are used for informative purposes only, and the list included here is not exhaustive

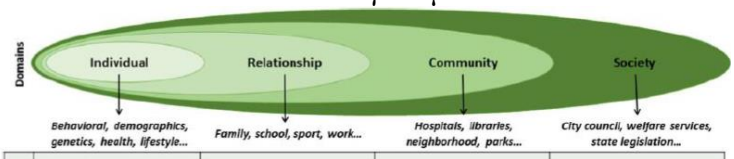
Το OHDSI χρησιμοποιεί το OMOP (Observational Medical Outcomes Partnership) που είναι ένα κοινό μοντέλο δεδομένων, το οποίο διαθέτει μία σειρά εφαρμογών για την βελτιστοποίηση της ενσωμάτωσης των EHR, της αξιολόγησης για την ποιότητα και τον καθαρισμό των δεδομένων το ACHILLES. Επίσης, διαθέτει τυποποιημένο λεξιολόγιο για το OMOP, για την αναζήτηση των δεδομένων το ATLAS και την ανάλυση τους το CYCLOPS (Prosperi *et al*., 2018).

2.2.2 Σχεδιασμοί μελέτης (hollow learning, shallow design)

Οι Prosperi *et al.* (2018) αναφέρουν πως με την πρόοδο της τεχνολογίας και τη σύνδεση των δεδομένων, η έρευνα ενός τομέα αντικαθίσταται από πολυεπίπεδες και πολυτομεακές μελέτες. Έτσι, αυτή η αύξηση της πολυπλοκότητας και της ετερογένειας των μελετών επηρεάζει το σχεδιασμό τους, τόσο των παρατηρητικών όσο και των προοπτικών μελετών. Ιδιαίτερα, για τις παρατηρητικές μελέτες, υπάρχει τεράστιος δυνητικά διαθέσιμος όγκος δεδομένων, αλλά η

πρόσβαση και η χρήση αυτών των ετερογενών πηγών δεδομένων είναι αναγκαίο να εξορθολογιστεί για την αντιμετώπιση της μεροληψίας, τον εντοπισμό αξιοποιήσιμων εισροών και να ληφθούν υπόψη οι δεοντολογικές ανάγκες. Σε έρευνες στον τομέα της ψυχολογίας, έχει προταθεί ότι οι μελέτες με βάση τα δεδομένα θα πρέπει να καθοδηγούνται από μία αιτιολογική θεωρία σχετικά με τον σχεδιασμό της μελέτης. Αυτές οι θεωριές βασίζονται στην αξιολόγηση των επιστημονικών αποδεικτικών στοιχείων ως «causal pathways» της νόσου. Ο υβριδισμός της χρήσης της θεωρίας για καθοδήγηση του σχεδιασμού ("top-down" προσέγγιση) με την έρευνα με βάση τα δεδομένα ("bottom-up" προσέγγιση) μπορεί να είναι πολύ χρήσιμος για την ανάπτυξη πολυεπίπεδων και πολυτομεακών μοντέλων πρόβλεψης, που περιλαμβάνουν ατομικά και πληθυσμιακά επίπεδα. Υπάρχουν διάφορα εννοιολογικά μοντέλα που μπορούν να χρησιμοποιηθούν, όπως το κοινωνικο-οικολογικό μοντέλο ή το πολλαπλών αιτιών μοντέλο. Στο Σχήμα 2.3 (Prosperi *et al.*, 2018) παρουσιάζεται το κοινωνικό-οικολογικό μοντέλο, οι τομείς από όπου αντλούνται οι πληροφορίες και μία σειρά από πηγές δεδομένων (κυρίως διαθέσιμες στις Ηνωμένες Πολιτείες) που μπορούν να χρησιμοποιηθούν για την εξαγωγή σχετικών χαρακτηριστικών που αφορούν τις διαστάσεις των τομέων.

Σχήμα 2.3
Κοινωνικο-οικολογικό μοντέλο



Το πλεονέκτημα της χρήσης ενός θεωρητικού μοντέλου είναι ότι είναι δυνατή η αποδόμηση του μοντέλου πρόβλεψης για τον έλεγχο υποθέσεων ή τον εντοπισμό νέων τομέων που χρήζουν περαιτέρω διερεύνησης. Για παράδειγμα, ας υποθέσουμε ότι χρησιμοποιούμε το κοινωνικό-οικολογικό μοντέλο και ενσωματώνουμε το ατομικό επίπεδο των EHR και τους γενετικούς δείκτες με κοινωνικούς (σε επίπεδο κοινότητας) και οικολογικούς δείκτες, σε ένα συγκεκριμένο χρονικό ορίζοντα, για να προσδιορίσουμε τον πληθυσμό κινδύνου οξέος ή χρόνιου άσθματος. Ορισμένες μεταβλητές ατομικού ή κοινοτικού επιπέδου μπορεί να βρεθεί ότι συμβάλλουν σε αυξημένο κίνδυνο και μέσω διατομεακών αλληλεπιδράσεων, το ποσοστό της διακύμανσης που εξηγείται μπορεί να αυξηθεί. Ακόμη, οι μεταβλητές σε κάθε τομέα μπορούν να εξεταστούν για να διαπιστωθεί εάν είναι εφαρμόσιμες ή αμετάβλητες (π.χ. περιβαλλοντικές εκθέσεις έναντι γενετικής) και δεοντολογικά αξιοποιήσιμες ή όχι (π.χ. η βαθμολογία στέρησης σε μία γειτονιά έναντι του φυλετικού προφίλ). Οι πληροφορίες αυτές μπορούν να αξιοποιηθούν για τον προσδιορισμό ενός κατάλληλου μοντέλου κινδύνου και να επιλεγούν παράγοντες που μπορούν να τροποποιηθούν για να μειωθεί ο κίνδυνος της νόσου.

Ένα από τα μεγαλύτερα εμπόδια στο σχεδιασμό μελετών, ιδίως για τις παρατηρητικές μελέτες ή τις αναδρομικές μελέτες, αφορά τον αποτελεσματικό εντοπισμό και την αντιμετώπιση της μεροληψίας. Με τα Big Data, το ζήτημα αυτό είναι σοβαρό, λόγω της ετερογένειας που υπάρχει στην συλλογή των δεδομένων, την επαλήθευση της πηγής και τη δειγματοληπτική μεροληψία μεταξύ άλλων. Οι ερευνητές πρέπει να προσέχουν "ότι τα Big Data αποτελούν υποκατάστατο και όχι συμπλήρωμα της παραδοσιακής συλλογής και ανάλυσης των δεδομένων". Με τα EHR, η μεροληψία υπερκαλύπτει την τυχαιοποίηση. Τα EHR δεδομένα είναι εγγενώς μεροληπτικά από την δομή του πληθυσμού των ασθενών, τη συχνότητα των επισκέψεων υγειονομικής περίθαλψης, τα διαγνωστικά κριτήρια και τους τρόπους περίθαλψης. Επιπροσθέτως, τα αρχεία συνταγογράφησης των φαρμάκων αντικατοπτρίζουν ως επί το πλείστον την ένδειξη ή την πρωτοπαθής μεροληψία. Ακόμη και οι πιο προηγμένες στατιστικές μέθοδοι δεν μπορούν να διαχωρίσουν την μεροληψία, αλλά μπορούν να εκπαιδευτούν με μεγάλη ακρίβεια. Ως εκ τούτου, η χρήση μεθόδων Deep Learning με ακατέργαστα EHR δεδομένα μπορεί να είναι μια πολύ κακή ιδέα, παρόλο που αποδίδει εκπληκτικές προβλέψεις. (Prosperi *et al.*, 2018). Στην πραγματικότητα, «οι μεροληπτικοί αλγόριθμοι είναι παντού και κανένας δεν φαίνεται να νοιάζεται» (Knight, 2017). Το συγκεκριμένο πρόβλημα δεν είναι καινούργιο και γίνεται επικίνδυνο αν χρησιμοποιηθεί για την λήψη αποφάσεων. Για παράδειγμα, ο αλγόριθμος που χρησιμοποιήθηκε από την ProPublica (Angwin *et al.*, 2016) για την αξιολόγηση του προφίλ των σωφρονιστικών παραβατών για εναλλακτικές κυρώσεις (COMPAS), ένα εργαλείο που χρησιμοποιείται για την πρόβλεψη του κινδύνου υποτροπής ενός ατόμου, αποτελεί ένα σοβαρό παράδειγμα μεροληπτικού αλγορίθμου (Prosperi *et al.*, 2018).

2.3 Πηγές και χρήσεις τύπων δεδομένων που χρησιμοποιούνται για την δημόσια υγεία

Σε αυτήν την ενότητα θα επικεντρωθούμε στον ορισμό των τύπων δεδομένων που χρησιμοποιούνται από την Δημόσια Υγεία. Παρέχεται ένας σύντομος ορισμός για κάθε τύπο δεδομένων, γίνεται συζήτηση για τις πιθανές πηγές των δεδομένων και υπάρχουν παραδείγματα για το πως μπορούν να χρησιμοποιηθούν αυτά τα δεδομένα.

2.3.1 Omics Data

Τα Omics Data μελετούν τα χαρακτηριστικά των μοριακών προφίλ (π.χ. γονίδια, πρωτεΐνες, μεταβολίτες) για την καλύτερη κατανόηση της σχέσης μεταξύ των μορίων ενός οργανισμού. Τα δεδομένα που συνήθως μελετώνται περιλαμβάνουν γονίδια, χημικές ενώσεις, πρωτεΐνες, μεταβολίτες και υδατάνθρακες. Η μελέτη των βιολογικών παραγόντων και των αλληλεπιδράσεών τους, συμπεριλαμβανομένων των σχέσεων μεταξύ των διαφόρων τύπων "-omics", είναι ιδιαίτερα σημαντική. Οι μελέτες αυτές μπορούν να οδηγήσουν σε διάφορες

γνώσεις, όπως η κατανόηση του κατά πόσον υπάρχουν διαφορετικά υποείδη μιας νόσου και η μελέτη παραγόντων που μπορεί να κάνουν ένα πιθανό φάρμακο πιο αποτελεσματικό. Παρακάτω ακολουθούν πιθανές πηγές των Omics Data:

Τα δεδομένα αυτά λαμβάνονται κυρίως από πηγές όπως οι επόμενης γενιάς τεχνολογίες αλληλουχίας (next generation sequencing - NGS). Οι τεχνολογίες NGS αναλύουν παράλληλα πολλαπλές αλυσίδες DNA, επιταχύνοντας την διαδικασία (μπορούμε να έχουμε σε σειρά ολόκληρο το ανθρώπινο DNA μέσα σε μια ημέρα) και τις καθιστούν πιο αποδοτικές από πλευράς κόστους. Ένα παράδειγμα των τεχνολογιών NGS είναι η βαθιά αλληλουχία, η οποία περιλαμβάνει την αλληλουχία του ίδιου του γονιδίου με στόχο πολλές φορές την ανίχνευση σπάνιων μεταλλάξεων ή παραλλαγών σε ένα κύτταρο. Παρακάτω ακολουθούν παραδείγματα χρήσης των Omics Data:

Μία πιθανή χρήση των Omics Data περιλαμβάνει την μελέτη γενετικών χαρακτηριστικών για την προσαρμογή των θεραπειών. Για παράδειγμα, πολλές μελέτες μπορούν να συσχετίσουν ασθένειες όπως η παχυσαρκία και η κυστική ίνωση με γενετικούς παράγοντες. Επίσης, αρκετές μεταδοτικές ασθένειες όπως η γρίπη, μπορούν να έχουν γενετικούς παράγοντες που έχουν επιπτώσεις στην αντοχή ενός ατόμου στην θεραπεία. Ένα άλλο παράδειγμα χρήσης των Omics Data είναι η ανίχνευση των βιοδεικτών, οι οποίοι μπορούν να βοηθήσουν τους ερευνητές να κατανοήσουν καλύτερα ορισμένες ασθένειες. Η ανίχνευση βιοδεικτών μπορεί επίσης να έχει μεγάλη σημασία στην βελτίωση των κλινικών δοκιμών. Ένα εργαλείο για την παρακολούθηση βιοδεικτών μπορεί να επιτρέψει τη μέτρηση των διαφορών στους ασθενείς σε βιολογικό επίπεδο, επιτρέποντας στους ερευνητές να εκτιμήσουν τις επιδράσεις ενός φαρμάκου πολύ ταχύτερα και με μεγαλύτερη ακρίβεια από ό,τι επιτρέπουν οι τρέχουσες μέθοδοι. Τα Omics Data και οι γονιδιωματικοί βιοδείκτες, που συνδέονται με υπολογιστικούς παράγοντες όπως οι αλγόριθμοι μηχανικής μάθησης, έχουν τη δυνατότητα να ανακαλύψουν νέες συσχετίσεις μεταξύ μοριακών προφίλ και άλλων κλινικών μεταβλητών που δεν θα αξιολογούνταν διαφορετικά και μπορεί να οδηγήσει στον εντοπισμό νέων κλινικά σημαντικών υποείδων σε κλινικές δοκιμές (Velmovitsky *et al.*, 2021).

2.3.2 Κλινικά δεδομένα

Τα κλινικά δεδομένα συλλέγονται από τους ασθενείς κατά τη διάρκεια των θεραπειών ή σε κλινικές δοκιμές, συνήθως σε μια ιατρική μονάδα. Παραδείγματα κλινικών δεδομένων περιλαμβάνουν διάφορες εργαστηριακές εξετάσεις, όπως αξονικές τομογραφίες, ηλεκτροκαρδιογραφήματα και ακτίνες Χ. Ορισμένα είδη κλινικών δεδομένων μπορούν να συλλεχθούν από άτομα που χρησιμοποιούν συσκευές υγείας, όπως ασύρματες ζυγαριές, smartwatches και όργανα μέτρησης της αρτηριακής πίεσης. Ωστόσο, αυτές οι πηγές και οι χρήσεις των κλινικών δεδομένων δεν παρακολουθούνται επίσημα από ιατρικές μονάδες και δεν καταχωρούνται στα EHR των ασθενών.

Παρακάτω ακολουθούν πιθανές πηγές των κλινικών δεδομένων:

- EHR
- Administrative data
- Claims data

- Μητρώα ασθενών/ασθενειών
- Έρευνες υγείας και δεδομένα κλινικών δοκιμών.

Παρακάτω ακολουθούν παραδείγματα χρήσης των κλινικών δεδομένων.

Τα κλινικά δεδομένα μπορούν να χρησιμοποιηθούν για την πρόβλεψη των διαγνώσεων των ατόμων, την αντίδραση και τις ανταπόκριση που έχουν τα άτομα σε διάφορες θεραπείες και των αποτελεσμάτων επιβίωσης. Μάλιστα, έχουν υπάρξει διάφορα ερευνητικά προγράμματα που χρησιμοποιούν Deep Learning για τον σκοπό αυτό (Velmovitsky *et al.*, 2021). Για παράδειγμα, ένας ερευνητής στο Trinity College εκπαίδευσε ένα μοντέλο Μηχανικής Μάθησης σε 110 σαρώσεις μαγνητικής τομογραφίας, όπου πέτυχε ακρίβεια 90% στην έγκαιρη διάγνωση της αμυοτροφικής πλευρικής σκλήρυνσης (Savage, 2017). Οι Gulshan *et al.* (2016) ανέπτυξαν Deep Learning αλγόριθμους που ξεπέρασαν τις επιδόσεις των ειδικών στην ανίχνευση της διαβητικής αμφιβληστροειδοπάθειας και του διαβητικού οιδήματος ωχράς κηλίδας σε εικόνες του βυθού του αμφιβληστροειδούς.

Τα κλινικά δεδομένα μπορούν επίσης να συνδυαστούν με τεχνολογίες απομακρυσμένης παρακολούθησης για την παρακολούθηση ασθενών εντός και εκτός νοσοκομείων, χωρίς να υπάρχει η ανάγκη της φυσικής επαφής, η οποία είναι απαραίτητη κατά τη διάρκεια της σημερινής πανδημίας COVID-19 (Velmovitsky *et al.*, 2021). Για παράδειγμα, οι Dhillon *et al*. (2018) προτείνουν ένα σύστημα που επεξεργάζεται διάφορα συμβάντα από φορητούς αισθητήρες τα οποία συλλέγουν δεδομένα ECG (καρδιογραφήματα-electrocardiogram), EEG (ηλεκτροεγκαφαλογράφος - Electroencephalogram) και αρτηριακής πίεσης. Ακόμη, αυτό το σύστημα έχει τη δυνατότητα να παρέχει συναγερμούς σε περίπτωση επείγοντος συμβάντος.

2.3.3 Social Data

Τα κοινωνικά δεδομένα (Social Data) είναι οι πληροφορίες που μοιράζονται δημοσίως στα μέσα κοινωνικής δικτύωσης, συμπεριλαμβανομένων και των πληροφοριών σχετικά με την τοποθεσία του χρήστη, τον κύκλο των φίλων που έχει, καθώς και την γλώσσα επικοινωνίας που χρησιμοποιείται. Επεκτείνοντας την έννοια των κοινωνικών δεδομένων, η χρήση και το νόημα ορισμένων δεδομένων είναι στενά συνδεδεμένα με τις κοινωνικές αλληλεπιδράσεις (όπως τα δεδομένα συμπεριφοράς, παρόλο που δεν παράχθηκαν ούτε συλλέχθηκαν μέσω των μέσων κοινωνικής δικτύωσης). Παρακάτω ακολουθούν πιθανές πηγές των κοινωνικών δεδομένων:

Πιθανές πηγές των κοινωνικών δεδομένων αποτελούν διάφορα μέσα κοινωνικής δικτύωσης όπως είναι το Facebook, το Twitter και το Google (π.χ. ιστορικό αναζήτησης, Google trends), καθώς και δεδομένα που αντιπροσωπεύουν κοινωνικές αλληλεπιδράσεις.

Παρακάτω ακολουθούν παραδείγματα χρήσης των κοινωνικών δεδομένων.

Τα κοινωνικά δεδομένα μπορούν να χρησιμοποιηθούν για την εκτίμηση των χαρακτηριστικών της συμπεριφοράς ενός ατόμου. Για παράδειγμα, το 20% των ασθενών με χρόνιες παθήσεις μοιράζονται τις εμπειρίες τους στο διαδίκτυο με άλλους ασθενείς, διημιουργώντας διαδικτυακές κοινότητες στα μέσα κοινωνικής δικτύωσης. Ακόμη, τα δεδομένα αυτά μπορούν να αξιοποιηθούν για σκοπούς υγείας. Οι εικόνες τροφίμων που βρίσκονται στο Instagram μπορούν να χρησιμοποιηθούν για την μελέτη της διατροφικής συμπεριφοράς των εφήβων

(Velmovitsky *et al.*, 2021). Οι Deiner *et al.* (2016) χρησιμοποιήσαν δεδομένα του Twitter για να εντοπίσουν την ύπαρξη ατόμων οι οποίοι είναι διαγνωσμένοι με επιπεφυκίτιδα. Επιπλέον, οι πιο προηγμένες μελέτες είναι σε θέση να εντοπίσουν διάφορους δείκτες, ακόμα και την πρόβλεψη της διάγνωσης. Για παράδειγμα οι Reece και Danforth (2017) εφάρμοσαν μεθόδους Μηχανικής Μάθησης για τον εντοπισμό δεικτών κατάθλιψης χρησιμοποιώντας φωτογραφίες του Instagram, οι Jain *et al.* (2015) χρησιμοποιήσαν δεδομένα του Twitter για την πρόβλεψη της αϋπνίας. Επιπροσθέτως, οι Odlum και Yoon (2015) έδειξαν ότι μία αύξηση των δεδομένων στο Twitter σχετικά με τον Έμπολα υπέδειξε αυξημένη συχνότητα εμφάνισης του ιού στη Νιγηρία τρεις ημέρες πριν από την ειδοποίηση που μετέφεραν οι ειδήσεις και επτά ημέρες πριν από τις επίσημες προειδοποιήσεις των Κέντρων Ελέγχου Ασθενειών. Ωστόσο, πρέπει να είναι κανείς προσεκτικός όσον αφορά τη χρήση των κοινωνικών δεδομένων, καθώς πολλά από αυτά τα ερευνητικά αποτελέσματα μπορεί να μην είναι επαναλήψιμα. Επιπλέον, οι ερευνητές πρέπει να προσέχουν την ιδιωτικότητα των χρηστών, καθώς τα δεδομένα αυτά μπορεί να μην έχουν συλλέχθει ως μέρος μιας ερευνητικής μελέτης (Velmovitsky *et al.*, 2021).

2.3.4 PGHD Data

Τα PGHD ή «αυτοπαραγόμενα» (real-world) δεδομένα περιλαμβάνουν πληροφορίες σχετικά με την υγεία και την συμπεριφορά των ατόμων, όπου τα δεδομένα αυτά συλλέγονται μέσω προσωπικών smart συσκευών. Πολλοί τύποι δεδομένων στη βιβλιογραφία που σχετίζονται με τα προσωπικά κλινικά δεδομένα, όπως το βάρος και ο καρδιακός ρυθμός, δεν χρησιμοποιούνται επισήμως για κλινική χρήση, αλλά μπορούν να χρησιμοποιηθούν από τα άτομα για την αυτοπαρακολούθηση της υγείας και των στόχων που έχουν θέση ως προς την συμπεριφορά τους στην καθημερινή τους ζωή (π.χ. διακοπή του καπνίσματος, κατανάλωση νερού, άσκηση κ.λπ.). Πιθανές πηγές των PGHD δεδομένων αποτελούν δεδομένα που συλλέγονται μέσω προσωπικών κινητών/wearable συσκευών, όπως smartphones, smartwatches, smart θερμοστάτες.

Παρακάτω ακολουθούν παραδείγματα χρήσης των PGHD δεδομένων:
Εφαρμογές όπως το Apple Health και το Google Fit μπορούν να συλλέγουν και να διαχείριζονται ατομικά δεδομένα, όπως η ρουτίνα άσκησης, η πρόσληψη θερμίδων, ο καρδιακός ρυθμός και το βάρος. Υπάρχουν ανάμεικτα στοιχεία σχετικά με την ακρίβεια των δεδομένων που συλλέγονται μέσω smart αισθητήρων και υπάρχουν πολλοί παράγοντες που μπορούν να το επηρεάσουν. Μια πρόσφατη συστηματική ανασκόπηση των wearables συσκευών διαπίστωσε ότι τα δεδομένα μπορεί να είναι υποβαθμισμένα ή υπερεκτιμημένα σε αρκετές συσκευές και μοντέλα και μια μελέτη που συνέκρινε το FitBit Flex και το ActiGraph GT3X+ διαπίστωσε ότι το Fitbit υποεκτιμούσε σημαντικά τα βήματα σε κανονικές συνθήκες. Εκτός από την αυτοδιαχείριση της υγείας, οι smart συσκευές μπορούν επίσης να βοηθήσουν τους ερευνητές να συλλέξουν δεδομένα για μελέτες υγείας. Για παράδειγμα, το ResearchKit είναι ένα πρόγραμμα που αναπτύχθηκε από την Apple για να βοηθήσει τους ερευνητές να δημιουργήσουν εφαρμογές που προσλαμβάνουν συμμετέχοντες σε μελέτες και συλλέγουν δεδομένα υγείας από συνδεδεμένες συσκευές. Ένα παράδειγμα είναι το mPower, μια εφαρμογή iOS που μετρά την ισορροπία, την δεξιοσύνη και τον βηματισμό με επιταχυνσιόμετρο και το

γυροσκόπιο του iPhone για την καλύτερη κατανόηση της νόσου του Πάρκινσον. Έχει περισσότερους από 10.000 χρήστες (το 93% δεν συμμετείχε ποτέ σε καμία μελέτη) και έγινε η μεγαλύτερη, στην ιστορία, μελέτη για τη νόσο του Πάρκινσον. Φυσικά, πρέπει να υπάρξουν πολλές ακόμη μελέτες σχετικά με το κατά πόσο η ακρίβεια των δεδομένων αυτών είναι επαρκής για την υποστήριξη της έρευνας (Velmovitsky *et al.*, 2021).

2.3.5 Περιβαλλοντικά Δεδομένα

Τα περιβαλλοντικά δεδομένα σχετίζονται με πληροφορίες που συλλέγονται από το γενικό πλαίσιο μέσα στο οποίο βρίσκονται τα άτομα και πληθυσμοί. Παραδείγματα περιβαλλοντικών δεδομένων είναι η ατμοσφαιρική ρύπανση, οι αλλαγές της θερμοκρασίας, η ποιότητα του νερού, τα ποσοστά εγκληματικότητας και το πόσο φιλική είναι μία γειτονιά για να περπατήσεις. Παρακάτω ακολουθούν πιθανές πηγές των περιβαλλοντικών δεδομένων:

- Αισθητήρες ατμοσφαιρικής ρύπανσης και του καιρού
- GPS (Global Positioning System)
- GIS (Geographic Information System)
- Δεδομένα της αστυνομίας
- Βάσεις δεδομένων των μέσων μαζικής μεταφοράς
- Διάφορες εφαρμογές των κινητών τηλεφώνων.

Παρακάτω ακολουθούν παραδείγματα χρήσης των περιβαλλοντικών δεδομένων:
Είναι γνωστό ότι ορισμένοι περιβαλλοντικοί παράγοντες μπορούν άμεσα να επηρεάσουν την υγεία μας, όπως η ποιότητα του αέρα και του νερού. Φυσικές και ανθρωπογενείς επιπτώσεις στους φυσικούς πόρους σχετίζονται με διάφορες οξείες και χρόνιες ασθένειες. Οι επιπτώσεις αυτές μπορούν να αξιολογηθούν για την ανάπτυξη καλύτερων στρατηγικών υγειονομικής περίθαλψης για άτομα και πληθυσμούς, αν και οι τελευταίοι συνδέονται συχνότερα με αυτές τις ανησυχίες. Οι αισθητήρες που διατίθενται στο εμπόριο, όπως οι Ecobee smart θερμοστάτες, για για παράδειγμα, αποτελούν νέες λύσεις με στόχο την συλλογή διάφορων περιβαλλοντικών μεταβλητών των δεδομένων. Άλλοι παράγοντες που σχετίζονται με το περιβάλλον, είναι η συχνότητα της εγκληματικότητας, των μετακινήσεων και του πολεοδομικού σχεδιασμού και μπορούν να έχουν σημαντικό αντίκτυπο στην υγεία του ατόμου και του πληθυσμού. Πολλές μελέτες έχουν δείξει ότι παράγοντες όπως η καθημερινή μετακίνηση, μπορούν να έχουν αρνητική επίπτωση στην υγεία, επηρεάζοντας το μεταβολισμό μας (π.χ. αύξηση των επιπέδων σακχάρου και χοληστερόλης), τη στάση του σώματος και τον ύπνο (Velmovitsky *et al.*, 2021).

2.3.6 Δημογραφικά δεδομένα

Αποτελούν όλα τα δεδομένα εκείνα που σχετίζονται με την περιγραφή ατόμων ή πληθυσμών που είναι υπό μελέτη, όπως η ηλικία, το φύλο, η εθνικότητα, η εκπαίδευση, η εργασιακή απασχόληση και το εισόδημα.

Παρακάτω ακολουθούν πιθανές πηγές των δημογραφικών δεδομένων:

Υπάρχουν πολλοί τρόποι για την απόκτηση δημογραφικών δεδομένων. Οι ίδιοι οι συμμετέχοντες μπορούν να αναφέρουν τα δημογραφικά τους χαρακτηριστικά όταν διεξάγεται

μια μελέτη. Οι ερευνητές μπορούν επίσης να έχουν πρόσβαση σε δεδομένα από μεγάλες έρευνες και απογραφές, όπου συνήθως συλλέγουν αυτές τις πληροφορίες. Οι πληροφορίες αυτές μπορούν επίσης να προέρχονται από κλινικά αρχεία των ατόμων (με αυτόν τον τρόπο, μπορούν να θεωρηθούν ως μορφή κλινικών δεδομένων) και επίσης να συλλέγονται μέσω των μέσων κοινωνικής δικτύωσης. Παρακάτω ακολουθούν παραδείγματα χρήσης των δημογραφικών δεδομένων:

Τα δημογραφικά δεδομένα χρησιμοποιούνται συνήθως για τη διαστρωμάτωση των συμμετεχόντων σε μια μελέτη σύμφωνα με ορισμένα χαρακτηριστικά τους, για παράδειγμα για τον εντοπισμό αν υπάρχει ένας συγκεκριμένος παράγοντας ή όχι στο δείγμα των συμμετεχόντων με το ίδιο χαρακτηριστικό. Η χρήση των δημογραφικών δεδομένων μπορεί επίσης να οδηγήσει τους ερευνητές να εντοπίσουν και να αποφύγουν τυχόν μεροληψίες ή σύγχυση, καθώς και να εξετάσουν την εγκυρότητα της μελέτης, διότι οι αναλύσεις που πραγματοποιούνται σε έναν συγκεκριμένο πληθυσμό μπορεί να μην είναι δυνατό να μεταφραστούν άμεσα σε πληθυσμούς με διαφορετικά χαρακτηριστικά. Με αυτόν τον τρόπο, οι δημογραφικές πληροφορίες λειτουργούν ως ένα εργαλείο κανονικοποίησης, για να διασφαλιστεί ότι τα δεδομένα που συλλέγονται από τα άτομα, μπορούν να συγκριθούν όσον αφορά τη μεταβλητή ή το γεγονός που ενδιαφέρει τον ερευνητή στην ίδια βάση, ελαχιστοποιώντας την πιθανότητα να μην υπάρχουν άλλοι παράγοντες που να προκαλούν εσφαλμένη ανάλυση ή αποτελέσματα (Velmovitsky *et al.*, 2021).

Στον παρακάτω πίνακα (2.4 - Velmovitsky *et al.*, 2021) συνοψίζονται όλα αυτά που έχουμε αναφέρει παραπάνω για τους τύπους δεδομένων που χρησιμοποιούνται από την Δημόσια Υγεία.

Σχήμα 2.4
Τύποι Δεδομένων

TABLE 1 | Data Types Identified in the Literature Review.

Data Types	Omics	Clinical	Social	PGHD	Environmental	Demographic
Definition	The study that explores the roles, relationships, and actions of the various types of molecules	Data collected through the course of treatment or in the processes of clinical trials	Information publicly shared on social media and related to personal health data collected by the individual	Information on the health and behavior of individuals collected through personal smart devices	Information gathered from the context in which individuals and populations are immersed	Information describing attributes of the population under study
Examples	Genomics, epigenomics, proteonomics, transcriptomics, etc.	EHR, laboratory tests, MRI, CT Scans, Administrative data, etc.	Social media posts, GPS location, data generated through smartwatches, smartbands, etc.	Smart personal devices data such as sleep, heart rate, and physical activity	Natural resources quality, temperature, crime rates, traffic, walkability of neighborhoods, etc.	Age, sex, education, income, ethnicity, employment, etc.
Possible Uses	Oncology and genetics studies, pharmacogenomics, omic biomarkers, clinical trials improvement	Predictive medicine, trends and correlation identification, clinical trials improvement, false alarm mitigation	Social media use, behavior and social habits assessment (e.g., quit smoking, lose weight, etc.)	Health self-management and research into behavioral aspects of an individual (e.g., dietary intake tracking, vital signs log, etc.)	Air and water quality monitoring, traffic impact measuring, social factors impact on life quality	Stratifying populations under study in groups with the same attributes, preventing biases and confounding, and serving as a normalization tool for comparing data points in a study

Παρακάτω εξετάζουμε την αλληλεπίδραση των προαναφερθέντων τύπων των δεδομένων, δηλαδή τους συνδυασμούς των τύπων των δεδομένων στην Δημόσια Υγεία και την χρήση που έχουν.

2.3.7 Κλινικά δεδομένα

Οι Chimmula και Zhang (2020) χρησιμοποίησαν ένα σχετικά μικρό σύνολο δεδομένων από ανθρώπους που προέρχονται από τον Καναδά, που αφορά θανάτους και ασθενείς που ανάρρωσαν με σκοπό να μοντελοποιήσουν την μετάδοση του ιού, επιτυγχάνοντας ακρίβεια 93,4% για βραχυπρόθεσμες προβλέψεις (αν και οι ερευνητές διαπίστωσαν ότι η πανδημία θα συνεχιζόταν στον Καναδά μέχρι τον Ιούνιο του 2020, κάτι που δεν ήταν η περίπτωση, γεγονός που υποδηλώνει ότι οι μακροπρόθεσμες προβλέψεις δεν είναι έγκυρες- ένα μεγάλο σύνολο δεδομένων μπορεί ενδεχομένως να αντιπροσωπεύει πιο ακριβείς πληροφορίες).

2.3.8 Omics και Κλινικά δεδομένα

Η πρόληψη αποτελεί μείζονα στόχο της δημόσιας υγείας. Ένα ουσιαστικό βήμα για να επιτύχουμε την πρόληψη είναι ο εντοπισμός των παραγόντων κινδύνου για τις ασθένειες, οι οποίοι εκτείνονται από γονιδιωματικούς παράγοντες έως τη σωματική δραστηριότητα, το κάπνισμα και την ρύπανση. Οι προηγμένες αναλυτικές τεχνικές μπορούν να βοηθήσουν στον εντοπισμό των παραγόντων κινδύνου και των συσχετίσεων μεταξύ των μεταβλητών. Ένα παράδειγμα περιλαμβάνει κρατικά προγράμματα δημόσιας υγείας στην ΗΠΑ, τα οποία ελέγχουν περισσότερα από 4 εκατομμύρια νεογέννητα ετησίως για την ανίχνευση γενετικών ή μεταβολικών παθήσεων (Velmovitsky *et al.*, 2021).

2.3.9 Κλινικά και Κοινωνικά δεδομένα

Έχει παρατηρηθεί ότι τα δεδομένα αναζήτησης στο Google έχουν χρησιμοποιηθεί για τον εντοπισμό αυτοκτονικού ιδεασμού, ενώ τα δεδομένα του Twitter μπορούν να χρησιμοποιηθούν για την αξιολόγηση της αϋπνίας σε ένα πληθυσμιακό επίπεδο. Ένα άλλο παράδειγμα χρήσης πολλαπλών δεδομένων για τη δημόσια υγεία είναι η στόχευση αποτελεσματικών παρεμβάσεων σε συγκεκριμένες υποομάδες του πληθυσμού, όπως έχει προταθεί από τους Barret *et al.* (2013). Ακόμη, κατά τη διάρκεια της πανδημίας COVID-19, είδαμε αρκετές δημοσιεύσεις που χρησιμοποιούν προηγμένη ανάλυση για να προσπαθήσουν να προβλέψουν την εξάπλωση του COVID-19 σε μια συγκεκριμένη περιοχή (Velmovitsky *et al.*, 2021). Για παράδειγμα, οι Ayyoubzadeh *et al.* (2020) ανέλυσαν ένα σύνολο δεδομένων που αποτελούνταν από την ημερήσια επίπτωση των COVID-19 στο Ιράν και τα Google Trends δεδομένα, αναζητώντας συγκεκριμένα για όρους αναζήτησης που εμφανίζονται σε ερωτήματα όπως "corona", "πώληση αντισηπτικού" και "απολυμαντικό χεριών" (ως παράδειγμα ενσωμάτωσης των κλινικών και κοινωνικών δεδομένων). Παρόλο που δεν αναφέρθηκε η ακρίβεια, το μέσο τετραγωνικό σφάλμα του μοντέλου ήταν 27,187. Οι ερευνητές διαπίστωσαν ενδείξεις υπερπροσαρμογής στο μοντέλο λόγω του περιορισμένου όγκου των training data και πρότειναν ότι περισσότερα

δεδομένα θα οδηγούσαν σε καλύτερη ακρίβεια, υποδηλώνοντας και πάλι τις δυνατότητες των προηγμένων αναλύσεων σε συνδυασμό με μεγάλο όγκο δεδομένων. Επίσης, οι Qin *et al.* (2020) μπόρεσαν να προβλέψουν επιτυχώς τα κρούσματα COVID-19 χρησιμοποιώντας ευρετήρια αναζήτησης από την Baidu, που αποτελούν τις πιο δημοφιλείς μηχανές αναζήτησης της Κίνας. Αναμφισβήτητα, οι περιπτώσεις αυτές θα μπορούσαν να χαρακτηριστούν ως περιπτώσεις δημόσιας υγείας λόγω της χρήσης μεγάλων δεδομένων και της ακριβέστερης κατανόησης για τη συμπεριφορά των ατόμων. Αποφασίστηκε να συμπεριληφθούν σε αυτήν την ενότητα καθώς στοχεύουν σε ολόκληρους πληθυσμούς σε αντίθεση με συγκεκριμένες ομάδες ή περιοχές και επομένως μπορεί να μην αντιπροσωπεύουν μια πραγματική προσέγγιση (Velmovitsky *et al.*, 2021).

2.3.10 Κλινικά, Κοινωνικά και Περιβαλλοντικά δεδομένα

Οι Ram *et al.* (2015) χρησιμοποίησαν δεδομένα από μέσα κοινωνικής δικτύωσης (δεδομένα Twitter και Google αναζητήσεις) σε συνδυασμό με περιβαλλοντικά δεδομένα (δεδομένα της ποιότητα του αέρα) για να προβλέψουν με ακρίβεια κλινικά δεδομένα (ημερήσιες επισκέψεις στα επείγοντα περιστατικά για οξύ άσθμα).

2.3.11 Omics, Κλινικά, Κοινωνικά, PGHD και Περιβαλλοντικά δεδομένα

Οι Velmovitsky *et al.* (2021) αναφέρουν ότι η δημόσια υγεία έχει επαναπροσδιοριστεί την τελευταία δεκαετία, ενσωματώνοντας προβληματισμούς και ανησυχίες που υπερβαίνουν τις μεταδοτικές ασθένειες. Πολλοί πληθυσμοί πάσχουν, για παράδειγμα, από ενδημική παχυσαρκία και η συγκεκριμένη ασθένεια έχει αποτελέσει σημαντική πρόκληση για τους οργανισμούς της δημόσιας υγείας με σκοπό να σχεδιάσουν στρατηγικές για την καταπολέμησή της. Η παχυσαρκία είναι γνωστή ως μία πολυπαραγοντική ασθένεια που προκαλείται από γενετικούς, κοινωνικούς και περιβαλλοντικούς παράγοντες. Εάν τα δεδομένα από EHR και smart συσκευές (π.χ. Fitbit, Apple Watch) θα μπορούσαν να συνδυαστούν με κοινωνικά και περιβαλλοντικά δεδομένα, τότε θα ήταν εφικτό να μελετηθούν διάφοροι παράγοντες που σχετίζονται με τη σωματική δραστηριότητα (ή την έλλειψή της) και την παχυσαρκία σε έναν πληθυσμό. Για παράδειγμα, η δυνατότητα περιπάτου σε μια πόλη και η ποιότητα του περιβάλλοντος (π.χ, ατμοσφαιρική ρύπανση, ποσοστά εγκληματικότητας, διαθεσιμότητα μέσων μαζικής μεταφοράς, κυκλοφορία, κ.λπ.) μπορούν να επηρεάσουν το ποσοστό άσκησης που κάνει ένα άτομο. Επιπλέον, η παχυσαρκία στο κοινωνικό δίκτυο ενός ατόμου μπορεί να αποτελέσει προγνωστικό παράγοντα του BMI του ατόμου (π.χ. αν οι φίλοι του δεν βγαίνουν έξω και ασκούνται, ζώντας κυρίως καθιστική ζωή, το άτομο μπορεί να κάνει το ίδιο).

2.4 Στοχευμένες παρεμβάσεις-Προσωποποιημένοι παράγοντες κινδύνου

Σύμφωνα με τους Velmovitsky *et al.* (2021), μία από τις δυνατότητες των Big Data στην υγειονομική περίθαλψη είναι η εύρεση νέων παραγόντων κινδύνου για διάφορες ασθένειες. Το συγκεκριμένο ζήτημα είναι χρησιμό και αφορά την δημόσια υγεία, διότι η ανακάλυψη των παραγόντων κινδύνου μπορεί να βοηθήσει στην ανάπτυξη νέων στρατηγικών με στόχο την πρόληψη των ασθενειών και την αλλαγή της συμπεριφοράς στους πληθυσμούς. Μαζικά σύνολα δεδομένων από ετερογενείς πηγές και συμπεριλαμβανομένων των τύπων των δεδομένων που περιγράφηκαν παραπάνω, επιτρέπουν αναλύσεις σε πληθυσμιακό, υποπληθυσμιακό και προσωπικό επίπεδο, οι οποίες μπορούν να οδηγήσουν στην ανακάλυψη εξατομικευμένων παραγόντων κινδύνου που λαμβάνουν υπόψη ατομικές μεταβλητές και χαρακτηριστικά.

Χωρίς τα Big Data, θα ήταν εξαιρετικά δύσκολο να εφαρμοστεί στρωματοποίηση και να εντοποιστούν αυτοί οι παράγοντες κινδύνου. Για παράδειγμα, ορισμένοι παράγοντες μπορεί να είναι ωφέλιμοι για ορισμένους ανθρώπους αλλά επιβλαβείς για άλλους, οπότε η συνολική επίδραση που θα μετρηθεί σε αυτόν τον πληθυσμό θα είναι μηδενική. Μόνο με την συλλογή μεγάλου όγκου και μεγάλης ποικιλίας δεδομένων, θα μπορέσουμε να υπολογίσουμε τέτοιες στατιστικές αλληλεπιδράσεις. Με άλλα λόγια, μόνο με δεδομένα σε πληθυσμιακό επίπεδο είναι δυνατόν να εντοπιστούν σχέσεις που θα είναι επωφελείς για μία ακριβής προσέγγιση. Ακόμη, μία ενδιαφέρουσα περίπτωση των περιβαλλοντικών δεδομένων για την ανίχνευση εξατομικευμένων παραγόντων κινδύνου, είναι η χρήση των Big Data για την πρόβλεψη περιβαλλοντικών παραγόντων, όπως η συγκέντρωση συγκεκριμένων σωματιδίων στον αέρα. Επιπλέον, ο εντοπισμός των περιβαλλοντικών παραγόντων κινδύνου μπορεί να βοηθήσει στον προσδιορισμό του ποιος διατρέχει μεγαλύτερο κίνδυνο έκθεσης (π.χ. μέλλουσες μητέρες), επιτρέποντας την εξατομικευμένη θεραπεία, καθώς και να οδηγήσει στην πραγματοποίηση στοχευμένων παρεμβάσεων με στόχο την βελτίωση της υγείας του πληθυσμού (π.χ. εφαρμογή πολιτικών σε κάποιο αστικό περιβάλλον). Η ανάλυση και οι γνώσεις που αποκτώνται από τη χρήση των Big Data για την ανάλυση και την ανίχνευση περιβαλλοντικών παραγόντων μπορεί να βοηθήσει τόσο τα άτομα σε εξατομικευμένο επίπεδο (όπως η ανίχνευση των επιπέδων ρύπανσης στη γειτονιά τους) όσο και να υποστηρίξει στοχευμένες παρεμβάσεις σε επίπεδο πληθυσμού (π.χ. μια ολόκληρη πόλη).

Παραπάνω συζητήθηκε πως οι αναλυτικές μέθοδοι που εφαρμόζονται σε PGHD και κοινωνικά δεδομένα (και ενδεχομένως σε συνδυασμό με άλλους τύπους δεδομένων, όπως τα κλινικά δεδομένα) χρησιμοποιούνται για την γρήγορη ανάλυση των δεδομένων, αποκτώντας εικόνα της συμπεριφοράς των ατόμων. Αυτές οι γνώσεις μπορούν να κοινοποιηθούν πίσω στον χρήστη (για παράδειγμα, μέσω τηλεφωνικών ειδοποιήσεων) δημιουργώντας έναν «βρόχο ανάδρασης», δηλαδή με την συλλογή και την ανάλυση των δεδομένων, οι γνώσεις και οι πληροφορίες που αποκτώνται από τα δεδομένα σχετικά με την υγεία των χρηστών μεταδίδονται πίσω σε αυτούς και ενδεχομένως ακόμη να υπάρχουν συστάσεις για την βελτίωση της υγείας τους. Αυτό μπορεί να βελτιώσει σημαντικά την υγεία των ασθενών, παρέχοντας σε αυτούς προτάσεις για την βελτίωση της υγείας τους με βάση τις ατομικές συμπεριφορές του. Οι βρόχοι

αυτοί αποτελούν μία περίπτωση χρήσης των Big Data για την Ιατρική. Ωστόσο, μπορούν επίσης να θεωρηθούν ως μια μορφή στοχευμένης παρέμβασης στη δημόσια υγεία, καθώς εφαρμόζονται σε ένα μεγάλο ποσοστό του πληθυσμού με πολύ συγκεκριμένο και εξατομικευμένο τρόπο. Στο παρελθόν, οι παρεμβάσεις για τη βελτίωση της έκθεσης σε παράγοντες κινδύνου διαφόρων ασθενειών, σήμαινε ότι οι γιατροί έδιναν συστάσεις στους ασθενείς (π.χ. να σταματήσουν το κάπνισμα ή το ποτό, να είναι σωματικά δραστήριοι). Με τους πραγματικούς χρόνου βρόχους ανάδρασης που ενεργοποιούνται από smart συσκευές που συλλέγουν και μεταδίδουν δεδομένα, καθώς και με την εφαρμογή προηγμένων αναλυτικών μεθόδων, συμπεριλαμβανομένων των μεθόδων τεχνητής νοημοσύνης, οι παρεμβάσεις αυτές μπορούν να φτάσουν σε πολύ μεγαλύτερο δείγμα του πληθυσμού σε ελάχιστο έως μηδενικό χρόνο. Επιπλέον, αυτές οι παρεμβάσεις μπορούν να θεωρηθούν ακριβείς, καθώς βασίζονται σε δεδομένα που συλλέγονται σε ατομικό επίπεδο. Ως εκ τούτου, οι ίδιοι τύποι δεδομένων και χρήσεις αυτών για την ιατρική, μπορούν να είναι χρήσιμες για παρεμβάσεις στη δημόσια υγεία. Οι Jain *et al.* (2015) υπογραμμίζουν ότι τα φορητά επιταχυνσιόμετρα μπορούν να παρακολουθούν τα λειτουργικά αποτελέσματα σε άτομα με νευρολογικές διαταραχές και να παρέχουν δεδομένα για τον σχεδιασμό προγραμμάτων αποκατάστασης. Αυτό μπορεί να θεωρηθεί ως ένα παράδειγμα της δημόσιας υγείας, καθώς οι πληροφορίες που περιέχονται σε βρόχους ανάδρασης μπορούν να παρέχουν παρεμβάσεις σε μια υποομάδα του πληθυσμού (Velmovitsky *et al.*, 2021).

2.5 Πρόβλεψη κινδύνου

Για να μπορέσει η ιατρική να έχει ακρίβεια πρέπει να βασίζεται σε τεκμηριωμένα, πλούσια και ποικίλα δεδομένα που συγκεντρώνονται από διάφορες πηγές. Αυτό σημαίνει ότι πρέπει να ξεπεράσουμε τα παραδοσιακά όρια του τι θεωρούνται δεδομένα υγείας (omics και κλινικά δεδομένα) με σκοπό να συγκεντρωθούν και να ενσωματωθούν κοινωνικοί και περιβαλλοντικοί παράγοντες της υγείας. Η ενσωμάτωση πλήθους δεδομένων από διαφορετικές πηγές θα οδηγήσει στην καλύτερη κατανόηση του τι καθιστά μία θεραπεία αποτελεσματική για ένα άτομο ή μία ομάδα ατόμων. Αυτό θα φέρει επίσης την ιατρική πιο κοντά στη δημόσια υγεία, καθώς η κοινωνικές και οι περιβαλλοντικές μεταβλητές συνήθως συλλέγονται, αναλύονται και χρησιμοποιούνται σε επίπεδο πληθυσμού (Velmovitsky *et al.*, 2021). Για παράδειγμα, οι Yu *et al.* (2019) για να εκτιμήσουν με ακρίβεια την έκθεση σε αιωρούμενα σωματίδια στον αέρα χρησιμοποίησαν δεδομένα του Google Maps. Ερευνητές όπως οι Prosperi *et al.* (2018) έθεσαν την πρόληψη ασθενειών ως έναν από τους στόχους για την ακρίβεια της ιατρικής. Κάποιος μπορεί να ισχυριστεί ότι η ιατρική ασχολείται με τα συγκεκριμένα χαρακτηριστικά ενός ατόμου προκειμένου να προστατεύσει το άτομο αυτό από τις ασθένειες, αλλά όταν αυτή η γνώση εφαρμόζεται σε μια κοινότητα ατόμων με παρόμοια χαρακτηριστικά, η ιατρική συγκλίνει στην δημόσια υγεία. Στο ίδιο πνεύμα, η παραδοσιακή δημόσια ανάλυση της δημόσιας υγείας εξέταζε συνήθως δεδομένα σε επίπεδο πληθυσμού, συμπεριλαμβανομένων κοινωνικών, περιβαλλοντικών και κλινικών δεδομένων. Συμπεριλαμβανομένου περισσότερων δεδομένων για κάθε άτομο, όπως τα omics, μπορεί να

βελτιώσει τις προσπάθειες πρόληψης, καθώς μπορούν να στοχεύσουν σε πιο συγκεκριμένες υποομάδες ενός πληθυσμού, καθώς και να μπορέσει να βοηθήσει τους ερευνητές να κατανοήσουν καλύτερα τα χαρακτηριστικά των ασθενειών. Επιπροσθέτως, τα μοντέλα που δημιουργούνται με τη χρήση αυτής της μεγάλης ποικιλίας πληροφοριών μπορούν να χρησιμοποιηθούν για την πρόβλεψη κινδύνων (Velmovitsky *et al.,* 2021). Όπως περιγράφεται από τον Dolley (2018), οι πρώτες προσπάθειες για τη χρήση Big Data για σκοπούς της δημόσιας υγείας με τα δεδομένα του Google Flu Trends κατέληξαν σε αποτυχία, οδηγώντας τους ερευνητές να ενσωματώσουν περισσότερα πηγές δεδομένων.

2.6 Παρακολούθηση των ασθενειών

Επίσης, οι νέες πηγές και οι τύποι δεδομένων μπορούν να οδηγήσουν σε αυξημένη παρακολούθηση και επιτήρηση των ασθενειών. Με την σύνδεση νέων τεχνολογιών, οι ερευνητές μπορούν να μελετήσουν τις κινήσεις, τα μοτίβα και τις συμπεριφορές ατόμων και πληθυσμών, παρακολουθώντας τα προσβεβλημένα από κάποια ασθένεια άτομα. Η βασική διαφορά μεταξύ των παραδοσιακών προσεγγίσεων της παρακολούθησης και των Big Data είναι ότι τα Big Data μπορούν να οδηγήσουν σε επιτήρηση ασθενειών σε σχεδόν πραγματικό χρόνο. Όπως προτείνει ο Dolley (2018), "η πρόσβαση σε τεράστιους όγκους δεδομένων σε πραγματικό χρόνο που παράγονται από ανθρώπους, φαίνεται ταυτόχρονα να αποτελεί ιδανική αποθήκη σημάτων για τον εντοπισμό και την παρακολούθηση των προσβεβλημένων ατόμων". Οι Ginsberg *et al.* (2009) χρησιμοποιούν κλινικά δεδομένα (τον αριθμό των γιατρών σε μια περιοχή), και κοινωνικά δεδομένα (ερωτήματα αναζήτησης) για να εκτιμήσουν την πιθανότητα ότι μια επίσκεψη σε γιατρό στην εν λόγω περιοχή σχετίζεται με τη γρίπη. Ένα ενδιαφέρον παράδειγμα που αξιοποιείται μια ακριβής προσέγγιση στη δημόσια υγεία παρουσιάζεται από τους Arora et al. (2020), όπου χώρισαν την Ινδία σε διάφορες περιοχές ανάλογα με τη σοβαρότητα, μέτρια και ήπια ανάλογα με τον αριθμό των κρουσμάτων COVID-19 και δημιούργησαν ξεχωριστά μοντέλα για κάθε ένα από τα κρατίδια και τα ενωτικά εδάφη. Σε αντίθεση με την πρόβλεψη των κρουσμάτων του COVID-19 που περιγράφονται παραπάνω , η έρευνα αυτή έχει μία ακριβής προσέγγιση για τη δημόσια υγεία που στοχεύει σε περιοχές εντός ενός μεγαλύτερου πληθυσμού ή μιας περιοχής. Σημαντικό είναι να αναφερθεί ότι η παρακολούθηση των ασθενειών καθώς και η διαχείρηση οποιουδήποτε τύπου δεδομένων των ασθενών, πρέπει να αξιολογούνται όσον αφορά την προστασία της ιδιωτικότητας των δεδομένων των ασθενών και την ανωνυμία τους (Velmovitsky *et al.,* 2021) .

2.7 Γενετική επιδημιολογία

Η ευρεία διαθεσιμότητα των μεθόδων αλληλούχισης σε συνδυασμό με τη δραστική μείωση του κόστους ήταν σε μεγάλο βαθμό υπεύθυνες για την άνοδο και την εξέλιξη της Ιατρικής. Σήμερα, η αλληλούχιση σε όλο το γονιδίωμα μπορεί να κοστίσει περίπου 1.000 δολάρια, από σχεδόν 98 εκατομμύρια δολάρια που κόστιζε το 2001 (Whole Genome Sequencing). Επιπλέον, αρκετές εμπορικές εταιρείες προσφέρουν υπηρεσίες που παρέχουν μερική ανάλυση της αλληλουχίας του γονιδιώματος για λίγο πάνω από 100 δολάρια, μαζί με χαρτογράφηση δημογραφικών χαρακτηριστικών και συγκεκριμένων ασθενειών (αλλά με μη αποδεδειγμένη κλινική χρησιμότητα), κάτι το οποίο εγείρει ανησυχίες που σχετίζονται με την προστασία της ιδιωτικότητας των πληροφοριών υγείας. Παρόλο που ένα γονιδίωμα παραμένει σχετικά αμετάβλητο κατά τη διάρκεια της ζωής, ένας γονιδιωματικός έλεγχος που λαμβάνεται κατά τη γέννηση ή πριν από τη γέννηση θα είναι βέλτιστος για την ακριβέστερη πρόβλεψη μίας νόσου. Παρά τον αρχικό ενθουσιασμό στην Ιατρική για τη γενετική, τα αποτελέσματα ήταν απογοητευτικά και δεν απέδωσαν όπως πίστευαν μέχρι στιγμής. Η προβλεπτική ικανότητα και η επακόλουθη κλινική χρησιμότητα της εκτίμησης του κινδύνου από τις γενετικές παραλλαγές έχει βρεθεί ότι είναι μέτρια για πολλές ασθένειες και οι μελέτες συσχέτισης σε επίπεδο γονιδιώματος (Genome-Wide Association Studies - GWAS) δεν έχουν οδηγήσει στην κατανόηση των γενετικών μηχανισμών που διέπουν την ανάπτυξη πολλών ασθενειών. Μεταξύ των μειωνεκτημάτων των GWAS, ένα από αυτά είναι το πρόβλημα της κληρονομικότητας. Η κληρονομικότητα είναι ένα μέτρο του ποσοστού της φαινοτυπικής διακύμανσης μεταξύ των ανθρώπων που εξηγείται από τη γενετική ποικιλομορφία, για την οποία οι μεμονωμένες γενετικές παραλλαγές δεν μπορούν να εξηγήσουν μεγάλο μέρος της κληρονομικότητας των ασθενειών, συμπεριφορών και άλλων φαινοτύπων. Ένας άλλος περιορισμός των GWAS αφορά τη μελέτη ενός μόνο φαινότυπου ή αποτελέσματος (συχνά ανακριβής) και ο υπολογισμός για ετερογενείς φαινότυπους θα απαιτούσε τεράστιες μελέτες σε μέγεθος. Άλλα ζητήματα των GWAS περιλαμβάνουν το σχεδιασμό, την ισχύ, την αποτυχία αναπαραγωγής και τους στατιστικούς περιορισμούς. Στην πράξη, εφαρμόζεται μόνο η μονομεταβλητή και η πολυμεταβλητή γραμμική παλινδρόμηση. Η εξέταση των αλληλεπιδράσεων γονιδίου-γονιδίου και συμπεριλαμβανομένου και άλλων μεταβλητών αντί των βασικών δημογραφικών ή κλινικών χαρακτηριστικών, σπάνια γίνεται και συνήθως είναι υπολογιστικά δύσκολη. Υπάρχουν πολύ λίγα παραδείγματα κοινών ασθενειών με υψηλές επιδράσεις γενετικών παραλλαγών που επηρεάζουν ασθένειες υψηλής συχνότητας και οι κοινές γενετικές παραλλαγές έχουν συνήθως χαμηλή προγνωστική ικανότητα. Όσο πιο σπάνια είναι μια γενετική παραλλαγή, τόσο πιο δύσκολο είναι να υπάρχει εξακρίβωση του μεγέθου της επίδρασης. Υπάρχουν σπάνιες αλληλόμορφες υψηλής επίδρασης γενετικές παραλλαγές που προκαλούν Μεντελικές ασθένειες και μια πληθώρα παραλλαγών χαμηλής συχνότητας με μέτριες επιδράσεις. Οι σπάνιες παραλλαγές χαμηλής επίδρασης είναι πολύ δύσκολο να βρεθούν και μπορεί να είναι κλινικά άσχετες, εκτός εάν εμπλέκονται σε περισσότερο πολύπλοκα μονοπάτια. Στην πραγματικότητα, είναι γνωστό ότι τα γονιδιακά ζεύγη έκφρασης μπορούν να συσχετίζονται από κοινού με μια φαινότυπη ασθένεια και οι αλληλεπιδράσεις υψηλότερης

τάξης πιθανόν να παίζουν έναν ρόλο. Λίγοι αλγόριθμοι έχουν προταθεί για την αναζήτηση γονιδίων που εκφράζονται από κοινού και οι υπάρχουσες μέθοδοι είναι υπολογιστικά αναποτελεσματικές. Ωστόσο, αυτά τα ζητήματα ευθύνονται μερικώς που η ιατρική δεν έχει ακόμη εκπληρώσει τις αρχικές της υποσχέσεις. Πράγματι, για να είναι έγκυρα και αποτελεσματικά τα μοντέλα της δημόσιας υγείας, η ενσωμάτωση και η δοκιμή παραγόντων πέραν των γενετικών είναι καθοριστικής σημασίας.

Ενώ η γενετική παραμένει ως επί το πλείστον στατική με την πάροδο του χρόνου, άλλοι σχετιζόμενοι με την υγεία παράγοντες μεταβάλλονται συνεχώς και πρέπει να αξιολογούνται περιοδικά. Η επιγενετική, π.χ. τα δεδομένα μεθυλίωσης, τα οποία έχουν μια χρονική συνιστώσα, μπορούν να συμβάλουν σε ένα σχετικό τμήμα της ανεξήγητης κληρονομικότητας. Φθηνότερη και ταχύτερη παραγωγή ακολουθιακών δεδομένων με την επόμενη γενιά τεχνολογιών αλληλούχισης έχει ανοίξει την μεταγενέστερη εποχή των GWAS, επιτρέποντας έναν εντελώς νέο κόσμο των omics. Η επανάσταση των GWAS, και αναμφισβήτητα ο κορεσμός, έφερε μια πληθώρα από επιγονιδιωματικές συσχετίσεις σε επίπεδο μεθυλίωσης, μεταγραφικό, μικροβιακό και μελέτες συσχέτισης σε επίπεδο περιβάλλοντος. Ενδιαφέρον παρουσιάζει το γεγονός ότι οι μελέτες συσχέτισης σε επίπεδο φαινομένου αντιστρέφουν τον κανόνα, καθώς όλες οι καταστάσεις υγείας που εντοπίζονται στο ιατρικό ιστορικό χρησιμοποιούνται ως μεταβλητές και συσχετίζονται με μεμονωμένα γενετικά χαρακτηριστικά. Τα genomics, τα transcriptomics, τα metabolomics και όλες τα άλλα -omics μπορούν να θεωρηθούν ως τομείς εισόδου σε ένα μοντέλο πρόβλεψης. Η συγχώνευση δύο ή περισσοτέρων συσχετισμένων μελετών είναι το επόμενο βήμα προς έναν καλύτερο χαρακτηρισμό των μηχανισμών και των κινδύνων της νόσου. Ωστόσο, προκύπτουν προκλήσεις για την μοντελοποίηση και τον υπολογισμό με την ενοποίηση πολλών τομέων, λόγω των αυξημένων διαστάσεων, της μεταβλητής ετερογένειας, της σύγχυσης και της αιτιότητας. Οι τυποποιήσεις της συσχέτισης των μελετών σε διατομεακή κλίμακα, υπό τον γενικό όρο των multiomics, έχουν προταθεί. Παρά τις φθηνότερη και ταχύτερη παραγωγή ακολουθιακών δεδομένων, οι περισσότερες από τις μελέτες των multiomics περιορίζονται από μικρά δείγματα. Γενικά, όσο περισσότερο ετερογενή είναι τα πειραματικά δεδομένα που θα παραχθούν ή οι πηγές δεδομένων που πρέπει να συμπεριληφθούν στη μελέτη, τόσο πιο δύσκολο είναι να επιτευχθεί μεγαλύτερο μέγεθος δείγματος. Η πιο ενδιαφέρουσα χρησιμότητα των multiomics, αντί για πρόβλεψη των αποτελεσμάτων της υγείας, είναι η "μη επιβλεπόμενη" ανάλυσή τους, δηλαδή ο εντοπισμός προτύπων/ενδότυπων που μπορούν να βοηθήσουν στην αποκάλυψη βιολογικών παραγόντων και τελικά να επαναπροσδιορίσουν φάσματα και φαινότυπους ασθενειών. Ωστόσο, υπάρχουν αυξανόμενες ενδείξεις ότι για να διασφαλιστεί η ακρίβεια στην δημόσια υγεία για να τηρούν τις υποσχέσεις τους σε όλο το φάσμα της περίθαλψης, πρέπει να προχωρήσουμε πέρα από τα -omics.

Στην εποχή μας, οι μελέτες πολλαπλών τομέων πρέπει να επεκταθούν πέρα από τα "omics" δεδομένα και να εξετάσουν άλλους τομείς της ζωής ενός ατόμου. Συγκεκριμένα, τη γενετική, τη συμπεριφορά, τους κοινωνικούς, τους περιβαλλοντικούς και τους κλινικούς τομείς της ζωής. Θεωρείται ότι αυτοί είναι οι πέντε τομείς που επηρεάζουν την υγεία. Επιπλέον, η πρόσβαση στο διαδίκτυο και η ευρέως διαδεδομένη διαθεσιμότητα και χρήση των smartphone τεχνολογιών υποδηλώνουν ότι ο κλινικός τομέας μπορεί να να ενισχυθεί σημαντικά με

δεδομένα που δημιουργούνται από τους ασθενείς, όπως δεδομένα σωματικής δραστηριότητας, διατροφής, γλυκόζης αίματος, αρτηριακής πίεσης και άλλες παρόμοιες μεταβλητές που μπορούν να συλλέγονται απρόσκοπτα με τη χρήση smartphones και φορητών συσκευών. Επιπλέον, τα εν λόγω εργαλεία, σε συνδυασμό με τις πλατφόρμες κοινωνικών δικτύων παρέχουν ένα παράθυρο στη συμπεριφορά και τους κοινωνικούς τομείς της υγείας, που είναι περιβάλλοντα πλούσια σε δεδομένα και πρέπει να ληφθούν υπόψη στο πλαίσιο της δημόσιας υγείας, ώστε να δημιουργηθεί ένας "ψηφιακός φαινότυπος" της νόσου. Για παράδειγμα, έχουν χρησιμοποιηθεί εικόνες από το Instagram για την εξακρίβωση των διατροφικών συνηθειών αντί για ημερολόγιο διατροφής ή ερωτηματολόγια διατροφικής πρόσληψης, τα οποία μπορεί να είναι ανακριβή, πολύπλοκα και χρονοβόρα. Επίσης, το Instagram, έχει χρησιμοποιηθεί για τον εντοπισμό προγνωστικών δεικτών κατάθλιψης. Ακόμη, τα δεδομένα που συλλέγονται από το Twitter μπορούν να χρησιμοποιηθούν για τον χαρακτηρισμό και την πρόβλεψη τύπων αϋπνίας. Η έρευνα στον περιβαλλοντικό τομέα έχει δείξει ότι το περιβάλλον στο οποίο ζούμε επηρεάζει την υγεία και τη θνησιμότητά μας. Ωστόσο, η έρευνα που χρησιμοποιεί μη παραδοσιακά δεδομένα που σχετίζονται με την υγεία από αυτούς τους τομείς, έχουν διεξαχθεί με κάποια επιτυχία καθώς και με κάποια αμφισβήτηση (Prosperi *et al*., 2018).

2.8 Το avatar Υγείας

Το avatar υγείας είναι μια εικονική αναπαράσταση ενός ατόμου με όλες τις σχετικές πληροφορίες για την υγεία του (Σχήμα 2.5 - Prosperi *et al*., 2018) που περιέχει έξυπνους τρόπους για τη διαχείριση και την πρόβλεψη της μελλοντικής κατάστασης της υγείας του. Ακόμη και με την ευρεία χρήση των ηλεκτρονικών αρχείων υγείας (EHR) και των ολοκληρωμένων αποθηκευτικών χώρων για τα δεδομένα, τα άτομα είναι γενικά αποκομμένα από τις πληροφορίες για την υγεία τους και οι ευκαιρίες για ενεργή συμμετοχή στην έρευνα παραμένουν περιορισμένες, παρά τις υπάρχουσες πρωτοβουλίες όπως το HealthKit της Apple. Γνωστά εμπόδια στη σύνδεση και την αποτελεσματική αξιοποίηση των πληροφοριών υγείας σε διαφορετικούς διαδικτυακούς τόπους επιβραδύνουν την έρευνα στον τομέα της υγειονομικής περίθαλψης και την ανάπτυξη εξατομικευμένης περίθαλψης. Περαιτέρω, τα EHR δεν ενσωματώνονται μεταφραστικά με διαγνωστικά ή θεραπευτικά εργαλεία βελτιστοποίησης. Ένας γιατρός μπορεί να λάβει και να μεταφέρει εργαστηριακά αποτελέσματα σε απευθείας σύνδεση, αλλά στη συνέχεια οι διαγνώσεις γίνονται συχνά με παραδοσιακό τρόπο, με βάση τα δεδομένα του μέσου όρου του πληθυσμού. Ο προσωπικός φάκελος υγείας (PHR) είναι μια συγκέντρωση όλων των πληροφοριών υγείας από διάφορους παρόχους υγειονομικής περίθαλψης ή άλλες πηγές που είναι αποθηκευμένες στο cloud, που είναι άμεσα προσβάσιμες και στην ιδιοκτησία του κάθε ατόμου. Ο PHR είναι συμπληρωματικός στα EHR, ο οποίος συνήθως αποθηκεύεται στον πάροχο, με λογισμικό των προμηθευτών, όπως το Epic ή το Cerner. Ωστόσο, το avatar υγείας δεν θα πρέπει απλουστευτικά να ταυτίζεται με το PHR, καθώς το PHR είναι εγγενώς παθητικό, με μικρή συμμετοχή από τον ασθενή. Προτείνεται ένα μοντέλο του σύγχρονου avatar υγείας όπως θα πρέπει να είναι σε μια εποχή μεγάλων συνόλων δεδομένων που δημιουργούνται από τους ασθενείς. Ένα άτομο μπορεί να δει τις πληροφορίες

υγείας του χρησιμοποιώντας ένα PHR του παρόχου, αλλά δεν μπορεί εύκολα να συγχωνεύσει τις πληροφορίες με δεδομένα από άλλους παρόχους, ούτε να ζητήσει από έναν πάροχο να ανεβάσει τα δεδομένα του εύκολα από τα EHR στο PHR κατά τη διάρκεια μίας επίσκεψη στον γιατρό, π.χ. μέσω μιας εφαρμογής για smartphone. Ένας έξυπνος αλγόριθμος που αντιστοιχεί τους ανθρώπους σε ερευνητικές μελέτες με βάση το πλήρες ιατρικό ιστορικό τους δεν υπάρχει ακόμη. Τόσο οι γιατροί όσο και οι ασθενείς που ενδιαφέρονται για τη διάγνωση με τη βοήθεια ενός υπολογιστή, πρέπει συνήθως να ανεβάζουν πληροφορίες σε μία τρίτη υπηρεσία (π.χ. για να αναλύσουν την ευαισθησία στα αντιβιοτικά). Τέλος, οι κοινοποιήσεις δεδομένων είναι δύσκολες, όχι μόνο από την άποψη των βημάτων που απαιτούνται για τον σεβασμό των αρχών δεοντολογίας, της πρακτικής και της προστασίας των ατόμων, τα οποία είναι απαραίτητα και θα μπορούσαν να εκσυγχρονιστούν, αλλά και επειδή τα μόνα δεδομένα που θεωρούνται αξιόπιστα είναι εκείνα που προέρχονται από τα EHR. Αυτό σημαίνει ότι οι μεγάλες ανταλλαγές δεδομένων συμβαίνουν αποκλειστικά και μόνο σε επίπεδο πληθυσμού μέσω θεσμικών ή εταιρικών συνδέσμων. Ακόμη, η έρευνα και ανάλυση που ακολουθούν δεν είναι εκσυγχρονισμένες, τα πολυαναμενόμενα ερευνητικά αντικείμενα, που είναι πλούσιες συσσωρεύσεις πόρων που συγκεντρώνουν δεδομένα, μεθόδους και ανθρώπους σε επιστημονικές έρευνες, αλλά βρίσκονται ακόμη σε νηπιακό στάδιο. Η ενσωμάτωση διαφορετικών τύπων και πηγών δεδομένων θα πρέπει να διατηρεί το αρχικό πλαίσιο και νόημα, ενώ χαρτογραφώντας με νόημα τις σχέσεις τους με άλλες μεταβλητές που σχετίζονται με την υγεία, θα πρέπει να είναι ευέλικτη και ολοκληρωμένη.

Σχήμα 2.5
Health Avatar

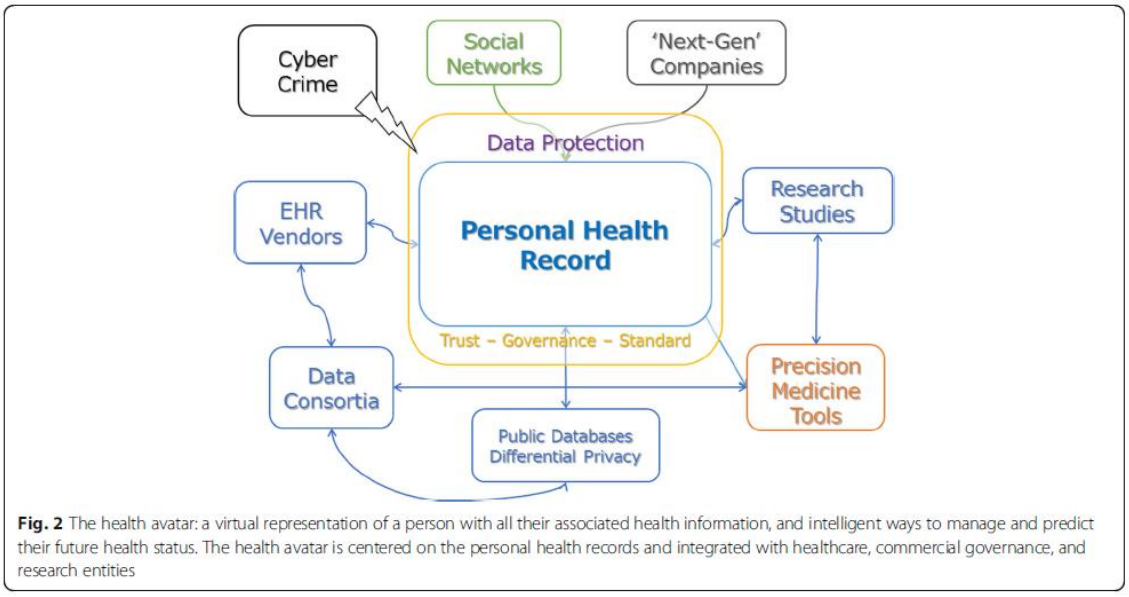


Fig. 2 The health avatar: a virtual representation of a person with all their associated health information, and intelligent ways to manage and predict their future health status. The health avatar is centered on the personal health records and integrated with healthcare, commercial governance, and research entities

Η ενσωμάτωση φυσικών δεδομένων των EHR απαιτεί τεράστιες προσπάθειες και πόρους, αλλά επί του παρόντος είναι η πιο επιτυχημένη προσέγγιση για τη σύνδεση πληροφοριών υγείας, επειδή υποστηρίζεται από αυστηρά πρότυπα διακυβέρνησης και σταθερή υποδομή. Προσπάθειες όπως το εθνικό ασθενοκεντρικό κλινικής έρευνας δίκτυο είναι ένα

χαρακτηριστικό παράδειγμα. Η κοινή χρήση δεδομένων για την αντιστοίχιση των συμμετεχόντων στην έρευνα, ένα από τα πολυαναμενόμενα προνόμια που παρέχει το NIH (National Institutes of Health), που επιτέλους γίνεται να αξιοποιηθεί μέσω του ResearchMatch. Το avatar υγείας θα πρέπει να συνδέει όλους άλλα και νέους τύπους δεδομένων που σχετίζονται με την υγεία, τη γονιδιωματική, έως τις μυριάδες των -omics, με βάση την κινητή και την φορητή τεχνολογία και τις περιβαλλοντικές πηγές. Τα δεδομένα αυτά καταγράφουν πληροφορίες από άλλους τομείς που επηρεάζουν την υγεία πολύ περισσότερο από ότι η κλινική φροντίδα και μόνο. Τέτοιες ενσωματώσεις έχουν ήδη αρχίσει να πραγματοποιούνται σε όλο τον κόσμο με συστήματα υγειονομικής περίθαλψης όπως το Geisinger που διεξάγουν γενετική αλληλουχία και επιστρέφουν κάποια από τα αποτελέσματα στους ασθενείς και με πρωτοβουλίες όπως τα electronic Medical Records and Genomics (eMERGE) και τα Implementing Genomics in Practice (IGNITE) δίκτυα. Ωστόσο, οι προσπάθειες αυτές έχουν περιοριστεί στη γονιδιωματική.

Γενικότερα, το avatar υγείας θα πρέπει να είναι σε θέση να συνδέεται και να εκμεταλλεύεται πληροφορίες που δεν ανήκουν στα EHR που είναι δυνητικά χρήσιμα για την αξιολόγηση της υγείας, ακόμη και αν προέρχονται από εξαιρετικά αδόμητες πηγές, όπως τα μέσα κοινωνικής δικτύωσης. Ένα παράδειγμα αποτελεί η πρόσφατη συνεργασία της Epic με την Apple για να επιτρέψει στο HealthKit της Apple να εμφανίζει τα δεδομένα EHR του ασθενούς. Το App Orchard της Epic επιτρέπει επίσης τη συλλογή φορητών δεδομένων τεχνολογίας και την αποθήκευση στα EHR. Για παράδειγμα, ένα εργαλείο τεχνητής νοημοσύνης θα μπορούσε να επεξεργαστεί εικόνες από το Instagram και τις αναρτήσεις στο Facebook/Twitter ώστε να διαπιστωθούν οι διατροφικές συνήθειες. Οι πληροφορίες αυτές μπορούν στη συνέχεια να χρησιμοποιηθούν για τη συμπλήρωση ένος ερωτηματολογίου διατροφής, κωδικοποιημένο σε κάποιο είδος δομημένης πληροφορίας και να αποθηκευτούν στα EHR. Η μετακίνηση από το ατομικό επίπεδο, οι περιβαλλοντικές πληροφορίες που αφορούν το άτομο, για παράδειγμα μέσω της εξακρίβωσης του τόπου κατοικίας ή τον γεωγραφικό εντοπισμό μέσω κινητού τηλεφώνου, θα μπορούσαν επίσης να συμπληρώσουν πεδία των EHR, αποθηκεύοντας πληροφορίες όπως η έκθεση σε αλλεργιογόνα και ρύπους. Ωστόσο, η ταξινόμηση μη τυποποιημένων δεδομένων, π.χ. στιγμιαία οικολογική αξιολόγηση μέσω Twitter, Facebook ή παρακολούθηση GPS μέσω smartphone, είναι επιρρεπής σε σοβαρά προβλήματα προστασίας της ιδιωτικής ζωής και σε ζητήματα ασφάλειας. Η ολοκλήρωση δεδομένων και ακόμη περισσότερο η κοινή χρήση τους πρέπει να είναι ασφαλής ώστε να ανταποκρίνεται στη λαϊκή υποστήριξη. Υπό αυτή την έννοια, η έρευνα στην διαφορετική ιδιωτικότητα αποσκοπεί στην ανάπτυξη νέων αλγορίθμων όχι μόνο για την προστασία των ταυτοτήτων, αλλά και για τη δημιουργία συγκαλυμμένων ή συνθετικών δεδομένων που μπορούν να διαμοιραστούν δημόσια και να χρησιμοποιηθούν ελεύθερα για προκαταρκτική έρευνα. Ενώ η ιδιωτικότητα έχει διευκολύνει την κοινή χρήση δεδομένων, παραμένει πρόκληση η ασφαλής ανωνυμία των δεδομένων με την ταυτόχρονη διατήρηση όλων των πολυμεταβλητών στατιστικών ιδιοτήτων τους. Η ατομοκεντρική προσέγγιση του avatar υγείας μπορεί να διευκολύνει την αντιστοίχιση ατόμων με ερευνητικά προγράμματα, με ασαφή όρια μεταξύ της κλινικής περίθαλψης και της έρευνας, με σεβασμό στη δεοντολογία αλλά και εκσυγχρονισμό των απόψεων της συγκατάθεσης μετά από ενημέρωση. Όσον αφορά τα ενεργά

χαρακτηριστικά, δηλαδή όχι μόνο την αποθήκευση δεδομένων, το avatar υγείας θα έπρεπε να διαθέτει σύνδεση με εξατομικευμένα προγνωστικά εργαλεία για την κατάσταση της υγείας. Στο πλαίσιο των κατάλληλων δεοντολογιών και συγκαταθέσεων, μετά από ενημέρωση, το avatar υγείας θα μπορούσε να τροφοδοτεί άμεσα πληροφορίες υγείας σε ατομικό επίπεδο σε πολλαπλά ερευνητικά προγράμματα για τη δημιουργία νέων και ακριβέστερων εργαλείων της ιατρικής. Αυτό θα απαιτήσει μέτρα προστασίας της ιδιωτικής ζωής και των δεδομένων για την αποφυγή κλοπής και κατάχρησης ταυτότητας ή δεδομένων. Περαιτέρω, η ευρεία πρόσβαση σε δεδομένα που παράγονται από τους ασθενείς, μαζί με την ενσωμάτωση με κλινικές βάσεις και βάσεις δεδομένων υγείας παρέχουν μια μοναδική ευκαιρία για την επέκταση της ιατρικής σε επίπεδο πληθυσμού (Prosperi *et al*., 2018).

2.9 Μοντέλα πρόβλεψης: Ερμηνευσιμότητα έναντι της απόδοσης

Μία σημαντική πρόκληση στην χρήση των Big Data για την Δημόσια Υγεία είναι η χρήση των συναγόμενων μοντέλων, δηλαδή «τα Big Data οδηγούν σε «μεγάλα» μοντέλα». Τα «μεγάλα» μοντέλα αποτελούνται από πολλές μεταβλητές και με μη-γραμμικούς ή ιδιαίτερα πολύπλοκους τρόπους. Μάλιστα, τέτοια μοντέλα της Μηχανικής Μάθησης μπορούν εύκολα να αποδώσουν ερμηνεύσιμα αποτελέσματα ή άριστη πρόβλεψη, αλλά όχι απαραίτητα και τα δύο ταυτόχρονα. Παρά τη δυνητικά υψηλότερη ακρίβεια στην πρόβλεψη διαγνώσεων ασθενειών και αποτελεσμάτων υγείας, πολλές μέθοδοι της Μηχανικής Μάθησης συνήθως θεωρούνται αδιαφανείς για τον τελικό χρήστη και χαρακτηρίζονται ως «black-boxes». Στον αντίποδα, τα «white-boxes» μοντέλα είναι ερμηνεύσιμα από τον άνθρωπο, όπως τα scores κινδύνου ή οι διαγνωστικοί κανόνες. Αν και τα «black-box» μοντέλα είναι πιθανόν να παρέχουν έναν πολύ ακριβή υπολογισμό της πιθανότητας ενός συμβάντος ή αποτελέσματος, αντιμετωπίζονται συχνά με σκεπτικισμό λόγω της έλλειψης εξέτασης των αιτιωδών διαδρομών (Σχήμα 2.6- Prosperi *et al*., 2018).

Σχήμα 2.6
White και Black boxes

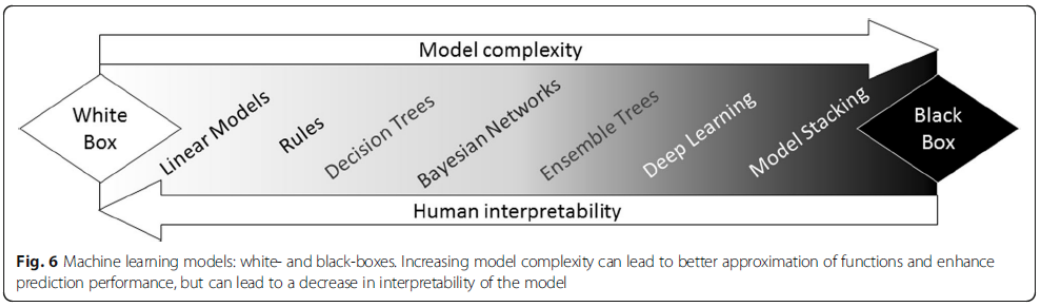


Fig. 6 Machine learning models: white- and black-boxes. Increasing model complexity can lead to better approximation of functions and enhance prediction performance, but can lead to a decrease in interpretability of the model

Ωστόσο, όταν ενσωματωθούν απρόσκοπτα στα EHR με στόχο την υποστήριξη του συστήματος σε κλινικές αποφάσεις και αν μπορούν να εντοπίσουν κλινικά εφαρμόσιμα χαρακτηριστικά, τότε μπορούν να γίνουν περισσότερο αποδεκτά. Η διαχείριση του συμβιβασμού μεταξύ της ερμηνευσιμότητας και της απόδοσης της πρόβλεψης συχνά παραμελείται με την ανάπτυξη πλαισίων για την προγνωστική ανάλυση, αλλά μπορεί να είναι

κρίσιμη για την ανάπτυξη των μοντέλων στην κλινική πρακτική. Ένας πιθανός τρόπος για την εξισορρόπηση μεταξύ των «white» και «black boxes» είναι η χρήση μίας πιο σύνθετης στρατηγικής γνωστή ως super learning ή στοίβαξης και την αποδόμηση των δομικών στοιχείων του. Ουσιαστικά, η προσέγγιση της super learning προσαρμόζει και στοιβάζει πολλά διαφορετικά μοντέλα μαζί στα δεδομένα και επιλέγει τον καλύτερο σταθμισμένο συνδυασμό. Αν και οι προσεγγίσεις της super learning θεωρείται ότι έχουν μέγιστη ακρίβεια πρόβλεψης και ελάχιστη ερμηνευσιμότητα, η αποδόμηση σε συνιστώσες είναι ένα απαραίτητο βήμα για την ερμηνευσιμότητα και, συνεπώς, την κλινική χρησιμότητα. Αυτό μπορεί να επεκταθεί για τη δοκιμή διαφόρων domains που πρέπει να συμπεριληφθούν στο μοντέλο, να βελτιστοποιηθεί το μοντέλο και να καθοδηγήσει τις μελλοντικές διερευνήσεις των δεδομένων (Σχήμα 2.7- Prosperi *et al.*, 2018).

Σχήμα 2.7

Πολυπλοκότητα των μοντέλων σε διάφορα domain

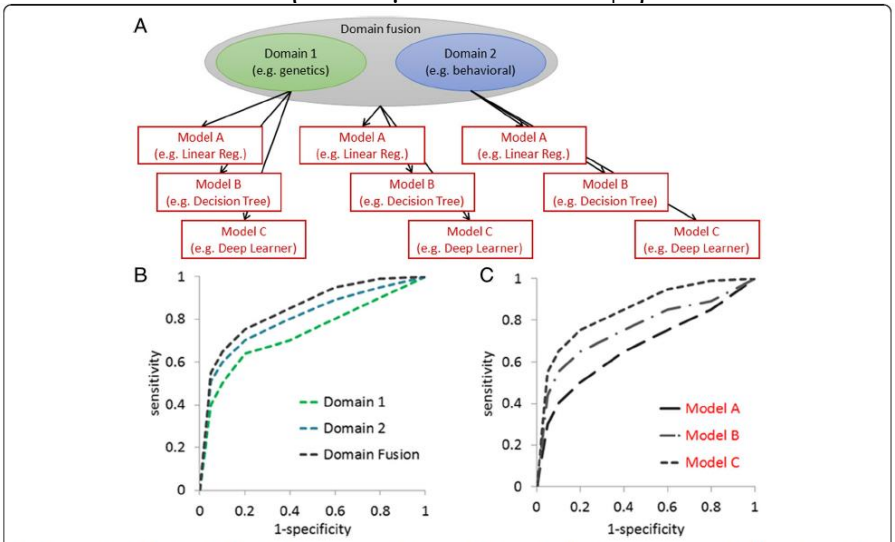


Fig. 7 Domain-guided and complexity-guided model selection. **a** Hypothetical data set with two domains and their merged domain, on which models of increasing complexity (linear regression, decision tree, and deep learner) are fit and compared. For example, using the receiver-operating characteristic, (**b**) the predictive performance of a prime model can be assessed using a single domain or merged domains, or (**c**) different models can be compared within the same domain space

Για παράδειγμα, ο αλγόριθμος μάθησης super ICU (SICULA) έχει κατασκευαστεί για την πρόβλεψη της θνησιμότητας. Τα Post-hoc εργαλεία για τον εντοπισμό της σημασίας των επιμέρους μεταβλητών μπορούν να διασπάσουν την πολυπλοκότητα των «black-box» μοντέλων όπως το random forest ή τα νευρωνικά δίκτυα. Η πολυπλοκότητα των μοντέλων δεν ορίζεται καθολικά, αλλά οι δείκτες όπως η Vapnik-Chervonenkis διάσταση μπορούν να χρησιμοποιηθούν. Κατά την επιλογή των μοντέλων με βάση την πολυπλοκότητά τους, υπάρχουν δύο πλεονεκτήματα: 1) κατώτατα όρια επιδόσεων μπορούν να οριστούν με βάση την κλινική χρησιμότητα, έτσι ώστε ένα πιο ερμηνεύσιμο μοντέλο που είναι λιγότερο ακριβές από ένα πιο πολύπλοκο θα μπορούσε να επιλεγεί εάν πληροί την απαιτούμενη ευαισθησία (sensitivity ή την ειδικότητα (specificity) και 2) η απλούστευση του μοντέλου και η ερμηνεία μπορεί να βοηθήσει στην κατανόηση ευρημάτων για την ανάπτυξη νέων αιτιολογικών υποθέσεων. Παρ' όλα αυτά, η εικόνα δεν είναι τόσο απλή, καθώς δεν υπάρχει καμία εγγύηση ότι οι συνδυασμένες πληροφορίες που προκαλούνται από έναν super learner θα είναι εύκολο να αποδομηθούν, εάν τα τελικά μοντέλα είναι deep neural networks, θα εξακολουθούν να είναι

πολύ δύσκολο να ερμηνευθούν. Η ερμηνευσιμότητα των πολύπλοκων μοντέλων θα εξακολουθεί να περιορίζεται από την εγγενή ερμηνευσιμότητα των υποκείμενων συνιστωσών και συναρτήσεων. Η downstream ανάλυση όπως η κατάταξη της σημασίας των μεταβλητών ή η μερική εξάρτηση τους μπορεί να είναι χρήσιμες, αλλά οι λύσεις αυτές εξαρτώνται σε μεγάλο βαθμό από το μοντέλο και μπορεί να μεροληπτούν λόγω πολυάριθμων παραγόντων (όπως η πολυσυγγραμικότητα των μεταβλητών).

Εκτός από τη σημασιολογική διαλειτουργικότητα, απαιτείται η διαλειτουργικότητα των φάσεων μοντελοποίησης. Χρησιμοποιώντας τυποποιημένα επίπεδα εννοιολογικής διαλειτουργικότητας, η διαλειτουργικότητα της μοντελοποίησης μπορεί να οριστεί ως "ρεαλιστική διαλειτουργικότητα", δηλαδή την επίγνωση των μεθόδων και των διαδικασιών, που βρίσκονται πάνω από τη σημασιολογική διαλειτουργικότητα. Οι Reps *et al.* (2018) εισήγαγαν ένα τυποποιημένο πλαίσιο που αξιοποιεί το OHDSI και το OMOP όχι μόνο για την διαφάνεια στον ορισμό του προβλήματος και την επιλογή των κατάλληλων συνόλων δεδομένων, αλλά και για την κατασκευή μεταβλητών από τα παρατηρητικά δεδομένα, την εκμάθηση του προβλεπτικού μοντέλου και την επικύρωση της απόδοσης του μοντέλου.

Για να είναι χρήσιμο οποιοδήποτε μοντέλο για τη δημόσια υγεία, θα πρέπει να είναι ανθεκτικό στο θόρυβο και γενικεύσιμο. Επίσης, θα πρέπει να παρουσιάζονται με διαφάνεια όσον αφορά την απόδοσή τους και την αναπαραγωγιμότητα και οι βιβλιοθήκες λογισμικού για τη διαφοροποημένη ιδιωτικότητα θα πρέπει να επιβάλλονται ως γενικά πρότυπα ώστε να διευκολύνεται η ανταλλαγή δεδομένων και η αναπαραγωγή των εργασιών. Κατά την αξιοποίηση αυτών των προτύπων, πρέπει να να εξεταστεί κατά πόσον τα ευρήματα υπερβαίνουν τη στατιστική σημαντικότητα και σηματοδοτούν πεδία κλινικής σημασίας. Ακόμη, ο προσδιορισμός των παραγόντων κινδύνου που είναι αμετάβλητοι είναι ανέφικτοι για παρεμβάσεις και σε περιπτώσεις ασθενειών όπου δεν υπάρχουν θεραπείες, η ακρίβεια των διαγνώσεων της νόσου δεν θα επηρεάσει τις αποφάσεις της κλινικής θεραπείας. Ωστόσο, η πρόσθετη διορατικότητα σχετικά με τους μηχανισμούς των εξέλιξης της νόσου μπορεί να αποκτηθεί. Η σύνδεση και η συστηματοποίηση των δεδομένων σε πολλαπλούς τομείς της ζωής έχει τη δυνατότητα να αυξήσει την εκπαίδευση των ασθενών και τη συμμετοχή στην υγειονομική περίθαλψη. Αυτό με τη σειρά του θα μπορούσε να οδηγήσει σε βελτίωση της ενδυνάμωσης των ασθενών και στη λήψη κοινών αποφάσεων, οι οποίες σχετίζονται με τη βελτίωση των αποτελέσματων της υγείας. Με τη δημιουργία ενός σημείου πρόσβασης για τα άτομα να βλέπουν τα EHR τους και άλλες μεταβλητές που μπορεί να επηρεάζουν την υγεία τους, το avatar της υγείας ενδυναμώνει τους ασθενείς να αναλάβουν δράση. Η επίδραση αυτής της ενδυναμωσης έχει αποδειχθεί ότι τροποποιεί τις συμπεριφορές υγείας των ασθενών για τη μείωση του κινδύνου ρευματοειδούς αρθρίτιδας και να κάνουν προετοιμασίες για την ασθένεια στο μέλλον. Επιπλέον, τα avatar υγείας μπορούν να είναι χώροι για την αύξηση των διαθέσιμων εγκαταστάσεων υγειονομικής περίθαλψης και την ευκολία σύνδεσης με τη φροντίδα, κατι το οποίο βρίσκεται επί του παρόντος σε εξέλιξη. Επιπλέον, δοκιμάζεται με την βοήθεια της φορητής τεχνολογίας, να μπορεί να ανιχνέυεται η κολπική μαρμαρυγή και να υπάρχει άμεση σύνδεση με γιατρό μέσω κινητών συσκευών. Εκτός από την επίδραση στους γιατρούς για την υποστήριξη κλινικών αποφάσεων και στην ενδυνάμωση των ασθενών, το avatar υγείας μπορεί να είναι ένας έξυπνος αλγόριθμος που αντιστοιχεί τους ανθρώπους με τις ερευνητικές μελέτες

με βάση το πλήρες ιατρικό ιστορικό τους και άλλους σχετικούς με την υγεία παράγοντες. Αυτό θα επιτρέψει στους ερευνητές να προσεγγίσουν τεράστιους σε αριθμό πληθυσμούς και να επιτρέψει την εφαρμογή νέων σχεδιασμών των μελετών, όπως η εξέταση σπάνιων ανεπιθύμητων ενεργειών ενός φαρμάκου που οι τυχαιοποιημένες κλινικές δοκιμές δεν μπορούν να έχουν επαρκή ισχύ για την ανίχνευση. Η δημόσια υγεία εξελίσσεται σε ένα εγχείρημα πολλαπλών τομέων και πολλών ενδιαφερομένων. Το Food and Drug Administration δοκιμάζει πιλοτικά προγράμματα ψηφιακού λογισμικού υγείας. Οι εταιρείες που βρίσκονται εκτός του τομέα της υγειονομικής περίθαλψης ασχολούνται τώρα με τη δημιουργία προγραμμάτων υγειονομικής περίθαλψης για τους υπαλλήλους τους. Ωστόσο, ορισμένα βασικά εμπόδια παραμένουν ακόμη ανοιχτά: Τα μοντέλα πρόβλεψης της μελλοντικής κατάστασης της υγείας δεν είναι ακόμη ακριβή και η δυνατότητά τους να αναλάβουν δράση, δηλαδή να αλλάξουν τις πιθανότητες να εμφανιστεί μια ασθένεια, είναι ακόμη λιγότερο υπολογίσιμη. Ακόμη, στη δημόσια υγεία παρουσιάζεται έλλειψη πλαισίωσης στο κοινωνικό και οικολογικό περιβάλλον και της ενσωμάτωσης με την ηθική και της χάραξης πολιτικής. Τέλος, η οικονομική προσιτότητα, η εμπιστοσύνη και η εκπαίδευση των μαζών σε αυτό το νέο παράδειγμα της ιατρικής θα πρέπει να αντιμετωπιστεί σύντομα (Prosperi *et al*., 2018).

2.10 Ο ρόλος των Big Data και της Στατιστικής Μηχανικής Μάθησης στην Διατροφική επιδημιολογία

Σύμφωνα με τους Morgenstern *et al.* (2021) η διατροφή αναμφισβήτητα κατέχει έναν πολύ σημαντικό ρόλο στην προαγωγή της υγείας. Μάλιστα, παρατηρήθηκε ότι η κακή διατροφή ξεπέρασε πρόσφατα το κάπνισμα που θεωρούνταν ο κύριος παράγοντας κινδύνου για τη νοσηρότητα από μη μεταδοτικές ασθένειες και την θνησιμότητα σύμφωνα με το παγκόσμιο ερευνητικό πρόγραμμα Global Burden of Disease Study. Ως εκ τούτου, οι συνεχιζόμενες προσπάθειες για την κατανόηση των επιπτώσεων της κακής διατροφής στην υγεία πρέπει να αποτελούν ύψιστη προτεραιότητα στις προσπάθειες που γίνονται για τη βελτίωση της δημόσιας υγείας. Μεγάλη πρόοδος στην κατανόηση της διατροφής έχει σημειωθεί κατά τον τελευταίο μισό αιώνα, με επιστημονικά ευρήματα στην διατροφική επιδημιολογία που οδηγεί σε αλλαγές πολιτικής, όπως η απαγόρευση των τρανς λιπαρών σε πολλές χώρες. Παρά τη μεγάλη πρόοδο, οι προκλήσεις παραμένουν, δηλαδή υπάρχουν διάφορα εμπόδια ως προς το να έχουμε ακριβείς μετρήσεις στην διατροφή, κατάλληλη μοντελοποίηση της πολυπλοκότητας της διατροφής, πολυσυγγραμμικότητα και σύγχηση των καταλοίπων. Ο στόχος αυτού του κεφαλαίου είναι να επανεξετάσουμε τον τρόπο με τον οποίο η εφαρμογή μεγάλων συνόλων δεδομένων και η Μηχανική Μάθηση μπορούν να βοηθήσουν στην αντιμετώπιση των προκλήσεων στον τομέα της διατροφικής επιδημιολογίας, με έμφαση στο σφάλμα μέτρησης, στη διατροφική πολυπλοκότητα, στη σύγχυση, στην πρόβλεψη ασθενειών και στις συμπερασματικές μελέτες (Πίνακας 2.1 - Morgenstern *et al.*, 2021).

Πίνακας 2.1
Εφαρμογές των Big Data και της Μηχανικής Μάθησης στη διατροφική επιδημιολογία

TABLE 1 Summary of major potential applications of big data and machine learning to nutritional epidemiology

	Measurement error	Dietary complexity	Confounding	Disease prediction	Inferential studies
Potential big data and machine learning applications	New measurement methods Frequent repeated measures Increased statistical power Increased precision Decreased regression dilution bias	Including more complex and comprehensive dietary exposures Improved modeling of interactions and nonlinearity	New data sources could reduce unmeasured confounding Greater opportunities for use of negative controls and instrumental variables Machine learning methods applied in inferential frameworks	Improved disease predictions with greater incorporation of complex dietary exposures, nonlinearity, and interactions in predictive models	Less biased estimation of causal effects Hypothesis generation with methods for interpreting machine learning models
Limitations	Evaluation of the validity and precision of new measurement methods is in early stages Many proposed methods still rely on self-report Selection bias Investments in big data infrastructure and expertise would be needed Privacy concerns must be addressed	Limited interpretability of unsupervised and supervised machine learning methods High sample sizes needed to reliably model nonadditive and nonlinear relations	Potential for limited interpretability of machine learning–derived covariates Potential for worsening model bias and variance if there is data-driven inclusion of covariates in models	Potential for overfitting Limited interpretability of models Careful validation required to ensure reliable predictions May not enhance performance relative to traditional models (e.g., if interactions and nonlinearity are not very important)	Potential for inaccurate data-driven conclusions Interpretability of machine learning models remains limited

2.10.1 Μοντελοποίηση της πολυπλοκότητας της διατροφής

Η διατροφή είναι ένα πολύπλοκο θέμα, γεγονός που καθιστά να είναι σημαντικός ορισμός των μεταβλητών και ο προσδιορισμός των μοντέλων. Τα τρόφιμα δεν καταναλώνονται μεμονωμένα αλλά, μάλλον, σε ποικίλους συνδυασμούς και αναλογίες. Εάν ένα στοιχείο μειωθεί ή αυξηθεί, άλλα μέρη της διατροφής πρέπει να αλλάξουν αντίστοιχα για να ικανοποιηθούν οι συνολικές ενεργειακές ανάγκες. Επιπλέον, τα θρεπτικά συστατικά και τα τρόφιμα μπορούν να αλληλεπιδρούν μεταξύ τους με συνεργατικούς και ανταγωνιστικούς τρόπους, καθιστώντας το "ολόκληρο" πολύ διαφορετικό από το άθροισμα των μερών του. Δεδομένης αυτής της πολυπλοκότητας, οι προσεγγίσεις για τη διαμόρφωση της διατροφής μπορούν να επικεντρωθούν σε μεμονωμένα θρεπτικά συστατικά, τρόφιμα, ομάδες τροφίμων ή σε διατροφικά μοτίβα. Τα τρέχοντα διατροφικά μοτίβα βασίζονται συχνά σε μία προηγούμενη γνώση σημαντικών πτυχών της διατροφής και συμπυκνώνονται σε μονοδιάστατες μετρήσεις, όπως το σκορ της μεσογειακής διατροφής, (εναλλακτικός) δείκτης υγιεινής διατροφής (Alternative Healthy Eating Index - AHEI) ή το σκορ των διατροφικών προσεγγίσεων για τη διακοπή της υπέρτασης (Dietary Approaches to Stop Hypertension - DASH). Όταν συμπυκνώνονται σε μονοδιάστατες βαθμολογίες, ο πολυδιάστατος χαρακτήρας των διατροφικών προτύπων χάνεται. Αυτά τα διατροφικά πρότυπα μπορεί να ευθύνονται για κάποια συνέργεια, αλλά μόνο όταν οι αλληλεπιδράσεις είναι γνωστές και λαμβάνονται υπόψη κατά τη διάρκεια της κατασκευής της βαθμολογίας. Τέτοιες αλληλεπιδράσεις είναι σπάνια

γνωστές. Επιπλέον, στις μελέτες των θρεπτικών συστατικών, των τροφίμων και των ομάδων των τροφίμων, οι αλληλεπιδράσεις συχνά υποθέτουν σιωπηρά ότι απουσιάζουν στις προδιαγραφές του μοντέλου. Οι αλληλεπιδράσεις θα μπορούσαν να συμπεριληφθούν σε παραμετρικά μοντέλα, αλλά μόνο εάν είναι γνωστές εκ των προτέρων. Τέλος, πολλές διατροφικές επιδημιολογικές μελέτες υποθέτουν γραμμικά μοντέλα συσχετίσεων μεταξύ διατροφής και ασθένειας. Υπάρχουν αναδυόμενες ενδείξεις ότι οι μη γραμμικές σχέσεις μπορεί να είναι πιο συχνές απ' ό,τι πιστεύονταν προηγουμένως. Για παράδειγμα, το αλάτι, οι υδατάνθρακες και τα λίπη μπορεί όλες να έχουν σχέσεις σχήματος U ή J με τις καρδιαγγειακές παθήσεις. Επιπλέον, υποστηρίζεται ότι υπάρχουν διάφορες αλληλεπιδράσεις στην διατροφική επιδημιολογία. Για παράδειγμα, η επίδραση του αλατιού στην υπέρταση φαίνεται να μετριάζεται από το κάλιο και η περιεκτικότητα της διατροφής σε απλούς υδατάνθρακες. Ακατάλληλος προσδιορισμός των μοντέλων λόγω εσφαλμένων ή ελλιπών χαρακτηρισμών, υποθέσεων σχετικά με τις αλληλεπιδράσεις, και υποθέσεις σχετικά με τη γραμμικότητα μπορεί να οδηγήσουν σε συγκαλυμμένες ή ψευδείς συσχετίσεις και μεροληπτικές εκτιμήσεις των αποτελεσμάτων (Morgenstern *et al.*, 2021).

2.10.2 Μέθοδοι Μηχανικής Μάθησης για τη μοντελοποίηση της πολυπλοκότητας της διατροφής σε σχέση με τη νόσο

Η μηχανική μάθηση θα μπορούσε να συμπεριλάβει πιο σύνθετες και πολυπληθέστερες διατροφικές επεξηγηματικές μεταβλητές στα διατροφικά επιδημιολογικά μοντέλα και να βοηθήσει στον εντοπισμό των προβλέψεων. Πολλές τεχνικές μείωσης της διαστασιμότητας χρησιμοποιούνται ήδη συχνά στη διατροφική επιδημιολογία, όπως η Ανάλυση Κύριων Συνιστωσών (PCA), k-means clustering, και η μερική παλινδρόμηση ελαχίστων τετραγώνων. Εκτός από την ομαδοποίηση k-means, μέθοδοι μείωσης της γραμμικής διαστασιμότητας, όπως αυτές, χρησιμοποιούνται τόσο στη μηχανική μάθηση όσο και στην κλασική στατιστική ανάλυση. Ωστόσο, έχει γίνει λιγότερη χρήση μη γραμμικών μεθόδων για την μείωση της διαστασιμότητας στη διατροφική επιδημιολογία, όπως οι autoencoders, η t-distributed stochastic neighbor embedding και η manifold learning. Αν και οι προσεγγίσεις αυτές μπορεί να δημιουργήσουν περισσότερο αντιπροσωπευτικά και ολοκληρωμένα διατροφικά πρότυπα, είναι επίσης πιθανό να πλήττονται από ακόμη χειρότερη ερμηνευσιμότητα από ό,τι οι γραμμικές μεθόδοι μείωσης της διαστασιμότητας. Αντίστοιχες προσεγγίσεις αναπτύσσονται για τη βελτίωση της ερμηνευσιμότητας των διαστάσεων που προκύπτουν. Οι μέθοδοι επιλογής των μεταβλητών είναι ένα άλλο μέσο για την αντιμετώπιση της πολυπλοκότητας της διατροφής. Αυτές οι μέθοδοι μπορούν να περιορίσουν τα μεγάλα διατροφικά δεδομένα σε ένα υποσύνολο πιο σχετικό με την πρόβλεψη του αποτελέσματος υγείας που μας ενδιαφέρει. Και πάλι, έχει υπάρξει κάποια χρήση αυτών των μεθόδων στη διατροφική επιδημιολογία, όπως η χρήση του least absolute shrinkage και του selection operator, οι οποίες βρέθηκε ότι προβλέπουν καλύτερα τους καρδιομεταβολικούς δείκτες με διατροφικά δεδομένα σε σύγκριση με τις παραδοσιακές μεθόδους. Άλλοι συνήθεις αλγόριθμοι επιλογής μεταβλητών, όπως τα regularized trees, οι genetic algorithms και η recursive feature elimination, έχουν χρησιμοποιηθεί λιγότερο. Επιπλέον, έχει χρησιμοποιηθεί ελάχιστα η Μηχανική Μάθηση για

την ανάλυση πολλαπλών επιπέδων κατηγοριοποίησης των τροφίμων, όπως στην περιοκτικότητα των μικροθρεπτικών και των μακροθρεπτικών συστατικών, σε συγκεκριμένους τύπους τροφίμων και ομάδες τροφίμων. Αυτό θα μπορούσε να επιτρέψει τις πιο προγνωστικές πτυχές της διατροφής να προσδιοριστούν εμπειρικά για ένα δεδομένο πρόβλημα.

Γνωρίζουμε μία μελέτη η οποία εφάρμοσε survival gradient boosted machines και survival random forests για την πρόβλεψη της καρδιαγγειακής θνησιμότητας με διατροφικά δεδομένα του National Health and Nutrition Examination Survey (NHANES), που περιλάμβαναν πολυεπίπεδα διατροφικά δεδομένα. Τα μοντέλα αυτά έδειξαν βελτιωμένο score πρόβλεψης και διαχωρισμού όταν συμπεριλάμβαναν και τις 103 διατροφικές μεταβλητές πάνω από τους παραδοσιακούς κλινικούς προγνωστικούς παράγοντες. Όταν οι μόνες προστιθέμενες διατροφικές μεταβλητές ήταν οι εκ των προτέρων διατροφικές ήταν οι εκ των προτέρων διατροφικές βαθμολογίες (MDS, Healthy Eating Index, AHEI και DASH), δεν υπήρξε καμία βελτίωση. Συνολικά, αν και οι αρχικές εφαρμογές φαίνεται να είναι πολλά υποσχόμενες, η χρήση της μηχανικής μάθησης χρησιμοποιείται περισσότερο στο πλαίσιο των υψηλών διαστάσεων και των μεγάλων διατροφικών δεδομένων. Με καμία αρχική επιμέλεια των μεταβλητών από εμπειρογνώμονες και προσεκτικό validation, σημαντικοί προγνωστικοί παράγοντες θα μπορούσαν να διαφύγουν και ασήμαντοι προγνωστικοί παράγοντες να τονιστούν εσφαλμένα. Εκτός από την καλύτερη αποτύπωση της διατροφής, η μηχανική μάθηση μπορεί να μοντελοποιήσει μη γραμμικές και μη προσθετικές σχέσεις με μεγαλύτερη ευελιξία. Επιπλέον, οι σχέσεις αυτές δεν χρειάζεται να είναι γνωστές εκ των προτέρων. Αν και περιορισμένες, υπάρχουν ορισμένες μελέτες που έχουν εφαρμόσει τη μηχανική μάθηση για να μοντελοποιήσουν πιο ευέλικτα τις σχέσεις διατροφής-υγείας. Για παράδειγμα, ένας stochastic gradient boosting regression αλγόριθμος χρησιμοποιήθηκε για να προβλέψει με ακρίβεια τις ατομικές γλυκαιμικές αντιδράσεις στα τρόφιμα με τις διατροφικές συνήθειες, τον τρόπο ζωής, τα ιατρικά, τα εργαστηριακά, τα ανθρωπομετρικά και τα microbiota δεδομένα. Το μοντέλο περιελάμβανε χιλιάδες μεταβλητές και έχοντας τη σημαντικότητα των μεταβλητών και τη μερική εξάρτηση τους, αυτό έδωσε την δυνατότητα να ερμηνεύσουν την συμβολή των μεταβλητών του μοντέλου στις προβλέψεις. Απροσδόκητα, το μοντέλο έδωσε μεγαλύτερη έμφαση στις μεταβλητές που σχετίζονται με τα microbiota δεδομένα. Η μελέτη αυτή ήταν μοναδική μεταξύ άλλων διατροφικών μελετών στη χρήση ενός εναλλακτικού αποτελέσματος με χαμηλό σφάλμα και με ασυνήθιστα ακριβείς διατροφικές μετρήσεις. Μια άλλη επιδημιολογική cohort μελέτη της διατροφής διαπίστωσε 22% αύξηση της ακρίβειας των καρδιομεταβολικών παραγόντων κινδύνου κατά τη σύγκριση του random forest αλγορίθμου σε σχέση με τη γραμμική παλινδρόμηση. Αυτή η μελέτη ενσωμάτωσε ανεξάρτητες μεταβλητές της διατροφής και χρησιμοποίησε PCA για τη μείωση της διαστασιμότητας. Μια άλλη πρόσφατη μελέτη εξέτασε τις συσχετίσεις μεταξύ της διατροφής και των ανεπιθύμητων εκβάσεων της εγκυμοσύνης χρησιμοποιώντας τον αλγόριθμο Super Learner (αλγόριθμος της Μηχανικής Μάθησης) για την εκτίμηση της μέγιστης πιθανοφάνειας, σε σύγκριση με τη λογιστική παλινδρόμηση. Υπήρχαν κυρίως μηδενικές συσχετίσεις στο μοντέλο της λογιστικής παλινδρόμησης. Αντίθετα, ο αλγόριθμος Super Learner έδειξε να υπάρχουν συσχετίσεις μεταξύ των λαχανικών, των φρούτων και των πρόωρων γεννήσεων. Επίσης, έδειξε να υπάρχει

ελάχιστη συσχέτιση μεταξύ της ηλικίας κύησης και των αποτελεσμάτων της προεκλαμψίας, εκτός από περισσότερες ακριβείς εκτιμήσεις.

Οι ερευνητές απέδωσαν αυτή τη διαφορά μεταξύ της μεθόδου της μηχανικής μάθησης και της λογιστικής παλινδρόμησης στη βελτιωμένη μοντελοποίηση της διατροφικής συνέργειας στο μοντέλο της μηχανικής μάθησης. Τέλος, η μελέτη που συζητήθηκε προηγουμένως χρησιμοποίησε μοντέλα μηχανικής μάθησης για την πρόβλεψη της θνησιμότητας από καρδιαγγειακές νόσους με διατροφικά δεδομένα του NHANES, κάτι που έδειξε βελτιωμένο προγνωστικό score και διαχωρισμό σε σύγκριση με μοντέλα αναλογικών κινδύνων του Cox. Ενδιαφέρον παρουσιάζει η προσθήκη των διατροφικών δεδομένων στο στατιστικό μοντέλο διότι δεν βελτίωσε τη διαχωριστική ικανότητα πρόβλεψης ή το score του, αλλά όταν τα δεδομένα προστέθηκαν στα μοντέλα μηχανικής μάθησης, και τα δύο μέτρα βελτιώθηκαν. Αυτό υποστηρίζει την πρόταση ότι τα μοντέλα μηχανικής μάθησης μπορούν να αξιοποιήσουν καλύτερα τα διατροφικά δεδομένα στη μοντελοποίηση των αποτελεσμάτων υγείας, ίσως τόσο με την ενσωμάτωση περισσότερων διατροφικών μεταβλητών και με τη συνεκτίμηση για μη γραμμικές και μη προσθετικές σχέσεις (Morgenstern *et al.*, 2021).

2.10.3 Βελτίωση της πρόβλεψης της νόσου

Σχετικά λίγα κλινικά μοντέλα πρόβλεψης ή μοντέλα πρόβλεψης της δημόσιας υγείας περιλαμβάνουν διατροφικά δεδομένα. Η ενσωμάτωση τέτοιων δεδομένων σε μοντέλα πρόβλεψης θα μπορούσαν να βελτιώσουν τις προβλέψεις των αποτελεσμάτων της υγείας. Τα μοντέλα πρόβλεψης διαφέρουν στις περισσότερες έρευνες στον τομέα της διατροφικής επιδημιολογίας, επειδή περιλαμβάνουν γενικά όλα τις μεταβλητές που θεωρούνται σχετικές με την πρόβλεψη ενός αποτελέσματος. Ενδιαφέρονται περισσότερο για τα συνολικά χαρακτηριστικά πρόβλεψης του μοντέλου παρά για τις συσχετίσεις μεμονωμένων μεταβλητών έκθεσης (π.χ. περιοχή κάτω από την καμπύλη ROC αντί του σχετικού κινδύνου) και ενδιαφέρονται λιγότερο για την ερμηνευσιμότητα. Τα μοντέλα πρόβλεψης για καρδιαγγειακές παθήσεις, ένα από τα σημαντικότερα πεδία της επιστήμης της διατροφής, έχουν εκτενώς μελετήθει τις τελευταίες πέντε δεκαετίες. Εργαλεία πρόβλεψης κινδύνου, όπως αυτό που αναπτύχθηκε αρχικά από τη μελέτη Framingham το 1967, εξακολουθούν να χρησιμοποιούνται συνήθως στην κλινική πρακτική για τον προσδιορισμό της ανάγκης για υπερτασικά φάρμακα και φάρμακα χοληστερόλης. Πιο πρόσφατα, έχουν αναπτυχθεί μοντέλα σε πληθυσμιακό επίπεδο που μπορούν να χρησιμοποιηθούν για την εφαρμογή προληπτικών παρεμβάσεων της δημόσιας υγείας, την ενημέρωση των φορέων για τη μελλοντική επιβάρυνση από ασθένειες και την αξιολόγηση της επίδρασης των ενεργειών στη δημόσια υγεία. Συνήθως, τα μοντέλα πρόβλεψης περιλαμβάνουν πολύ λίγα διατροφικά συστατικά, όπου όταν αυτά συμπεριλαμβάνονται, χρησιμοποιούνται συνήθως πολύ απλουστευμένοι διατροφικοί παράγοντες (π.χ. μόνο ένας μικρός αριθμός τροφίμων ή αναλογίες θρεπτικών συστατικών). Οι λόγοι για τον αποκλεισμό των διατροφικών μεταβλητών από τα σημερινά μοντέλα πρόβλεψης μπορεί να περιλαμβάνουν την απουσία διατροφικών δεδομένων σε πολλές κοινά

χρησιμοποιούμενες πηγές δεδομένων, δυσκολία στην συλλογή και περιορισμένη ή μηδενική πρόσθετη προγνωστική απόδοση (π.χ. λόγω υψηλού σφάλματος μέτρησης, χρήση υπεραπλουστευμένων scores των διατροφικών προτύπων ή ακατάλληλες μεθόδους μοντελοποίησης).

Συνεπώς, η συμπερίληψη διατροφικών δεδομένων σε μοντέλα πρόβλεψης, ιδίως σε συνδυασμό με τη χρήση νέων μεθόδων συλλογής δεδομένων και μεθόδων μηχανικής μάθησης, θα μπορούσε να αποτελέσει ένα σημαντικό και σε μεγάλο βαθμό μία δίοδο για τη βελτίωση των επιδόσεων. Όπως συζητήθηκε προηγουμένως, οι νέες μέθοδοι μέτρησης θα μπορούσαν να μετριάσουν το σφάλμα μέτρησης και να επιτρέψουν στα μοντέλα πρόβλεψης να επωφεληθούν από σχετικά μικρές συσχετίσεις. Επιπλέον, η χρήση μοντέλων μηχανικής μάθησης θα μπορούσε να αξιοποιήσει καλύτερα τις πολύπλοκες διατροφικές εκθέσεις, τις μη προσθετικές σχέσεις και τις μη γραμμικές συσχετίσεις για τη βελτίωση των μοντέλων πρόβλεψης. Μια πρόσφατη cohort μελέτη υποστηρίζει αυτή την ιδέα, διότι κατέδειξε βελτιώσεις των επιδόσεων πρόβλεψης για την θνησιμότητα από καρδιαγγειακές νόσους όταν συνδυάζονται διατροφικά δεδομένα με μεθόδους της μηχανικής μάθησης. Ένα ακόμη πλεονέκτημα της εφαρμογής της μηχανικής μάθησης είναι ότι το cross-validation καθιστά πολλούς αλγορίθμους πιο ανθεκτικούς στις επιδράσεις της πολυσυγγραμμικότητας στο πλαίσιο της πρόβλεψης. Επιπλέον, αυτό το εσωτερικό validation θα μπορούσε να επιτρέψει τον εντοπισμό διατροφικών μοτίβων και παραγόντων που είναι πιο σημαντικοί σε συγκεκριμένους πληθυσμούς για την πρόβλεψη συγκεκριμένων ασθενειών. Συνολικά, οι νέες πηγές δεδομένων και οι μέθοδοι μηχανικής μάθησης προσφέρουν δυνατότητες για τη βελτίωση των μοντέλων πρόβλεψης χρόνιων ασθενειών μέσω της ενσωμάτωσης διατροφικών δεδομένων (Morgenstern *et al.*, 2021).

2.10.4 Περιορισμοί των Big Data και της Μηχανικής Μάθησης στην πρόβλεψη ασθενειών

Παρά τις πιθανές θετικές επιπτώσεις στην προβλεπτική μοντελοποίηση, η εφαρμογή των Big Data και της μηχανικής μάθησης έχει αρκετές πιθανές παγίδες. Πρώτον, η επιλογή της μεροληψίας και το συστηματικό σφάλμα μέτρησης σε νέες πηγές δεδομένων αποτελούν μία ανησυχία. Εάν εξαιρεθούν από τα training data sets, οι ευάλωτοι πληθυσμοί θα μπορούσαν να περιθωριοποιηθούν περαιτέρω από τους αλγορίθμους πρόβλεψης που είναι ανακριβείς γι' αυτούς. Επιπλέον, δεδομένου ότι οι μέθοδοι της Μηχανικής Μάθησης είναι συνήθως μη θεωρητικοί και μερικές φορές μη κατανοητοί, είναι ευάλωτες σε περίπτωση που κάποια μορφή της διαδικασίας παραγωγής δεδομένων αλλάξει. Στην περίπτωση αυτή, μπορεί απροσδόκητα να γίνουν ανακριβείς, οπότε οι ερευνητές θα πρέπει να λαμβάνουν μέτρα για να διασφαλιστούν από αυτό το ενδεχόμενο. Μία άλλη σημαντική παρατήρηση είναι ότι τα πολύπλοκα μοντέλα της μηχανικής μάθησης δεν βελτιώνουν πάντα την πρόβλεψη. Είναι περισσότερο ευέλικτα από τα περισσότερα παραμετρικά μοντέλα παλινδρόμησης, όμως αυτό τα καθιστά πιο επιρρεπή στην υπερπροσαρμογή. Η υπερπροσαρμογή είναι ένα σφάλμα που μπορεί να συμβεί με τα πιο ευέλικτα μοντέλα όταν προσαρμόζονται πολύ στενά στα περιορισμένα παρατηρούμενα δεδομένα, το οποίο μπορεί να οδηγήσει σε χειρότερες επιδόσεις σε νέα δεδομένα. Το σχετικό

πλεονέκτημά τους εξαρτάται από τη σημασία των αλληλεπιδράσεων και της μη γραμμικότητας για ένα δεδομένο πρόβλημα. Ιδανικά, πολλά μοντέλα της μηχανικής μάθησης και στατιστικά μοντέλα θα πρέπει να δοκιμαστούν και να αξιολογηθούν χρησιμοποιώντας cross-validation για ένα δεδομένο πρόβλημα πρόβλεψης. Μη γραμμικά παραμετρικά στατιστικά μοντέλα, όπως τα κλασματικά πολυώνυμα και τα περιορισμένα κυβικά splines θα πρέπει επίσης να εξεταστούν. Ένα σχετικό ζήτημα με τις περισσότερες προσεγγίσεις της μηχανικής μάθησης είναι ότι συνήθως απαιτούν περισσότερες παρατηρήσεις ανά μεταβλητή για να γίνουν ισχυρές προβλέψεις. Ως εκ τούτου, μπορεί συχνά να μην είναι κατάλληλη η εφαρμογή των μεθόδων της μηχανικής μάθησης σε μικρότερα σύνολα δεδομένων. Εναλλακτικά, πολυάριθμοι μέθοδοι επιλογής των μεταβλητών και μείωσης της διαστασιμότητας μπορούν να χρησιμοποιηθούν, για τη μείωση του αριθμού των μεταβλητών που περιλαμβάνονται σε ένα μοντέλο. Επίσης, ορισμένοι supervised αλγόριθμοι της μηχανικής μάθησης, όπως η random forest, είναι σχετικά ανθεκτικοί στην παρουσία μη κατατοπιστικών μεταβλητών. Γενικά, οι στατιστικές τεχνικές αποδίδουν καλύτερα και είναι περισσότερο γενικεύσιμες σε καταστάσεις στις οποίες είναι διαθέσιμο μόνο ένα μικρό μέγεθος δείγματος και οι μη γραμμικές και μη προσθετικές σχέσεις δεν έχουν μεγάλη επιρροή. Τέλος, είναι σημαντικό να σημειωθεί ότι η μοντελοποίηση των αποτελεσμάτων υγείας διαφέρει από τους τομείς εφαρμογών στους οποίους αναπτύχθηκε αρχικά η μηχανική μάθηση. Από την άλλη πλευρά, στους ιατρικούς τομείς ένα σημαντικό ποσοστό του σφάλματος πρόβλεψης πιθανόν να προέρχεται από μη τροποποιημένη στοχαστικότητα, θέτοντας ένα χαμηλότερο ανώτατο όριο στην πιθανή ακρίβεια πρόβλεψης. Έτσι, στην έρευνα στον τομέα της υγείας, η αβεβαιότητα στις εκτιμήσεις και στις προβλέψεις της πιθανότητας είναι πιο σημαντική απ' ό,τι ήταν συχνά στη μηχανική μάθηση. Παρόλο που δεν γίνεται συχνά, οι εκτιμήσεις αβεβαιότητας μπορούν να προκύψουν για αναλύσεις μηχανικής μάθησης με τη χρήση επαναδειγματοληψίας και Μπεϋζιανές προσεγγίσεις. Τέλος, στο πλαίσιο της έρευνας στον τομέα της υγείας, είναι σημαντικό να επικεντρωθούμε κυρίως στο score ως προβλεπτικό μέτρο απόδοσης, το οποίο συνεπάγεται τη συμφωνία μεταξύ προβλεπόμενων και παρατηρούμενων πιθανοτήτων σε όλο το φάσμα κινδύνου. Αυτό έρχεται σε αντίθεση με την πιο συχνή χρήση διαχωριστικών μέτρων απόδοσης, όπως η περιοχή κάτω από την καμπύλη ROC (Morgenstern *et al.*, 2021).

2.10.5 Επαγωγικές μελέτες

Αν και οι περισσότερες έρευνες της μηχανικής μάθησης και των Big Data έχουν επικεντρωθεί στην πρόβλεψη ή την ταξινόμηση, θα μπορούσε επίσης να βοηθήσει στην ενημέρωση των επαγωγικών μελετών στον τομέα της διατροφικής επιδημιολογίας. Πρώτον, εάν είναι επιτυχής η μείωση της μη διαφορικής μέτρησης του σφάλματος, η αύξηση των μεγεθών των δειγμάτων και οι νέες μέθοδοι των διατροφικών μετρήσεων θα μπορούσαν να βοηθήσουν στην ανίχνευση μικρότερων μεγεθών επίδρασης και να μειωθούν οι επιπτώσεις της πολυσυγραμμικότητας στη σταθερότητα των συντελεστών. Επιπλέον, η εφαρμογή της μηχανικής μάθησης θα μπορούσε βοηθήσει στη δημιουργία υποθέσεων, καθώς οι μέθοδοι για την ερμηνεία πολύπλοκων αλγορίθμων βελτιώνονται. Ήδη, οι τρέχουσες τεχνικές, όπως η

permutation feature importance, οι συσσωρευμένες τοπικές επιδράσεις, τα διαγράμματα μερικής εξάρτησης, τα Shapley values, οι τοπικές ερμηνεύσιμες εξηγήσεις με βάση το μοντέλο και τα interaction h-statistics μπορούν να χρησιμοποιηθούν σε σχεδόν οποιαδήποτε μοντέλο μηχανικής μάθησης για την αποκάλυψη της μορφής των σχέσεων μεταξύ των προβλέψεων και των αποτελεσμάτων, καθώς και σε σημαντικές αλληλεπιδράσεις. Επιπλέον, η μείωση της διαστασιμότητας και οι μέθοδοι επιλογής των μεταβλητών μπορούν να χρησιμοποιηθούν για την εξαγωγή διατροφικών προτύπων και προγνωστικών διατροφικών παραγόντων για περαιτέρω μελέτη. Δεδομένου του υψηλού επιπέδου πολυπλοκότητας της διατροφής, αυτές οι διερευνητικές προσεγγίσεις μπορεί να είναι ιδιαίτερα χρήσιμες. Επίσης, ένα πλεονέκτημα των διατροφικών προτύπων και της επιλογής μεταβλητών με βάση τα δεδομένα είναι ότι μπορεί να αντικατοπτρίζουν περισσότερο τη σχετική διατροφική διακύμανση σε έναν τοπικό πληθυσμό από ό,τι οι εκ των προτέρων βαθμολογίες που αναπτύχθηκαν άλλου. Επιπλέον, εάν το σύνολο των διατροφικών δεδομένων ενσωματώνεται σε μια ανάλυση με μεθόδους της μηχανικής μάθησης, συμπεριλαμβανομένης της ταξινόμησης πολλαπλών επιπέδων των τροφίμων/θρεπτικών συστατικών, μπορεί να υπάρχουν πιθανές εξηγήσεις για τη διενέργεια επιλεκτικών αναλύσεων. Αυτό δεν θα ήταν πάντα σκόπιμο, διότι οι μελέτες που βασίζονται σε υποθέσεις θα απαιτούσαν μια πολύ πιο επιλεκτική ανάλυση, αλλά θα μπορούσε να είναι μια χρήσιμη προσέγγιση για διερευνητικές μελέτες.

Μια πρόσθετη σκέψη είναι ότι τα Big Data και η μηχανική μάθηση μπορεί να επιτρέψουν πιο ολοκληρωμένη και ακριβή ενσωμάτωση των συγχυτικών παραγόντων στην ανάλυση, μειώνοντας ενδεχομένως την σύγχυση των καταλοίπων. Τέλος, η μεγαλύτερη διαθεσιμότητα των Big Data μπορεί να επιτρέψει τη μελέτη περισσότερων μετα-διατροφικών παραγόντων, όπως ο χρόνος των γευμάτων, η προετοιμασία και οι μέθοδοι μαγειρέματος, οι κοινωνικές πλευρές του δείπνου, η τοποθεσία του φαγητού και άλλοι πρόσθετοι παράγοντες (π.χ. φαγητό κατά την παρακολούθηση τηλεόρασης). Η μηχανική μάθηση θα μπορούσε επίσης να ενισχύσει τις παρατηρητικές μελέτες που αναζητούν στοιχεία για αιτιώδεις σχέσεις στην διατροφική επιδημιολογία, εντός ενός πλαισίου πιθανών αποτελεσμάτων. Νέοι τρόποι χρήσης της μηχανικής μάθησης για τη μερική αυτοματοποίηση δημιουργίας scoer και επιλογής συγχυτικών παραγόντων από δεδομένα υψηλής διάστασης έχουν ήδη περιγραφεί. Επιπλέον, η στοχευμένη εκτίμηση μέγιστης πιθανοφάνειας (Targeted Maximum Likelihood Estimation - TMLE) μπορεί να χρησιμεύσει ως εναλλακτική λύση στα scores και την εκτίμηση της αιτιώδους επίδρασης με βάση τον υπολογισμό G, αφού ενσωματώνοουμε μεθόδους μηχανικής μάθησης, όπως ο Super Learner. Σε συνδυασμό με τον αλγόριθμο super Learner, η TMLE έχει επιδείξει λιγότερο μεροληπτική εκτίμηση των αιτιωδών επιδράσεων από τις παραδοσιακές προσεγγίσεις. Η κύρια διαφορά είναι η χρήση της μηχανικής μάθησης κατά τη διάρκεια μιας δευτερεύουσας φάσης για την καλύτερη εξισορρόπηση του συμβιβασμού μεροληψίας-διακύμανσης στην εκτίμηση του αιτιώδους αποτελέσματος. Όπως περιγράφεται προηγουμένως, μια αρχική εφαρμογή της TMLE στη διατροφική επιδημιολογία διαπίστωσε σχέσεις μεταξύ της πρόσληψης φρούτων και λαχανικών και των αποτελεσμάτων της εγκυμοσύνης που δεν αποκαλύφθηκαν με την λογιστική παλινδρόμηση. Οι εκτιμήσεις των αποτελεσμάτων της TMLE ήταν επίσης πιο ακριβείς (Morgenstern *et al.*, 2021).

2.10.6 Περιορισμοί των Big Data και της Μηχανικής Μάθησης για τις συμπερασματικές μελέτες

Αν και τα μεγάλα δεδομένα και η μηχανική μάθηση μπορεί να είναι χρήσιμα για την πληροφόρηση των επαγωγικών μελετών μέσω της δημιουργίας υποθέσεων και την εφαρμογή τους στα πλαίσια της αιτιώδους συμπερασματολογίας, δεν αρκούν από μόνα τους για την αιτιώδη συμπερασματολογία. Για τα οποιαδήποτε πειραματικά δεδομένα, υπάρχουν πολλά αιτιώδη μοντέλα που θα μπορούσαν να εξηγήσουν τις παρατηρούμενες σχέσεις. Ως εκ τούτου, η γνώση των ειδικών του τομέα είναι απαραίτητη για την ενημέρωση των εκ των προτέρων αιτιωδών μοντέλων, την ερμηνεία των αποτελεσμάτων που παράγονται από αλγορίθμους και την τοποθέτηση των ευρημάτων στο ευρύτερο πλαίσιο των αποδεικτικών στοιχείων. Ειδικότερα, αν και τα Big Data μπορούν να παρέχουν πρόσθετες δυνατότητες για τον έλεγχο μη μετρημένων συγχυτικών παραγόντων, χρήση αρνητικών ελέγχων και την εύρεση μεταβλητών, χωρίς επαρκή προνοητικότητα, ενέχει επίσης υψηλότερο κίνδυνο μεροληπτικής επίδρασης των εκτιμήσεων και συγκάλυψης άμεσων επιδράσεων μέσω συγκρουόμενων και μεσολαβητικών μεταβλητών στα μοντέλα. Περαιτέρω ζητήματα κατά τη χρήση των Big Data και της μηχανικής μάθησης για την πληροφόρηση στοιχείων για αιτιώδεις σχέσεις είναι η επιλογή μεροληψίας και το συστηματικό σφάλμα μέτρησης. Και τα δύο πρέπει να κατανοηθούν καλύτερα για να διασφαλιστούν έγκυρα και γενικεύσιμα αποτελέσματα. Τελευταία, θα πρέπει να χρησιμοποιούνται μέθοδοι επιλογής των μεταβλητών στο πλαίσιο αυτό με προσοχή. Εάν αυτές οι τεχνικές χρησιμοποιηθούν για τον καθορισμό ενός τελικού μοντέλου, ιδίως εάν η μεταβλητή απόκρισης χρησιμοποιήθηκε κατά τη διάρκεια επιλογής των μεταβλητών, υπάρχει μεγάλος κίνδυνος ανακριβών συμπερασμάτων (Morgenstern *et al.*, 2021).

Κ Ε Φ Α Λ Α Ι Ο 3

Εφαρμογή μεθόδων Στατιστικής Μηχανικής Μάθησης στη δημόσια υγεία

Ο σκοπός αυτού του κεφαλαίου είναι να μπορέσουμε να δούμε την χρησιμότητα της Στατιστικής Μηχανικής Μάθησης σε θέματα που αφορούν την δημόσια υγεία, βλέποντας ορισμένες μελέτες πάνω στις οποίες χρησιμοποιήθηκαν τεχνικές της Στατιστικής Μηχανικής Μάθησης.

3.1 Εκτίμηση της έκθεσης στον αμίαντο σε βιομηχανίες της Ιταλίας

Ο εντοπισμός και η παρακολούθηση του καρκίνου αποτελούν σημαντικές πτυχές της προστασίας της υγείας. Η πρώτη μελέτη, " Estimation of Occupational Exposure to Asbestos in Italy by the Linkage of Mesothelioma Registry (ReNaM) and National Insurance Archives Methodology and Results", των Airoldi *et al.* (2020), αποτελεί μια έρευνα όπου δημιουργεί έναν κατάλογο βιομηχανιών σχετικά με την έκθεση στον αμίαντο και εντοπίζει περιπτώσεις καρκίνου από τα άτομα τα οποία εργάζονταν σε αυτές τις βιομηχανίες (Chan & Chang, 2020).

Ο ιταλικός νόμος για την προστασία των εργαζομένων (D.Leg. 81/2008) περιλαμβάνει ένα σύστημα για την καταγραφή των καρκίνων που εντοπίζονται σε εργαζομένους με βάση τρεις δραστηριότητες: Το Εθνικό Μητρώο Μεσοθηλιώματος (National Mesothelioma Registry - ReNaM), το εθνικό μητρώο καρκίνου του ιγμορείου (Sinonasal Cancer National Registry - ReNaTuns) και ένα σύστημα για τον εντοπισμό συστάδων του καρκίνου εργασιακής προέλευσης (Occupational Cancer Monitoring-OCM). Το ReNaM και το ReNaTuns ακολουθούν την ίδια μεθοδολογία. Δηλαδή σε κάθε περιοχή, ένα ειδικό επιχειρησιακό κέντρο εντοπίζει τα κρούσματα στα νοσοκομεία, τους παίρνει συνεντεύξεις σχετικά με την εργασιακή και περιβαλλοντική έκθεση και αποστέλλει τα δεδομένα σε ένα εθνικό κέντρο για στατιστικές αναλύσεις και για την υποβολή των αναφορών. Η μεθοδολογία για το OCM είναι διαφορετική, επειδή το OCM στοχεύει σε όλους τους τύπους καρκίνου με υπάρχοντα στοιχεία εργασιακής αιτιολογίας, καθώς δεν θα ήταν δυνατόν να συλλεχθούν ατομικές πληροφορίες μέσω συνεντεύξεων για την εργασιακή έκθεση για όλα τα άτομα.

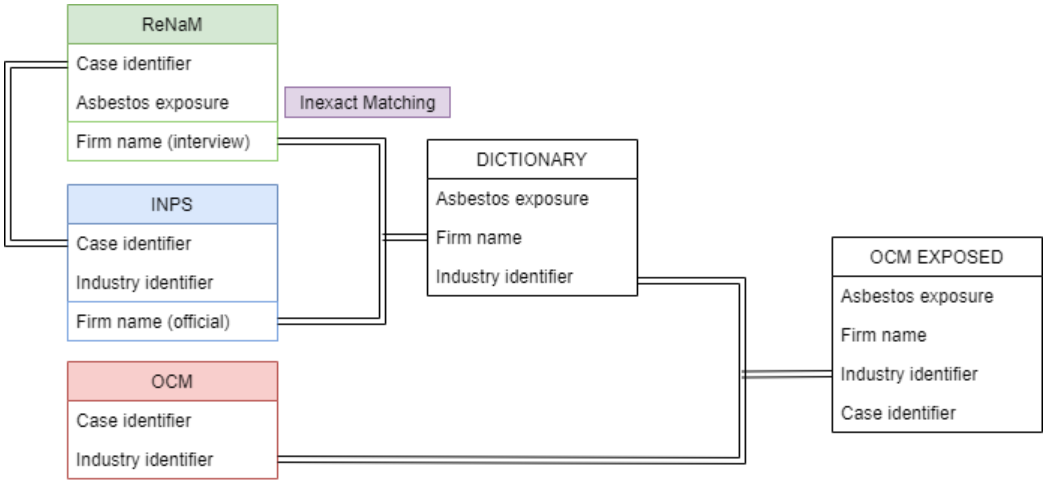
Στην Ιταλία κάθε χρόνο, εντοπίζονται 371.000 νέες περιπτώσεις κακοήθους νεοπλασμάτων, εκ των οποίων οι 39.000 περιπτώσεις είναι καρκίνος του πνεύμονα (Airoldi *et al.*, 2020). Ως εκ τούτου, επινοήθηκε ένα διαφορετικό μοντέλο λειτουργίας για το OCM, σύμφωνα με την έρευνα που διεξήχθη από τους Crosignani *et al.* (2006) και τους Oddone *et al.*

(2013), για τον εντοπισμό των περιπτώσεων καρκίνου που σχετίζονται με την εργασία. Οι περιπτώσεις καρκίνου και οι πληθυσμιακοί έλεγχοι συνδέονται με το αρχείο συνταξιοδοτικών εισφορών του ιταλικού εθνικού ινστιτούτου κοινωνικής ασφάλισης (Italian National Social Security Institute - INPS). Τα δεδομένα για κάθε έτος περιλαμβάνουν την απασχολούσα εταιρεία, καθώς και τον βιομηχανικό της τομέα, αλλά δεν παρέχουν πληροφορίες σχετικά με την πραγματική έκθεση. Οι Crosignani *et al.* (2006) και οι Oddone *et al.* (2013) ανέπτυξαν τη μεθοδολογία για τη χρήση αυτών των πληροφοριών σε επιδημιολογικές αναλύσεις για τη μέτρηση του κινδύνου ανά βιομηχανία και βιομηχανικό τομέα, με το σχεδιασμό ενός case-control study. Το μοντέλο σύνδεσης των αρχείων έχει το πλεονέκτημα ότι μπορεί να εφαρμοστεί σε μεγάλη κλίμακα με πολύ περιορισμένους πόρους, αλλά έχει αρκετά όρια. Συγκεκριμένα, οι πληροφορίες που παρέχονται από το αρχείο σύνδεσης περιορίζεται στον εντοπισμό των επιχειρήσεων στις οποίες εργαζόταν το άτομο, χωρίς να δίνονται πληροφοριές για την έκθεση στον κίνδυνο. Οι πραγματικές εκθέσεις πρέπει να εκτιμηθούν χρησιμοποιώντας εξωτερικές πληροφορίες, όπως JEM ή "ad hoc" έρευνες που διεξάγονται από τα τμήματα εργασιακής υγείας των τοπικών υγειονομικών αρχών. Στο OCM, οι καρκίνοι που προκαλούνται από την έκθεση στον αμίαντο παρουσιάζουν ιδιαίτερο ενδιαφέρον, δεδομένης της μεγάλης χρήσης υλικών αμιάντου στην Ιταλία, την έκταση της έκθεσης και τη δυνατότητα αποζημίωσης για ασθένειες που σχετίζονται με τον αμίαντο. Σύμφωνα με τον Διεθνή Οργανισμό Έρευνας για τον Καρκίνο (International Agency for Research on Cancer - IARC), οι καρκίνοι που προκαλούνται από την έκθεση στον αμίαντο περιλαμβάνουν καρκίνους του πνεύμονα, του λάρυγγα, των ωοθηκών και ενδεχομένως του φάρυγγα, του στομάχου και του παχέος εντέρου, καθώς και του μεσοθηλιώματος. Είναι εξαιρετικά σημαντικό, επομένως, η επινόηση μιας μεθόδου για την προκαταρκτική εκτίμηση της έκθεσης στον αμίαντο που εφαρμόζεται στον OCM. Τα δεδομένα του ReNaM παρέχουν μια πλούσια βάση δεδομένων για τις βιομηχανίες με ενδείξεις έκθεσης στον αμίαντο, η οποία θα μπορούσε να είναι χρήσιμη για το σκοπό αυτό. Όμως, οι πρώτες προσπάθειες χρήσης των πληροφοριών αυτών έδειξαν ότι οι πληροφορίες στη βάση δεδομένων ReNaM δεν μπορούσαν να χρησιμοποιηθούν άμεσα ως πρόσθετη είσοδος στη διαδικασία σύνδεσης, εξαιτίας έλλειψης τυποποίησης των πληροφοριών που αναφέρθηκαν κατά τη συνέντευξη από τις περιπτώσεις κακοήθους όγκου στο μεσοθηλίωμα (Malignant Mesothelioma-MM). Η παρούσα μελέτη αποσκοπεί στη βελτίωση της εκτίμησης της έκθεσης στον αμίαντο στο πλαίσιο της παρακολούθησης και ανακάλυψης των καρκίνων που σχετίζονται με την εργασία, χρησιμοποιώντας τις πληροφορίες σχετικά με την έκθεση στον αμίαντο που συλλέγονται στο ιταλικό μητρώο MM-ReNaM και την εφαρμογή τους στην αξιολόγηση της έκθεσης στον αμίαντο των άλλων περιπτώσεων καρκίνου. Στην παρούσα μελέτη, παρουσιάζεται ο σχεδιασμός και η εφαρμογή της μεθόδου που στόχευε (i) στον εντοπισμό των επιχειρήσεων με πιθανή έκθεση στον αμίαντο και (ii) στην εφαρμογή αυτού του καταλόγου επιχειρήσεων στις περιπτώσεις καρκίνου που εντοπίστηκαν μέσω της OCM παρακολούθησης.

Το έργο διεξήχθη ακολουθώντας τέσσερα στάδια: (1) οι περιπτώσεις MM που περιλαμβάνονται στη βάση δεδομένων ReNaM συνδέθηκαν με τη βάση δεδομένων INPS, προκειμένου να αποτυπωθούν τα αντίστοιχα ονόματα και τα αναγνωριστικά στοιχεία των επιχειρήσεων. (2) Τα ονόματα των επιχειρήσεων που αναφέρθηκαν για κάθε περίπτωση στη

συνέντευξη της ReNaM (αναφέρθηκαν με πιθανές ανακρίβειες κατά τη συνέντευξη των περιπτώσεων ΜΜ) συσχετίστηκαν με τις επωνυμίες των επιχειρήσεων που καταγράφηκαν στη βάση δεδομένων INPS. (3) Οι επιχειρήσεις με πιθανή έκθεση στον αμίαντο σύμφωνα με την αξιολόγηση της ReNaM απαριθμήθηκαν και ταυτοποιήθηκαν με τις επωνυμίες και τους κωδικούς των επιχειρήσεων INPS και την αξιολόγηση της έκθεσης στον αμίαντο. (4) Ο κατάλογος εφαρμόστηκε στα εργασιακά ιστορικά των περιπτώσεων καρκίνου από το OCM για την ένδειξη των επιχειρήσεων με πιθανή έκθεση στον αμίαντο. Πρέπει να θυμόμαστε πως για την κατανόηση του βήματος 4 το εργασιακό ιστορικό για τις περιπτώσεις καρκίνου από το OCM περιορίστηκε στις πληροφορίες από τη βάση δεδομένων INPS. Το Σχήμα 3.1 (Airoldi *et al.*, 2020) συνοψίζει τη ροή των διαφόρων αρχείων σύνδεσης των δραστηριοτήτων.

Σχήμα 3.1
Ροή αρχείων σύνδεσης των δραστηριοτήτων



Οι απαιτούμενες βάσεις δεδομένων ήταν: η βάση δεδομένων των ΜΜ περιπτώσεων (ReNaM) και η βάση δεδομένων των συνταξιοδοτικών εισφορών (INPS). Η βάση δεδομένων ReNaM παρείχε τα ονόματα και τις ημερομηνίες γέννησης των καταγεγραμμένων περιπτώσεων. Η βάση δεδομένων INPS περιλάμβανε όλες τις συνταξιοδοτικές εισφορές από ιδιωτικές επιχειρήσεις από το 1974 και μετά. Η βάση δεδομένων περιελάμβανε, για κάθε άτομο και συνταξιοδοτική εισφορά, το όνομα της κοινωνικής εγγραφής και τον κωδικό της επιχείρησης. Όμως, οι εισφορές του δημόσιου τομέα, της γεωργίας, στρατιωτικών δυνάμεων και αυτοαπασχολούμενων δεν περιλαμβάνονταν στη βάση δεδομένων INPS. Η σύνδεση πραγματοποιήθηκε ονομαστικά χρησιμοποιώντας τα ονόματα και τις ημερομηνίες γέννησης των περιπτώσεων ΜΜ που συνδέθηκαν με τη βάση δεδομένων INPS. Προκειμένου να μειωθεί το κόστος σύνδεσης, ο κατάλογος των ατόμων περιορίστηκε στις περιπτώσεις ΜΜ με πλήρη δημογραφικά στοιχεία και με στοιχεία των περιόδων εργασίας με έκθεση στον αμίαντο σύμφωνα με την αξιολόγηση ReNaM. Η μεταβλητή απόκρισης ήταν ο κατάλογος των επιχειρήσεων στις οποίες είχαν εργαστεί οι περιπτώσεις ΜΜ, σύμφωνα με το INPS.

Περιελάμβανε όλες τις επιχειρήσεις στις οποίες είχε εργαστεί μια περίπτωση ΜΜ και είχε καταβάλει συνταξιοδοτικές εισφορές, με τις ημερομηνίες εισόδου και απόλυσης. Το output αρχείο περιλάμβανε την εταιρική επωνυμία των επιχειρήσεων και τη διεύθυνση, τον κωδικό Φόρου Προστιθέμενης Αξίας (ΦΠΑ), τον κωδικό οικονομικής δραστηριότητας, καθώς και τα εργασιακά καθήκοντα του ατόμου σε μεγάλες κατηγορίες.

Τα δεδομένα που εισήχθησαν στο βήμα 2 ήταν τα αποτελέσματα του βήματος 1 και οι εγγραφές του εργασιακού ιστορικού των ΜΜ περιπτώσεων από το ReNaM. Η μεθοδολογία του ReNaM για την εκτίμηση της έκθεσης στον αμίαντο περιγράφηκε λεπτομερώς και συνοψίστηκε εδώ μόνο για την κατανόηση αυτού του βήματος της δραστηριότητας. Η βάση δεδομένων ReNaM περιλάμβανε πληροφορίες που αναφέρθηκαν με λέξεις κατά τη διάρκεια της συνέντευξης των περιπτώσεων ΜΜ, συμπεριλαμβανομένων των ονομάτων των επιχειρήσεων στις οποίες εργάζονταν. Όπως συμβαίνει πάντα στις συνεντεύξεις, υπήρχαν ανακρίβειες στη διατύπωση των ονομάτων των επιχειρήσεων, τα οποία συχνά αναφέρονταν διαφορετικά από τα εμπορικά τους ονόματα. Οι ειδικοί του χώρου της υγείας αξιολόγησαν το επίπεδο έκθεσης στον αμίαντο για κάθε επιχείρηση από τις περιγραφές των θέσεων εργασίας που συλλέχθηκαν κατά τη συνέντευξη, αλλά κράτησαν τα ονόματα των επιχειρήσεων χωρίς τυποποίηση. Για τους σκοπούς της έρευνας, έπρεπε να μεταφερθούν οι πληροφορίες σχετικά με την έκθεση στον αμίαντο που αξιολογήθηκαν από την ReNaM, στα ονόματα των επιχειρήσεων όπως στη βάση δεδομένων INPS. Χρειάστηκε να επινοηθεί και να εφαρμοστεί μια σύνθετη μεθοδολογία για να μεγιστοποιηθεί ο αριθμός των εγγραφών που αντιστοιχίζονται. Η κύρια πρόκληση ήταν η εφαρμογή προσεγγιστικής αντιστοίχισης που να αντιστοιχίζεται με κάθε όνομα επιχείρησης από το ReNaM, το αντίστοιχο όνομα από τον κατάλογο INPS. Για να συμβεί αυτό, πρώτα αξιολογήθηκαν και διορθώθηκαν τα δεδομένα, εφαρμόζοντας μια αναδρομική και επαναληπτική διαδικασία προεπεξεργασίας (text mining και text cleaning). Στη συνέχεια, συνδέθηκαν οι βάσεις δεδομένων χρησιμοποιώντας διαφορετικές προσεγγίσεις και τα αποτελέσματα στάλθηκαν σε διάφορα περιφερειακά κέντρα για έλεγχο από τοπικούς εμπειρογνώμονες. Αυτό το τελευταίο βήμα χρησιμοποιήθηκε για την αξιολόγηση και την παρακολούθηση της απόδοσης του αλγορίθμου, εξετάζοντας τα συνδεδεμένα και μη συνδεδεμένα δεδομένα. Επιπλέον, ελέγχθηκε ένα τυχαίο δείγμα εγγραφών για να εκτιμηθεί το ποσοστό των σωστών αντιστοιχίσεων. Η διαδικασία επαναλήφθηκε μετά τη λήψη των τροποποιήσεων από τις διάφορες περιοχές. Τελικά, δημιουργήθηκε ο τελικός κατάλογος των επιχειρήσεων που χρησιμοποιούν αμίαντο με την εταιρική τους επωνυμία και τον INPS κωδικό εγγραφής. Η διαδικασία εφαρμόστηκε χωριστά για κάθε περιφέρεια και τα επιμέρους αποτελέσματα που αποκτήθηκαν για κάθε μονάδα, χρησιμοποιήθηκαν για την αύξηση της ικανότητας και της επάρκειας του αλγορίθμου.

Η εξόρυξη κειμένου (text mining) βοήθησε να κατανοηθούν και να ανακαλυφθούν μοτίβα και ανωμαλίες των ονομάτων των επιχειρήσεων. Οι συμβολοσειρές με τα ονόματα των επιχειρήσεων χωρίστηκαν σε μεμονωμένες λέξεις με μια διαδικασία που ονομάζεται "tokenization". Ακόμη, πραγματοποιήθηκαν αναλύσεις με βάση την ομοιότητα των λέξεων, τη συχνότητα τους και τις διαφορές τους που παρουσιάστηκαν με κατάλληλους δείκτες. Οι όροι που τείνουν να συνυπάρχουν μαζί παρατηρήθηκαν επίσης με τη χρήση των bi-grams που βοηθούν στη διερεύνηση των ζευγών γειτονικών λέξεων.

Ο καθαρισμός του κειμένου πραγματοποιήθηκε με τη χρήση των πληροφοριών που παρείχε η εξόρυξη του κειμένου. Εκτελέστηκε προεπεξεργασία των δεδομένων του κειμένου, όπως η αφαίρεση των διαχωριστικών, των αριθμών και η μετατροπή του κειμένου σε κεφαλαία γράμματα.

Ορίστηκε ένας κατάλογος "stop words", δηλαδή κοινών όρων, οι οποίοι διαγράφηκαν από τις συμβολοσειρές. Τα δεδομένα διορθώθηκαν σύμφωνα με τους sound-like operators που εξέτασαν τις κλίσεις, τα συνώνυμα και το stemming που ήταν οι τροποποιήσεις μιας λέξης για να εκφράσουν διαφορετικές γραμματικές κατηγορίες όπως ο χρόνος, η πτώση, η φωνή, το πρόσωπο, ο αριθμός και το γένος. Μετά από αρκετές επαναλήψεις, τα δύο σύνολα δεδομένων ήταν έτοιμα για τη διαδικασία σύνδεσης "inexact matching". Η σύνδεση εγγραφών μεταξύ των ονομάτων επιχειρήσεων, όπως αναφέρεται στις βάσεις δεδομένων ReNaM και INPS, χρησιμοποιήθηκε για να προστεθεί η αξιολόγηση της έκθεσης στον αμίαντο από την ReNaM στις πληροφορίες του INPS. Ο αλγόριθμος σύνδεσης σχεδιάστηκε επιλέγοντας τις επιλογές που μεγιστοποιούσαν τον αριθμό των εγγραφών που συνδέονται εντός των περιφερειών. Η σύνδεση χρησιμοποίησε τις μεμονωμένες λέξεις (tokens) της επωνυμίας της επιχείρησης και όχι ολόκληρη τη συμβολοσειρά. Για το λόγο αυτό, χρησιμοποίηκε ο όρος "μη ακριβές ταίριασμα (inexact matching)". Πριν από τη σύνδεση, οι εγγραφές που ανέφεραν απασχόληση σε ορισμένες κατηγορίες, όπως ο γεωργικός τομέας, ο στρατός και οι πυροσβέστες, αποκλείστηκαν από τη βάση δεδομένων ReNaM επειδή οι κατηγορίες αυτές δεν περιλαμβάνονταν στη βάση δεδομένων INPS. Αρχικά, πραγματοποιήθηκε μια μη ακριβή αντιστοίχιση within-subject. Θεωρήθηκε ότι μία αντιστοίχιση είναι ορθή όταν τα δύο σύνολα δεδομένων είχαν τον ίδιο κωδικό περιπτώσεων και τουλάχιστον μία σημαντική λέξη της επωνυμίας της επιχείρησης. Στη συνέχεια, πραγματοποιήθηκε πλήρης ένωση χρησιμοποιώντας τις εγγραφές που δεν αντιστοιχίστηκαν στο προηγούμενο βήμα: Όλες οι υπόλοιπες εγγραφές από το ReNaM συνδυάστηκαν με όλες τις εγγραφές που ελήφθησαν από το INPS. Η διαδικασία ήταν χρονοβόρα και για να εφαρμοστεί στις μεγάλες βάσεις δεδομένων, χωρίστηκαν οι πίνακες σε υποομάδες των 10 γραμμών, επιλέγοντας μόνο τις εγγραφές που είχαν τουλάχιστον μία κοινή λέξη. Οι υπόλοιπες εγγραφές χρησιμοποιήθηκαν και επανελέγχονταν στο επόμενο βήμα. Επίσης, θεωρήθηκε μια αντιστοίχιση επιτυχής όταν 2 λέξεις του ονόματος της επιχείρησης ήταν ίσες. Η επιλογή της χρήσης 2 λέξεων ως κατώφλι σχετιζόταν με εμπειρικά κριτήρια: Η μία λέξη ήταν πολύ μικρή (υπέρβαση της ευαισθησίας- excess of sensitivity) και οι 3 λέξεις ήταν πολύ μεγάλες (υπέρβαση της ειδικότητας- excess of specificity). Σε ορισμένες περιπτώσεις, μετά την φάση της προεπεξεργασίας, οι συμβολοσειρές μειώθηκαν και αποτελούνταν από έναν μόνο όρο. Σε αυτό την περίπτωση, μια αντιστοίχιση θεωρήθηκε επιτυχής όταν η μοναδική λέξη που υπήρχε στη ReNaM βρέθηκε σε μια εγγραφή στην INPS. Τρίτον, στις υπόλοιπες μη συνδεδεμένες εγγραφές αναζητήθηκαν οι γνωστές εταιρείες που θα μπορούσαν να είναι καταχωρημένες με περισσότερα από ένα ονόματα ή με ένα ουσιαστικό που συνοδεύεται από άλλους όρους. Για παράδειγμα ήταν η Enel ή η Pirelli, μεγάλες επιχειρήσεις με μεγάλο αριθμό τμημάτων και υποομάδων. Για να αποφευχθεί η πιθανή απώλεια των εγγραφών, αναγκαστικά δημιουργήθηκε ένας κατάλογος με τα ονόματα των εταιρειών που ήταν γνωστά "εκ των προτέρων". Οι εναπομείνασες μη αντιστοιχισμένες εγγραφές στη ReNaM

αναδιοργανώθηκαν σε 3 ομάδες: "a priori", "γενικοί όροι", και "μη γενικός όρος", για να ελεγχθούν από τους περιφερειακούς εμπειρογνώμονες.

Το αποτέλεσμα των διαδικασιών συγχώνευσης ήταν ένας γενικός πίνακας ("λεξικό"), που περιελάμβανε όλες τις επιχειρήσεις με στοιχεία έκθεσης στον αμίαντο. Κάθε εγγραφή για την κάθε επιχείρηση περιελάμβανε: την εταιρική επωνυμία της επιχείρησης όπως προκύπτει από το INPS, τον αριθμό διοικητικής εγγραφής της, τον βαθμό έκθεσης στον αμίαντο σύμφωνα με τους εμπειρογνώμονες της ReNaM και την επωνυμία της επιχείρησης όπως αναφέρθηκε στις συνεντεύξεις της ReNaM. Οι διπλότυπες εγγραφές περιορίστηκαν σε μία εγγραφή, διατηρώντας την υψηλότερη βαθμολογία έκθεσης στον αμίαντο. Τα αποτελέσματα της διαδικασίας συγχώνευσης επικυρώθηκαν από περιφερειακούς εμπειρογνώμονες σε κάθε περιφερειακή μονάδα του ReNaM. Οι διευθύνσεις των επιχειρήσεων γεωκωδικοποιήθηκαν και οι συντεταγμένες γεωγραφικού πλάτους και μήκους χρησιμοποιήθηκαν για τη χωρική ανάλυση. Η εφαρμογή του γενικού πίνακα ("λεξικό") των επιχειρήσεων που εκτίθενται στον αμίαντο, πραγματοποιήθηκε για τις δύο ιταλικές περιφέρειες (Λάτσιο και Σικελία), σύμφωνα με τα στοιχεία του OCM που ήταν διαθέσιμα κατά τη στιγμή της έρευνας. Συμπεριλήφθηκαν όλοι οι τύποι καρκίνου που σχετίζονται ή πιθανώς σχετίζονται με τον αμίαντο, σύμφωνα με την IARC. Οι πληροφορίες που περιελάμβαναν για κάθε περίπτωση ήταν: η διάγνωση, η ημερομηνία διάγνωσης και το εργασιακό ιστορικό, όπως δόθηκε από το INPS (ονόματα επιχειρήσεων και κωδικοί). Το output ήταν οι περιπτώσεις με τουλάχιστον έναν αριθμό μητρώου επιχείρησης που υπήρχαν στο λεξικό, δηλαδή τα άτομα που εργάζονταν σε βιομηχανία που πιθανώς εκτέθηκε στον αμίαντο. Τα βήματα 1 έως 3 πραγματοποιήθηκαν με τη χρήση των πληροφοριών σχετικά με τα περιστατικά ΜΜ που συνέβησαν την περίοδο 1993-2012 σε 7 ιταλικές περιφέρειες ξεχωριστά (Πεδεμόντιο, Λομβαρδία, Τοσκάνη, Εμίλια-Ρομάνια, Λάτσιο, Απουλία και Σικελία). Το βήμα 4 αφορούσε μόνο δύο ιταλικές περιφέρειες που αντιστοιχούσαν στην επέκταση του προγράμματος OCM κατά τη διάρκεια της ανάλυσης: Χρησιμοποιήθηκε το εξιτήριο των νοσοκομείων του Λάτσιο κατά την περίοδο 2008-2015 και τα περιστατικά καρκίνου στην Ανατολική Σικελία κατά την περίοδο 2011-2014. Όλες οι αναλύσεις πραγματοποιήθηκαν με τη χρήση του SAS 9.3 (SAS Institute Inc., Cary, NC, ΗΠΑ), της έκδοσης 3.4.1 του R (R Core Team, Βιέννη, Αυστρία) και STATA 11 (StataCorp LLC, College Station, TX, ΗΠΑ). Η παρούσα μελέτη δεν περιελάμβανε καμία επαφή με τα άτομα και η χρήση προσωπικών πληροφοριών επιτρεπόταν από τον ιταλικό νόμο για την προστασία των εργαζομένων (D.Leg. 81/2008), επομένως, η έγκριση από την επιτροπή δεοντολογίας δεν απαιτήθηκε.

Για λόγους σαφήνειας, τα αποτελέσματα παρουσιάστηκαν χωριστά για τα διαφορετικά βήματα. Η βάση δεδομένων ReNaM περιελάμβανε περιπτώσεις ΜΜ που διαγνώστηκαν μεταξύ 1993 και 2012 στις επτά ιταλικές περιφέρειες που συμμετείχαν στην μελέτη. Στην παρούσα μελέτη, συμπεριλήφθηκαν 6057 περιπτώσεις: Οι υπόλοιπες περιπτώσεις αποκλείστηκαν επειδή δεν ανέφεραν πλήρη στοιχεία ταυτοποίησης ή ονόματα επιχειρήσεων ή δεν είχαν κανένα στοιχείο έκθεσης στον αμίαντο. Από αυτές, 4134 (68,25%) συνδέθηκαν επιτυχώς με την βάση δεδομένων INPS. Ο Πίνακας 3.1 (Airoldi *et al.,* 2020) παρουσιάζει τα αποτελέσματα ανά περιφέρεια. Ο πίνακας παρουσιάζει τον αριθμό των εγγραφών, που αντιστοιχούν για το ReNaM στον αριθμό των θέσεων απασχόλησης που αναφέρθηκαν κατά τη

συνέντευξη και για το INPS στον αριθμό των διαφορετικών επιχειρήσεων που έχουν καταγραφεί στη βάση δεδομένων. Ο αριθμός των εγγραφών ήταν μεγαλύτερος από τις περιπτώσεις, επειδή ένα άτομο θα μπορούσε να έχει εργαστεί σε περισσότερες από μία επιχειρήσεις. Οι περιφέρειες με τον υψηλότερο αριθμό περιπτώσεων καρκίνου ήταν η Λομβαρδία (2276) και το Πεδεμόντιο (1104), ενώ για τη Σικελία και τη Lazio παρατηρήθηκαν λιγότερα άτομα με ΜΜ, 256 και 289, αντίστοιχα, που αντιστοιχούν στις διαφορετικές περιπτώσεις του ΜΜ και την επικράτηση του επαγγέλματος σε κλάδους. Ο Πίνακας 3.2 (Airoldi *et al.*, 2020) περιγράφει τις περιπτώσεις ΜΜ ανάλογα με το επίπεδο της έκθεσης στον αμίαντο που αξιολογήθηκε από την ReNaM.

Πίνακας 3.1
Αποτελέσματα ανά περιφέρεια.

Region	ReNaM Cases	INPS Linked	ReNaM Records	INPS Records
Piedmont	1104	742 (67.21%)	2490	1885
Lombardy	2276	1595 (70.08%)	3965	4083
Emilia-Romagna	891	634 (71.16%)	1367	1906
Tuscany	906	568 (62.69%)	3465	1600
Lazio	289	189 (65.40%)	634	613
Apulia	335	208 (62.09%)	440	642
Sicily	256	198 (77.34%)	440	873
Total	**6057**	**4134 (68.25%)**	**12,801**	**11,602**

Πίνακας 3.2
Κατανομή των επιπέδων έκθεσης στον αμίαντο ανά περιοχή.

Region	Occupational Asbestos Exposure						
	Total	Certain		Probable		Possible	
	n	*n*	%	*n*	%	*n*	%
Piedmont	2490	1480	59.44	208	8.35	802	32.21
Lombardy	3965	2797	70.54	274	6.91	894	22.55
Emilia-Romagna	1367	883	64.59	239	17.48	245	17.92
Tuscany	3465	2464	71.11	388	11.2	613	17.69
Lazio	634	152	23.97	176	27.76	306	48.26
Apulia	440	204	46.36	96	21.82	140	31.82
Sicily	440	177	40.23	119	27.05	144	32.73
Total	**12,801**	**8157**	**63.72**	**1500**	**11.72**	**3144**	**24.56**

Τα δεδομένα που εισήχθησαν στο βήμα 2 ήταν τα αρχεία του ιστορικού εργασίας των περιπτώσεων ΜΜ. Τα αποτελέσματα της διαδικασίας που διεξήχθη στο βήμα 2 για την ανάλυση των καταλόγων των επιχειρήσεων παρουσιάζονται στον Πίνακα 3.3 (Airoldi *et al.* 2020). Στην βάση δεδομέων ReNaM, οι εγγραφές που αναφέρονται στον γεωργικό, στρατιωτικό, και πυροσβεστικό τομέα αποκλείστηκαν "εκ των προτέρων", επειδή οι εν λόγω εργασιακοί τομείς δεν περιλαμβάνονταν στο INPS και, ως εκ τούτου, η σύνδεση ήταν αδύνατη. Αυτές οι αποκλεισμένες εγγραφές κυμαίνονταν από 3% έως 25% στην βάση δεδομένων ReNaM, ανάλογα με την εξεταζόμενη περιοχή. Οι υπόλοιπες εγγραφές αναλύθηκαν και θεωρήθηκε επιτυχής η αντιστοίχιση όταν οι εγγραφές μοιράζονταν τουλάχιστον μία λέξη στα ίδια θέματα, τουλάχιστον δύο λέξεις και μία λέξη στη συμβολοσειρά μήκους ίση με ένα.

Πίνακας 3.3

Ο αριθμός των εγγραφών ReNaM και τα αποτελέσματα της σύνδεσης των εγγραφών ανά περιοχή

Region	ReNaM *n*	Excluded *n* (%)	Matched (By Matching Step)				Matched Total *n* (%)	Non-Matched *n* (%)
			1	2	3	4		
			n (%)	*n* (%)	*n* (%)	*n* (%)		
Piedmont	2490	185 (7.43)	861 (34.58)	156 (6.27)	168 (6.75)	119 (4.78)	1304 (52.37)	1001 (40.2)
Lombardy	3965	128 (3.23)	1324 (33.39)	271 (6.83)	106 (2.67)	95 (2.4)	1796 (45.3)	2041 (51.48)
Emilia-Romagna	1367	167 (12.22)	452 (33.07)	71 (5.19)	106 (7.75)	3 (0.22)	632 (46.23)	568 (41.55)
Tuscany	3465	364 (10.51)	755 (21.79)	117 (3.38)	146 (4.21)	21 (0.61)	1039 (29.99)	2062 (59.51)
Lazio	634	81 (12.78)	161 (25.39)	12 (1.89)	2 (0.32)	24 (3.79)	199 (31.39)	354 (55.84)
Apulia	440	112 (25.45)	117 (26.59)	7 (1.59)	22 (5)	9 (2.05)	155 (35.23)	173 (39.32)
Sicily	440	32 (7.27)	163 (37.05)	23 (5.23)	8 (1.82)	8 (1.82)	202 (45.91)	206 (46.82)
Total	**12,801**	**1069 (8.35)**	**3833 (29.94)**	**657 (5.13)**	**558 (4.36)**	**279 (2.18)**	**5327 (41.61)**	**6405 (50.04)**

Οι συνολικές αντιστοιχισμένες εγγραφές ήταν 5327 (41,61%), αλλά παρατηρήθηκε τεράστια διακύμανση μεταξύ των περιφερειών. Καλύτερη επίδοση παρατηρήθηκε στο Πεδεμόντιο (52,37%) και ακολούθησε η Emilia Romagna (46,23%), Σικελία (45,91%) και Λομβαρδία (45,30%), ενώ το ποσοστό των εγγραφών που συνδέθηκαν ήταν χαμηλότερο στη Τοσκάνη (29,99%), το Λάτσιο (31,39%) και την Απουλία (35,23%). Εάν δεν λάβουμε υπόψη τους "εκ των προτέρων" αποκλεισμούς στον παρονομαστή, το ποσοστό των επιτυχημένων αντιστοιχίσεων ήταν υψηλότερο, ιδίως για την Απουλία (47,26%), Emilia Romagna (52,67%), και Lazio (35,99%) (τα αποτελέσματα δεν καταγράφονται σε πίνακες).

Ο αριθμός των διαφορετικών επιχειρήσεων που προέκυψαν από τη διαδικασία σύνδεσης παρουσιάζεται στον Πίνακα 3.4 (Airoldi *et al.*, 2020), ανά περιφέρεια και επίπεδο έκθεσης στον αμίαντο. Επίσης, κάθε φορά που υπήρχαν διαφορετικές αξιολογήσεις για την ίδια επιχείρηση, θεωρήσαμε την υψηλότερη αξιολόγηση, και όταν η ίδια επιχείρηση εμφανιζόταν σε διαφορετικές περιοχές, τη λάβαμε υπόψη μόνο μία φορά. Οι περιφέρειες με περισσότερες περιπτώσεις ΜΜ συνεισέφεραν με περισσότερες επιχειρήσεις, και ιδίως η Λομβαρδία συνέβαλε με το 40% (1080 επιχειρήσεις) του εθνικού συνόλου. Ο συνολικός αριθμός των μοναδικών επιχειρήσεων που εντοπίστηκαν ήταν 2606: 1826 (70,07%) με ορισμένες, 222 (8,52%) με πιθανή και 558 (21,41%) με πιθανή έκθεση στον αμίαντο. Ο κατάλογος με τα ονόματα των εταιριών στο INPS αποτέλεσε το λεγόμενο "λεξικό", που χρησιμοποιήθηκε για την αξιολόγηση της έκθεσης στον αμίαντο σε επόμενο βήμα της διαδικασίας.

Πίνακας 3.4

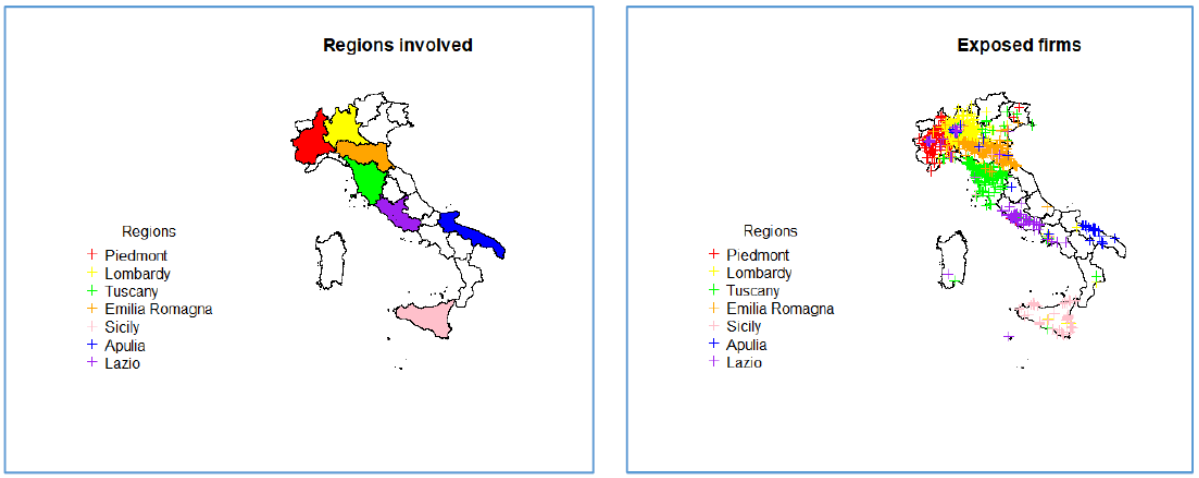
Ο αριθμός των διαφορετικών επιχειρήσεων.

Regions	Matched Records *n*	Firms *n*	Asbestos Exposure Rating of the Firm		
			Certain	Probable	Possible
			n (%)	*n* (%)	*n* (%)
Piedmont	1304	429	267 (62.24)	25 (5.83)	137 (31.93)
Lombardy	1796	1080	872 (80.74)	41 (3.80)	167 (15.46)
Emilia-Romagna	632	370	259 (70.00)	38 (10.27)	73 (19.73)
Tuscany	1039	537	420 (78.21)	45 (8.38)	72 (13.41)
Lazio	199	129	33 (25.58)	50 (38.76)	46 (35.66)
Apulia	155	99	50 (50.51)	10 (10.10)	39 (39.39)
Sicily	202	142	57 (40.14)	34 (23.94)	51 (35.92)
Total	**5208**	**2606**	**1826 (70.07)**	**222 (8.52)**	**558 (21.41)**

Ένα από τα πλεονεκτήματα αυτής της διαδικασίας είναι η παροχή καταλόγου των "επιχειρήσεων που εκτίθενται στον αμίαντο", ο οποίος δεν περιορίζεται στην εμπειρία των περιφερειακών εμπειρογνωμόνων. Στο Σχήμα 3.2 (Airoldi *et al.*, 2020), αναφέρονται οι δύο χάρτες της Ιταλίας με τα περιφερειακά όρια. Στον αριστερό χάρτη, επισημαίνονται οι εμπλεκόμενες περιφέρειες, ενώ στον δεξιό χάρτη η χωρική κατανομή των εκτεθειμένων επιχειρήσεων, αναφέρεται ξεχωριστά η περιοχή προέλευσης των ΜΜ.

Σχήμα 3.2

Ιταλικός χάρτης των περιοχών που συμμετείχαν στη μελέτη



Οι περιπτώσεις καρκίνου που περιλαμβάνονται σε αυτή τη δοκιμή για τον προκαταρκτικό προσδιορισμό των εκθέσεων σε αμίαντο στις επιχειρήσεις στο σύστημα OCM ήταν 35.010, 9848 (28,13%) από τη Σικελία και 25.162 (71,87%) από το Λάτσιο, που επιλέχθηκαν από το τοπικό μητρώο καρκίνου του πληθυσμού (Σικελία) ή από τις βάσεις δεδομένων των νοσοκομειακών εξιτηρίων (Λάτσιο).

Ο Πίνακας 3.5 (Airoldi *et al.*, 2020) αναφέρει την κατανομή ανά τύπο κακοήθους όγκου. Στην περίπτωση πολλαπλών κακοηθειών στο ίδιο άτομο, επιλέχθηκε μόνο η πρώτη κακοήθης νεοπλασία.

Πίνακας 3.5

Κατανομή των κακοήθων νεοπλασιών στα δεδομένα παρακολούθησης (Λάτσιο,Σικελία)

Malignant Neoplasm	Region			
	Lazio (n = 25,162)		Sicily (n = 9848)	
	n Cases	%	n Cases	%
Colorectal	9373	37.25	3812	38.71
Pharynx	605	2.4	137	1.39
Larynx	1288	5.12	573	5.82
Ovary	1490	5.92	464	4.71
Pleura	221	0.88	196	1.99
Lung	9590	38.11	3650	37.06
Nasopharynx	196	0.78	140	1.42
Stomach	2399	9.53	876	8.9

Οι περιπτώσεις που είχαν εργαστεί σε επιχειρήσεις, που επιλέχθηκαν, με έκθεση στον αμίαντο ήταν 1454 και 5092, από τη Σικελία και το Lazio, αντίστοιχα. Έδειξε ότι, εξαιρουμένων των κακοηθειών του υπεζωκότα, το 18,5% (6400) των περιπτώσεων καρκίνου που εξήχθησαν για το σύστημα OCM θα μπορούσε να σχετίζεται με την εργασιακή έκθεση στον αμίαντο (Πίνακας 3.6 - Airoldi *et al.*, 2020).

Πίνακας 3.6
Κατανομή των κακοήθων νεοπλασιών στα δεδομένα OCM (Λάτσιο,Σικελία)

Malignant Neoplasm	Region			
	Lazio Cases		Sicily Cases	
	n	Matched	*n*	Matched
Colorectal	9373	2020	3812	579
Pharynx	605	113	137	17
Larynx	1288	248	573	81
Ovary	1490	169	464	28
Pleura	221	83	196	53
Lung	9590	1953	3650	561
Nasopharynx	196	35	140	12
Stomach	2399	471	876	123
Total	**25,162**	**5092**	**9848**	**1454**

Περίπου το 50% των αντιστοιχισμένων εγγραφών ταυτοποιήθηκαν, χρησιμοποιώντας τις πληροφορίες λεξικού που ελήφθησαν από άλλες περιοχές: 1608 (51,59%) για τη Σικελία και 3463 (46,56%) για το Λάτσιο (Πίνακας 3.7 - Airoldi *et al.*, 2020).

Πίνακας 3.7
Αριθμός καταγραφών και περιπτώσεων OCM για τη Σικελία και το Λάτσιο.

Region	OCM		Matched			
	Records	Cases	Records			Cases
	n	*n*	All *n*	In Region *n* (%)	Out Region *n* (%)	*n*
Lazio	89,274	25,162	7437	3974 (53.44)	3463 (46.56)	5092
Sicily	47,602	9848	3117	1509 (48.41)	1608 (51.59)	1454
Total	**136,876**	**35,010**	**10,554**	**5483 (51.95)**	**5071 (48.05)**	**6546**

Η έκθεση σε ίνες αμιάντου προκαλεί μια σειρά από κακοήθειες και, ως εκ τούτου, είναι ιδιαίτερα σημαντικό για κάθε σύστημα ανίχνευσης του καρκίνου που σχετίζεται με την εργασία. Η επιτήρηση της υγείας του πληθυσμού που σχετίζεται με ασθένειες του αμιάντου έχουν διάφορους στόχους, όπως: (i) την ανίχνευση και την κοινοποίηση περιστατικών- (ii) την εκτίμηση αξιόπιστων επιδημιολογικών στοιχείων με τον εντοπισμό κρουσμάτων της νόσου- (iii) τον εντοπισμό πιθανών πηγών μόλυνσης που εξακολουθούν να υπάρχουν και την πρόληψη της έκθεσης στον αμίαντο, (iv) τον σχεδιασμό και την εφαρμογή πολιτικών δημόσιας υγείας- (v) την υποστήριξη της αποτελεσματικότητας των ασφαλιστικών συστημάτων και την αξιολόγηση των αποζημιώσεων, (vi) τη διάδοση των αποτελεσμάτων και την τεκμηρίωση του αντικτύπου μιας παρέμβασης. Η παρούσα μελέτη είχε ως στόχο να συμβάλει στην παρακολούθηση στην Ιταλία του καρκίνου που σχετίζεται με την εργασία με τον εντοπισμό

των επιχειρήσεων που εκτίθενται στον αμίαντο και τον εντοπισμό των περιπτώσεων καρκίνου που είχαν εργαστεί σε επιχειρήσεις που εκτέθηκαν στον αμίαντο. Η διερεύνηση της αιτιολογίας των περιστατικών που σχετίζονται με τον αμίαντο απαιτεί τεράστιους πόρους, δεδομένου του σχετικά χαμηλού αποδιδόμενου κλάσματος και του μεγάλου αριθμού. Οι στόχοι επιτεύχθηκαν με εγγραφές που συνδέει πληροφορίες από το ReNaM και το INPS μέσω μιας σύνθετης μεθοδολογίας που βασίζεται σε επεξεργασία πληροφοριών κειμένου και πιθανολογική αντιστοίχιση. Ένα από τα αποτελέσματα ήταν η προετοιμασία ενός καταλόγου επιχειρήσεων με πιθανή έκθεση σε αμίαντο στο παρελθόν, έτοιμου για χρήση ως εργαλείο διαλογής σε νέες ερευνητικές μελέτες. Ο κατάλογος αυτός εφαρμόστηκε στα πρώτα διαθέσιμα δεδομένα από το OCM, το σύστημα παρακολούθησης για τον εργασιακό καρκίνο που βρίσκεται υπό προετοιμασία. Η έρευνα χρησιμοποιεί σε μεγάλο βαθμό τη σύνδεση αρχείων, μια διαδικασία που γίνεται όλο και πιο συνηθισμένη στην στατιστική και στην ακαδημαϊκή έρευνα. Η σύνδεση αρχείων καθιστά δυνατό τον συνδυασμό δεδομένων από διαφορετικές πηγές για να απαντηθούν ερευνητικά ερωτήματα που είναι πολύ δύσκολο να απαντηθούν με τη χρήση δεδομένων από μία μόνο πηγή. Σε πολλές περιπτώσεις, η σύνδεση εγγραφών είναι ένας αποτελεσματικός τρόπος συλλογής δεδομένων και μπορεί να μειώσει την ταλαιπωρία της υποβολής ευαίσθητων ερωτήσεων. Η σύνδεση εγγραφών χρησιμοποιείται συνήθως στις σκανδιναβικές χώρες για την αξιολόγηση του κινδύνου του καρκίνου, ακόμη και σε πολύ περίπλοκους σχεδιασμούς μελετών. Η χρήση των συστημάτων παρακολούθησης των δεδομένων που συλλέγονται από τα θεσμικά οργανά αποτελεί τόσο πλεονέκτημα όσο και μειονέκτημα. Από τη μία πλευρά, μειώνει τη διάρκεια και το κόστος της μελέτης και επιτρέπει μια τυποποιημένη ταξινόμηση της νόσου, του εργασιακού ιστορικού και της εκτίμησης της έκθεσης. Από την άλλη πλευρά, οι βάσεις δεδομένων που δεν δημιουργήθηκαν για επιστημονικούς ή ερευνητικούς σκοπούς συχνά στερούνται κρίσιμων πληροφοριών. Συγκεκριμένα, η βάση δεδομένων INPS δεν περιλαμβάνει αρκετές λεπτομέρειες σχετικά με την επαγγελματική έκθεση. Η ReNaM χρησιμοποιήθηκε μόνο για τις περιπτώσεις με βέβαιη, πιθανή και πιθανή εργασιακή έκθεση όπως αξιολογήθηκαν από τα δεδομένα των συνεντεύξεων. Η βάση δεδομένων INPS απέκλεισε ορισμένους σημαντικούς εργασιακούς τομείς (γεωργία, δημόσια απασχόληση, ένοπλες δυνάμεις) και τους αυτοαπασχολούμενους, κυρίως τεχνίτες και καταστηματάρχες. Επιπλέον, περιλαμβάνει μόνο εισφορές από το 1974 και μετά. Η προεπεξεργασία των δεδομένων ήταν απαραίτητη για τη βελτιστοποίηση των πληροφοριών στις διοικητικές βάσεις δεδομένων πριν από τη σύνδεση. Η εφαρμογή τεχνικών εξόρυξης κειμένου και καθαρισμού κειμένου υιοθετήθηκε στις διάφορες επιτυχημένες βιοϊατρικές εφαρμογές τα τελευταία χρόνια. Εφαρμόστηκαν για την εξαγωγή ουσιαστικών πληροφοριών από τα ονόματα των επιχειρήσεων. Έγινε μεγάλη προσπάθεια για την προετοιμασία του καταλόγου των "stop words" και για τη διόρθωση και απλούστευση των λέξεων. Διάφορες διαδικασίες και αλγόριθμοι ήταν διαθέσιμα στο παρελθόν και ήταν συνηθισμένα και υλοποιήθηκαν στο κύριο στατιστικό λογισμικό, αλλά η πλειονότητα τους βασίζονταν στην αγγλική γλώσσα. Επομένως, έπρεπε να αναπτυχθούν αλγόριθμοι για την ιταλική γλώσσα. Αυτοί οι αλγόριθμοι είναι διαθέσιμοι για χρήση σε άλλα πλαίσια ή με άλλες γλώσσες.

Η κύρια πρόκληση στην παρούσα μελέτη ήταν ο ορισμός του κλειδιού σύνδεσης μεταξύ των ονομάτων των επιχειρήσεων που αναφέρονται με διαφορετική ακρίβεια στις δύο βάσεις δεδομένων, πιο απλά στις συνεντεύξεις της ReNaM και πιο αυστηρά στο INPS. Απορρίφθηκαν οι ντετερμινιστικές μέθοδοι σύνδεσης επειδή βασίζονται στην παρουσία κοινών χαρακτηριστικών μοναδικής ταυτοποίησης σε όλες τις πηγές. Προτιμήθηκε η πιθανολογική προσέγγιση χρησιμοποιώντας μη μοναδικά χαρακτηριστικά και υπολογίζοντας δείκτες ομοιότητας για συγκρίσεις ανά ζεύγη. Η ποιότητα των διαδικασιών σύνδεσης είναι δύσκολο να προσδιοριστεί και αυτό αποτελεί μείζον ζήτημα στη σύνδεση αρχείων. Κατά την ανάλυση της διαδικασίας και των αποτελεσμάτων της σύνδεσης των εγγραφών, τόσο οι χαμένες συνδέσεις (ψευδώς αρνητικές) όσο και οι ψευδείς συνδέσεις (ψευδώς θετικές) πρέπει να αντιμετωπιστούν. Οι αποτυχημένοι σύνδεσμοι είναι "ψευδώς αρνητικοί", που αντιστοιχούν σε απουσία "σημάτων" έκθεσης στον αμίαντο: οι αντίστοιχες επιχειρήσεις δεν θα συμπεριληφθούν στον κατάλογο των επιχειρήσεων που εκτίθενται στον αμίαντο. Καθώς η διαδικασία είναι επαναληπτική, το σφάλμα αυτό αναμένεται να μειωθεί με την μελλοντική επέκταση της διαδικασίας. Οι επιχειρήσεις που υποδεικνύονται από τη σύνσεση αλλά δεν εκτίθενται είναι οι "ψευδώς θετικές"και αποτελούν σήματα που μπορούν να διορθωθούν με την αναθεώρηση του τοπικού εμπειρογνώμονα. Ένας άλλος περιορισμός αφορά τις πολύ μεγάλες επιχειρήσεις με διαφορετικά εργοστάσια που δεν αναγνωρίζονται χωριστά από τη σύνδεση. Για όλους αυτούς τους λόγους, η διαδικασία πρέπει να θεωρηθεί μόνο ένα εργαλείο διαλογής με πληροφορίες που πρέπει να επικυρωθούν από τους τοπικούς εμπειρογνώμονες που αξιολογούν το πραγματικό εργασιακό ιστορικό των περιπτώσεων. Ο αλγόριθμος παρήγαγε έναν κατάλογο επιχειρήσεων με πιθανή έκθεση στον αμίαντο που συνδέθηκε με τα εργασιακά ιστορικά των περιπτώσεων καρκίνου που παρέχονται από το σύστημα OCM, για την ανίχνευση της συσχέτισής τους με την έκθεση στον αμίαντο. Τα αποτελέσματα έδειξαν ότι στο 18,5% των επιχειρήσεων όπου υπήρχαν περιπτώσεις του καρκίνου εκτός από τον ΜΜ, είχαν ένδειξη πιθανής έκθεσης σε αμίαντο. Επίσης, περίπου το 50% των συνδέσμων αφορούσε επιχειρήσεις σε περιοχές διαφορετικές από την περιοχή της κατοικίας του ατόμου. Η γραφική κατανομή των διευθύνσεων των επιχειρήσεων είναι ένα αποτελεσματικό εργαλείο για τη διερεύνηση αυτού του φαινομένου πτυχή, κάτι το οποίο θα είναι επίσης χρήσιμο για τη βελτίωση των πληροφοριών των συνεντεύξεων στο ReNaM. Η διαδικασία εξοικονομεί κόστος και μπορεί να επαναλαμβάνεται κάθε φορά που συγκεντρώνονται νέες πληροφορίες. Δεν είναι γνωστό αν υπάρχουν παρόμοιες δραστηριότητες που χρησιμοποιούν τις πληροφορίες από τα μητρώα μεσοθηλιώματος για τον εντοπισμό της έκθεσης στον αμίαντο άλλων τύπων κακοήθων όγκων. Η πρώτη επέκταση που μπορεί να προγραμματιστεί είναι η πανεθνική επέκταση του έργου, που τώρα περιορίζεται σε επτά περιφέρειες, και η αξιολόγηση της έκθεσης στον αμίαντο για το εθνικό σύστημα OCM (Airoldi *et al.,* 2020).

3.2 Παράγοντες κινδύνου που σχετίζονται με τα αποτελέσματα της θεραπείας rt-pa σε ασθενείς με οξύ ισχαιμικό εγκεφαλικό επεισόδιο

Στην έρευνα των Tseng *et.al* (2020) αναφέρεται ότι το ισχαιμικό εγκεφαλικό επεισόδιο είναι ο πιο κοινός τύπος εγκεφαλικού επεισοδίου και ευθύνεται για το 87% περίπου του συνόλου των εγκεφαλικών επεισοδίων παγκοσμίως, όπου μπορεί να προκαλέσει απώλεια λειτουργιών, όπως η ομιλία, η κίνηση και η ανάγνωση. Στην Ταϊβάν, σύμφωνα με τα στατιστικά στοιχεία του Υπουργείου Υγείας και Πρόνοιας (Ministry of Health and Welfare - MOHW), το εγκεφαλικό επεισόδιο είναι η τέταρτη υψηλότερη αιτία θανάτου, και παρόμοια με τα στοιχεία που δημοσίευσε η Αμερικανική Καρδιολογική Εταιρεία, περίπου το 70%-80% των ασθενών με εγκεφαλικό επεισόδιο είχαν ισχαιμικό εγκεφαλικό επεισόδιο. Η συνιστώμενη θεραπεία για τους επιλέξιμους ασθενείς με οξύ ισχαιμικό εγκεφαλικό επεισόδιο είναι η θρομβολυτική θεραπεία με χρήση ανασυνδυασμένου ενεργοποιητή πλασμινογόνου ιστικού τύπου (rt-PA) για τη διάλυση των θρόμβων του αίματος, ο οποίος εγκρίθηκε από τον Οργανισμό Τροφίμων και Φαρμάκων των ΗΠΑ το 1996 για ενδοφλέβια χρήση εντός 3 ωρών από το εγκεφαλικό επεισόδιο. Σύμφωνα με τα στοιχεία, οι κατευθυντήριες οδηγίες περιγράφουν πλήρως την αξιολόγηση και τη θεραπεία με ενδοφλέβιες θεραπείες rt-PA. Για παράδειγμα, οι κατευθυντήριες οδηγίες συνιστούν ενδοφλέβια χορήγηση αλτεπλάσης για επιλεγμένους ασθενείς που μπορούν να αντιμετωπιστούν εντός 3 ωρών ή 3-4,5 ωρών από την έναρξη των συμπτωμάτων του ισχαιμικού εγκεφαλικού επεισοδίου. Οι γιατροί θα πρέπει να επανεξετάζουν τα κριτήρια χρήσης, όπως η αρτηριακή πίεση να είναι <185/110 mmHg και αρχικά επίπεδα γλυκόζης να είναι >50 mg/dL για να καθορίσουν την επιλεξιμότητα των ασθενών. Η χρήση της rt-PA περιορίζεται επίσης από σημαντικές αντενδείξεις, συμπεριλαμβανομένης της πηκτικότητας, της πρόσφατης χειρουργικής επέμβασης ή του εγκεφαλικού επεισοδίου ή της κρανιοεγκεφαλική κάκωσης εντός των τελευταίων 3 μηνών. Προηγούμενες μελέτες δείχνουν ότι η θεραπεία με rt-PA μπορεί να βελτιώσει αποτελεσματικά τα νευρολογικά ελλείμματα σε ασθενείς με οξύ ισχαιμικό εγκεφαλικό επεισόδιο. Η πρώιμη θεραπεία με rt-PA προάγει την ανεξαρτησία και αυξάνει τα προσαρμοσμένα στην ποιότητα έτη ζωής (QALY) για τα θύματα εγκεφαλικού επεισοδίου. Η συνταγογράφηση του rt-PA μεταξύ 3 και 4,5 ωρών μετά την έναρξη βελτιώνει την κλινική έκβαση και τη λειτουργική αποκατάσταση σε ασθενείς με οξύ ισχαιμικό εγκεφαλικό επεισόδιο. Για την αξιολόγηση της μακροπρόθεσμης έκβασης, μια μελέτη έδειξε ότι οι ασθενείς με οξύ ισχαιμικό εγκεφαλικό επεισόδιο που έλαβαν θεραπεία με rt-PA εμφανίζουν μεγαλύτερη επιβίωση, καθυστερημένη επανεισαγωγή στο νοσοκομείο και μικρότερο χρόνο έως την ανεξαρτησία. Επιπλέον, οι ασθενείς με υψηλό κίνδυνο συμπτωματικής ενδοκρανιακής αιμορραγίας εξακολουθούν να επωφελούνται από την αλτεπλάση. Εκτός από το χρόνο μέχρι τη θεραπεία, η σοβαρότητα του εγκεφαλικού επεισοδίου και ο υποτύπος του εγκεφαλικού επεισοδίου μπορεί επίσης να επηρεάζουν το όφελος από τη θεραπεία με rt-PA. Ωστόσο, η θεραπεία με rt-PA απαιτεί προσεκτική εξέταση τόσο των κινδύνων όσο και των οφελών, καθώς η αιμορραγία είναι η πιο συχνή παρενέργεια, η οποία μπορεί να προκαλέσει την ελλιπή χρήση του rt-PA παρά τις συστάσεις των κατευθυντήριων οδηγιών. Παρά το χαμηλό ποσοστό χρήσης, μία έρευνα κατέληξε στο συμπέρασμα ότι τα

μεγαλύτερα νοσοκομεία είχαν περισσότερες πιθανότητες να χορηγήσουν rt-PA. Περαιτέρω προσπάθειες για βελτίωση της κατάλληλης χορήγησης του rtPA θα πρέπει να ενθαρρυνθούν. Προηγούμενες μελέτες χρησιμοποίησαν ηλεκτρονικά ιατρικά αρχεία για την εκτίμηση των παραγόντων κινδύνου που σχετίζονται με την αιμορραγία μετά τη θεραπεία με rt-PA. Οι ασθενείς που διαγνώστηκαν με υπερλιπιδαιμία, καρδιοεμβολή, σοβαρό εγκεφαλικό επεισόδιο ή προχωρημένη ηλικία διατρέχουν υψηλότερο κίνδυνο αιμορραγίας και η χρήση του rt-PA θα πρέπει να εξετάζεται προσεκτικά. Υπό την προϋπόθεση ότι όλοι οι ασθενείς είναι επιλέξιμοι για τη λήψη θεραπείας rt-PA με βάση τις κατευθυντήριες γραμμές, μόνο λίγες μελέτες χρησιμοποίησαν τόσο τα αποτελέσματα των εργαστηριακών εξετάσεων όσο και τα δεδομένα του ιατρικού ιστορικού για την αξιολόγηση των αποτελεσμάτων της θεραπείας. Ωστόσο, η χρήση τόσο των αποτελεσμάτων των εργαστηριακών εξετάσεων όσο και των δεδομένων του ιατρικού ιστορικού για τον εντοπισμό των παραγόντων κινδύνου που οδηγούν σε κακές εκβάσεις μετά από θεραπεία με rt-PA στον γενικό πληθυσμό εξακολουθεί να πρέπει να διερευνηθούν. Για τους ασθενείς στους οποίους χορηγήθηκε rt-PA μετά την έναρξη του εγκεφαλικού επεισοδίου, στόχος της παρούσας μελέτης ήταν να διερευνηθουν οι γενικοί παράγοντες κινδύνου που σχετίζονται με την κακή έκβαση της θεραπείας με rt-PA. Συνεπώς, η παρούσα μελέτη πραγματοποίησε ανάλυση για ασθενείς σε ομάδες ευνοϊκής και κακής έκβασης για να προσδιοριστεί ποιοι παράγοντες μπορεί να σχετίζονται με κακές ενδονοσοκομειακές εκβάσεις, κυρίως με τον ενδονοσοκομειακό θάνατο, την παραμονή στη μονάδα εντατικής θεραπείας (ΜΕΘ) και την παρατεταμένη διάρκεια παραμονής στο νοσοκομείο (LOS).

Η παρούσα μελέτη διεξήχθη με τη χρήση των συγκεντρωτικών ηλεκτρονικών ιατρικών αρχείων (EMR) από το Stroke Registry of the Chang-Gung Healthcare System (SRICHS) και την ερευνητική βάση δεδομένων Chang Gung (CGRD) από τα νοσοκομεία Chang Gung Memorial Hospitals (CGMHs), τη μεγαλύτερη ομάδα παρόχων υγειονομικής περίθαλψης στην Ταϊβάν. Αναλύθηκαν τα EMR των ασθενών που διαγνώστηκαν με οξύ ισχαιμικό εγκεφαλικό επεισόδιο στο CGMHs μεταξύ 2006 και 2016, συμπεριλαμβανομένων 7 παραρτημάτων των CGMHs που βρίσκονται στις πόλεις Linkou, Taipei, Taoyuan, Keelung, Yunlin, Chiayi και Kaohsiung από τη βόρεια έως τη νότια Ταϊβάν. Τα ιατρικά ιστορικά καθορίστηκαν από το ιστορικό διάγνωσης με βάση τους κωδικούς της διεθνούς ταξινόμησης ασθενειών (ICD) και εργαστηριακών, χρησιμοποιήθηκαν ως μεταβλητές ενδιαφέροντος για τον προσδιορισμό των παραγόντων κινδύνου. Το Chang Gung Medical Foundation Institutional Review Board ενέκρινε την παρούσα μελέτη (IRB no. 107-1113C) και χορήγησε απαλλαγές για τη συγκατάθεση των ασθενών.

Η παρούσα μελέτη περιελάμβανε ασθενείς που είχαν επισκεφθεί το τμήμα επειγόντων περιστατικών για μία από τις τρεις κορυφαίες διαγνώσεις που σχετίζονται με το εγκεφαλικό επεισόδιο (κωδικοί ICD-9-CM 433-436- κωδικοί ICD-10-CM I63, I65, I66 ή I679) και ακολούθως από νοσηλεία με κύρια διάγνωση οξέος ισχαιμικού εγκεφαλικού επεισοδίου (κωδικοί ICD-9-CM 433-434, ICD-10-CM κωδικοί I63) σε CGMHs από το 2006 έως το 2016. Στην παρούσα μελέτη χρησιμοποιήθηκαν τρία κριτήρια για τον ορισμό των κακών εκβάσεων, συμπεριλαμβανομένου του ενδονοσοκομειακού θανάτου, η παραμονή στη μονάδα εντατικής θεραπείας (ΜΕΘ) και η παρατεταμένη νοσηλεία. Ο ενδονοσοκομειακός θάνατος είναι ένα από

τα χειρότερα αποτελέσματα της κλινικής περίθαλψης και έχει αξιολογηθεί σε πολλές μελέτες. Η παραμονή στη ΜΕΘ και η παρατεταμένη νοσηλεία υποδηλώνουν ότι οι ασθενείς χρειάζονται κρίσιμη φροντίδα ή περισσότερο χρόνο για να ανακάμψουν μετά τη λήψη θεραπείας rt-PA. Ως ενδονοσοκομειακός θάνατος ορίστηκε ο θάνατος στο νοσοκομείο ή σε ίδρυμα φροντίδας που καταγράφηκε στο εξιτήριο. Σημειώνεται ότι 24 ασθενείς χωρίς καταγραφή στη δήλωση εξιτηρίου θεωρήθηκαν ως περιπτώσεις χωρίς ενδονοσοκομειακούς θανάτους. Στο CGMH, οι περισσότεροι από τους ασθενείς που έλαβαν θεραπεία με rt-PA εισήχθησαν στη ΜΕΘ για παρακολούθηση για 2 ημέρες. Ως εκ τούτου, όλοι οι ασθενείς που έλαβαν θεραπεία με rt-PA θα έχουν τουλάχιστον 2 ημέρες παραμονής στη ΜΕΘ. Για να εντοπίσουμε τους ασθενείς που χρειάζονται εισαγωγή στη ΜΕΘ λόγω κακής έκβασης, τέθηκε ως όριο οι 2 ημέρες για να αποκλειστούν οι ασθενείς που εισήχθησαν στη ΜΕΘ μόνο για παρακολούθηση. Το τελευταίο κριτήριο ήταν η παρατεταμένη νοσηλεία. Η περίοδος μεταξύ της εισαγωγής στο νοσοκομείο και του εξιτηρίου από το νοσοκομείο σε ένα επεισόδιο περίθαλψης υπολογίστηκε ως LOS. Ορίστηκε η παρατεταμένη νοσηλεία ως LOS >55 ημέρες.Το όριο των 55 ημερών ορίστηκε από το 90ο εκατοστημόριο του LOS στα δεδομένα. Εάν η χρονική περίοδος μεταξύ δύο νοσηλειών, επισκέψεων στο τμήμα επειγόντων περιστατικών και επισκέψεων στα εξωτερικά ιατρεία που σχετίζονταν με το εγκεφαλικό επεισόδιο ήταν μικρότερη από 3 ημέρες, οι δύο αυτές επισκέψεις συγχωνεύθηκαν σε ένα ενιαίο επεισόδιο περίθαλψης, το οποίο αποτελεί μέσο για την εφαρμογή συνεπών κανόνων για τις ιατρικές καταστάσεις για την εξαγωγή συμπερασμάτων για διακριτά επεισόδια περίθαλψης, σύμφωνα με προηγούμενη μελέτη. Οι ασθενείς που πληρούσαν ένα από τα κριτήρια,συμπεριλήφθηκαν στην ομάδα κακής έκβασης, αλλά οι ασθενείς που δεν πληρούσαν κανένα από τα κριτήρια συμπεριλήφθηκαν στην ομάδα ευνοϊκής έκβασης. Οι εργαστηριακές εξετάσεις που ήταν διαθέσιμες σε ποσοστό μεγαλύτερο του 80% των ασθενών με ισχαιμικό εγκεφαλικό επεισόδιο συμπεριλήφθηκαν ως μεταβλητή για περαιτέρω ανάλυση: γενική αίματος (CBC), κρεατινίνη, νάτριο, αλανίνη αμινοτρανσφεράσης (ALT), τα επίπεδα γλυκόζης στο αίμα και τα επίπεδα καλίου. Η εξέταση γλυκόζης μπορεί να πραγματοποιηθεί με τη χρήση μιας παραδοσιακής εξέταση γλυκόζης αίματος ή μιας δακτυλικής εξέτασης γλυκόζης αίματος. Για να συμπεριληφθούν όλα τα αποτελέσματα της εξέτασης γλυκόζης, συλλέχθηκαν τα αποτελέσματα τόσο της παραδοσιακής όσο και της δακτυλοσκοπικής εξέτασης γλυκόζης για να πραγματοποιηθεί περαιτέρω ανάλυση. Οι ασθενείς μπορούσαν να υποβληθούν σε διάφορες εργαστηριακές εξετάσεις κατά τη διάρκεια της νοσηλείας. Συμπεριλήφθηκαν μόνο οι εργαστηριακές εξετάσεις που είχαν συλλεχθεί εντός του εγκεφαλικού επεισοδίου και εκείνες που ήταν πιο κοντά στον χρόνο πριν από τη θεραπεία με rt-PA.

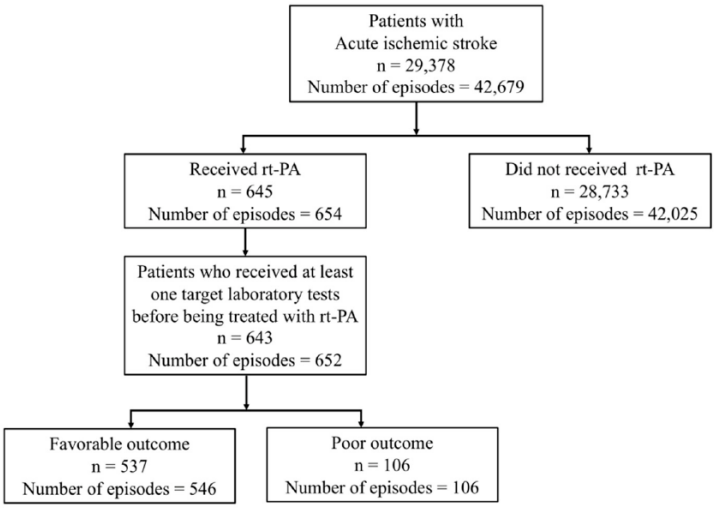
Για τη διερεύνηση του ιατρικού ιστορικού, ο δείκτης συννοσηρότητας Elixhauser που αναπτύχθηκε από το Healthcare Cost and Utilization Project (HCUP) εφαρμόστηκε στα αρχεία διαγνώσεων. Χρησιμοποιήθηκε το emr πακέτο της R , ένα εργαλείο για την ενσωμάτωση και την επεξεργασία των EMR, για την αξιολόγηση των μεταβλητών του ιατρικού ιστορικού για περαιτέρω ανάλυση. Αυτό το πακέτο μπορεί να χρησιμοποιηθεί για την ομαδοποίηση πολλαπλών κωδικών ICD σε μικρότερο αριθμό κλινικά σημαντικών κατηγοριών με την ταξινόμηση συννοσηρότητας Elixhauser. Για τον αποκλεισμό των ιατρικών ιστορικών που είχαν καταγραφεί σε λίγους μόνο ασθενείς στον πληθυσμό της παρούσας μελέτης, αποκλείστηκαν τα ιατρικά ιστορικά που καταγράφηκαν σε λιγότερο από το 5% των ασθενών

στις ομάδες κακής και ευνοϊκής έκβασης. Δηλαδή, συμπεριλήφθηκαν τα ιατρικά ιστορικά με ποσοστό εμφάνισης άνω του 5% για τους ασθενείς στις ομάδες κακής ή ευνοϊκής έκβασης. Χρησιμοποιήθηκε ο δείκτης σοβαρότητας εγκεφαλικού επεισοδίου (SSI) ως μέτρο της σοβαρότητας του εγκεφαλικού επεισοδίου. Η μελέτη των Sung και άλλων ερευνητών επιβεβαίωσε ότι ο SSI βάσει απαιτήσεων είναι έγκυρο υποκατάστατο της κλίμακας εγκεφαλικού επεισοδίου του εθνικού ινστιτούτου υγείας (NIHSS) για την εκτίμηση της σοβαρότητας του εγκεφαλικού επεισοδίου σε ασθενείς που νοσηλεύονται για οξύ ισχαιμικό εγκεφαλικό επεισόδιο. Υπολογίστηκαν οι κωδικοί χρέωσης της Εθνικής Ασφάλισης Υγείας της Ταϊβάν για το SSI και κατηγοριοποιήθηκαν σε τρεις βαρύτητες: ήπιο (SSI 5), μέτριο (5 < SSI 12) και σοβαρό (SSI > 12) εγκεφαλικό επεισόδιο σύμφωνα με τις προηγούμενες μελέτες. Επιπλέον, εφαρμόστηκαν δύο αναλύσεις ευαισθησίας για την αξιολόγηση της επίδρασης του προτεινόμενου ορισμού έκβασης και της μεθόδου που χρησιμοποιήθηκε για την ομαδοποίηση των αποτελεσμάτων των εργαστηριακών εξετάσεων. Η πρώτη ανάλυση ευαισθησίας όρισε μία κακή έκβαση ως ενδονοσοκομειακό θάνατο ή παραμονή στη ΜΕΘ. Επιπλέον, αντί να χρησιμοποιούνται οι πραγματικές τιμές των αποτελεσμάτων των εξετάσεων, στη δεύτερη ανάλυση ευαισθησίας, κατηγοριοποιήθηκαν τα αποτελέσματα των εργαστηριακών εξετάσεων ως φυσιολογικά, υψηλά ή χαμηλά σύμφωνα με τα ισχύοντα εύρη αναφοράς. Αυτή η μέθοδος μπορεί να αξιολογήσει την επίδραση της ομαδοποίησης των αποτελεσμάτων των εργαστηριακών εξετάσεων μέσω του εύρους αναφοράς. Οι μονομεταβλητές και οι πολυμεταβλητές αναλύσεις πραγματοποιήθηκαν και στις δύο αναλύσεις ευαισθησίας. Πραγματοποιήθηκε περιγραφική ανάλυση των χαρακτηριστικών των ασθενών στην ευνοϊκές και στις ομάδες κακής έκβασης. Οι συνεχείς μεταβλητές συνοψίστηκαν ως μέσοι όροι (τυπικές αποκλίσεις) ή διάμεσοι (ενδοτεταρτημοριακά εύρη) και οι διακριτές μεταβλητές συνοψίστηκαν ως συχνότητες και ποσοστά. Οι μέσοι όροι και οι διάμεσοι ελέγχθηκαν με το t-test του Student ή το τεστ Kruskal-Wallis, αντίστοιχα. Για την μονομεταβλητή ανάλυση των κατηγορικών μεταβλητών χρησιμοποιήθηκαν τα Chi-square tests ή τα Fisher's exact tests. Πραγματοποιήσαμε Lasso παλινδρόμηση στις πολυμεταβλητές αναλύσεις, η οποία μπορεί να αντιμετωπίσει την πολυσυγγραμμικότητα και είναι επίσης μια αυτοματοποιημένη μέθοδος επιλογής μεταβλητών. Τα δημογραφικά χαρακτηριστικά, τα αποτελέσματα των εργαστηριακών εξετάσεων και το ιατρικό ιστορικό, τα οποία ήταν διαφορετικά μεταξύ των ομάδων κακής και ευνοϊκής έκβασης (τιμή $p < 0,1$), διατηρήθηκαν ως εξαρτημένες μεταβλητές στο μοντέλο Lasso. Οι αναλύσεις πραγματοποιήθηκαν στην R (έκδοση 3.4.4, The R Foundation for Statistical Computing, http://www.r-project.org/, R Core Team, Βιέννη, Αυστρία). Όλοι οι στατιστικοί έλεγχοι ήταν αμφίπλευροι και η στατιστική σημαντικότητα ορίστηκε ως $p < 0,05$. Συνολικά, 42.679 περιστατικά περίθαλψης από 29.378 ασθενείς ήταν επιλέξιμα για τη μελέτη. Περίπου 2% των ασθενών (645 άτομα) έλαβαν rt-PA μετά την έναρξη του οξέος ισχαιμικού εγκεφαλικού επεισοδίου. Μεταξύ αυτών, τουλάχιστον μία εργαστηριακή εξέταση δόθηκε σε 652 επεισόδια περίθαλψης (643 ασθενείς) πριν από την rt-PA (Σχήμα 3.3 – Tseng *et.al,* 2020). Όπως φαίνεται στο Σχήμα 3.3, μεταξύ των ασθενών που διαγνώστηκαν με οξύ ισχαιμικό εγκεφαλικό επεισόδιο και τους χορηγήθηκε rt-PA, 537 είχαν ευνοϊκή έκβαση και 106 είχαν κακή έκβαση. Ο Πίνακας 3.8 περιγράφει τα δημογραφικά χαρακτηριστικά και τη σοβαρότητα του εγκεφαλικού επεισοδίου αυτών των ασθενών. Σε σύγκριση με τους ασθενείς με ευνοϊκή

έκβαση, οι ασθενείς με κακή έκβαση είχαν περισσότερες πιθανότητες να είναι μεγαλύτερης ηλικίας και να έχουν υψηλότερο SSI. Το τυποποιημένα σφάλματα της σοβαρότητας του εγκεφαλικού επεισοδίου παρουσιάζονται στον Πίνακα 3.8 (Tseng *et.al*, 2020). Η κατανομή του φύλου ήταν επίσης διαφορετική μεταξύ των δύο ομάδων. Μεταξύ των αποτελεσμάτων των εργαστηριακών εξετάσεων, το επίπεδο αιμοσφαιρίνης, η μέση συγκέντρωση σωματομετρικής αιμοσφαιρίνης (MCHC), ο αριθμός των αιμοπεταλίων και τα επίπεδα γλυκόζης ήταν διαφορετικά μεταξύ των ασθενών ευνοϊκής και κακής έκβασης (Πίνακας 3.9 - Tseng *et.al*, 2020). Υψηλότερες τιμές γλυκόζης, χαμηλότερα επίπεδα αιμοσφαιρίνης, χαμηλότερες MCHC, και χαμηλότερος αριθμός αιμοπεταλίων συσχετίστηκαν σημαντικά με τον αυξημένο κίνδυνο κακής έκβασης.

Σχήμα 3.3

Διάγραμμα ροής της μελέτης



Πίνακας 3.8

Μονομεταβλητή ανάλυση (Univariate analysis) των χαρακτηριστικών των ασθενών.

Patient Characteristics	Favorable Outcome (n = 546)	Poor Outcome (n = 106)	p Value
Sex (%)			0.013 *
Male	356 (65.2)	55 (51.9)	
Female	190 (34.8)	51 (48.1)	
Age (median [IQR [a], Q1–Q3 [b]])	66.00 [18, 56.00–74.00]	71.00 [18.75, 59.25–78.00]	0.001 *
SSI [c] (median [IQR, Q1–Q3])	9.65 [3.51, 8.28–11.79]	15.21 [11.34, 10.20–21.54]	<0.001 ***
Stroke severity (%)			<0.001 ***
Mild	99 (18.1)	14 (13.2)	
Moderate	313 (57.3)	20 (18.9)	
Severe	134 (24.5)	72 (67.9)	

* $p < 0.05$, *** $p < 0.001$, [a] IQR, interquartile range [first quartile, third quartile], [b] Q1–Q3, quartile 1–3, [c] Stroke severity index. SSI: stroke severity index.

Πίνακας 3.9

Μονομεταβλητή ανάλυση (Univariate analysis) των εργαστηριακών αποτελεσμάτων.

Table 2. Univariate analysis of the laboratory results.

Laboratory Tests	Favorable Outcome (n = 546)	Poor Outcome (n = 106)	p Value
Creatinine (median [IQR [a], Q1–Q3 [b]])	0.94 [0.36, 0.78–1.14]	0.95 [0.44, 0.79–1.23]	0.472
Hemoglobin (median [IQR, Q1–Q3])	14.20 [2.3, 13.00–15.30]	13.70 [2.6, 12.30–14.90]	0.015 *
Hematocrit (median [IQR, Q1–Q3])	41.70 [5.8, 38.70–44.50]	41.10 [6.2, 37.30–43.50]	0.053
MCH (mean corpuscular hemoglobin) (median [IQR, Q1–Q3])	30.50 [2.4, 29.30–31.70]	30.60 [2.4, 29.30–31.70]	0.826
MCHC (mean corpuscular hemoglobin concentration) (median [IQR, Q1–Q3])	33.90 [1.5, 33.20–34.70]	33.60 [1.8, 32.80–34.60]	0.024 *
MCV (mean corpuscular volume) (median [IQR, Q1–Q3])	89.60 [6, 86.50–92.50]	90.20 [6.2, 86.90–93.10]	0.259
Sodium (median [IQR, Q1–Q3])	139.00 [3.85, 137.15–141.00]	138.90 [3, 137.00–140.00]	0.116
Platelets (median [IQR, Q1–Q3])	206.00 [72, 170.00–242.00]	186.00 [64, 159.00–223.00]	0.007 **
RBCs (red blood cells) (median [IQR, Q1–Q3])	4.69 [0.67, 4.36–5.03]	4.69 [0.85, 4.15–5.00]	0.116
RDW (red cell distribution width) (median [IQR, Q1–Q3])	13.60 [3.6, 12.90–16.50]	13.60 [2.5, 12.80–15.30]	0.608
WBCs (white blood cells) (median [IQR, Q1–Q3])	7.90 [3.5, 6.40–9.90]	7.50 [3, 6.50–9.50]	0.639
ALT (alanine aminotransferase) (median [IQR, Q1–Q3])	23.00 [15, 17.00–32.00]	22.50 [17, 16.00–33.00]	0.694
Glucose (median [IQR, Q1–Q3])	128.00 [47, 110.00–157.00]	144.00 [56, 124.00–180.00]	<0.001 ***
Potassium (median [IQR, Q1–Q3])	3.70 [0.5, 3.44–3.94]	3.70 [0.42, 3.57–3.99]	0.139

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, [a] IQR, interquartile range [first quartile, third quartile], [b] Q1–Q3, quartile 1–3.

Το ιατρικό ιστορικό των ασθενών, το οποίο καθορίστηκε με βάση την ταξινόμηση συννοσηρότητας Elixhauser, συνοψίζονται στον Πίνακα 3.10 (Tseng *et.al*, 2020). Οι δύο ομάδες δεν διέφεραν ως προς το ιατρικό ιστορικό τους (όλες τα p-values>0,05).

Πίνακας 3.10

Μονομεταβλητή ανάλυση (Univariate analysis) των μεταβλητών του ιατρικού ιστορικού.

Medical Histories (%)	Favorable Outcome (n = 546)	Poor Outcome (n = 106)	p-Value
Deficiency anemias	11 (2.0)	6 (5.7)	0.068
Congestive heart failure	42 (7.7)	7 (6.6)	0.851
Diabetes without chronic complications	85 (15.6)	16 (15.1)	1
Hypertension, uncomplicated	187 (34.2)	33 (31.1)	0.611
Hypertension, complicated	21 (3.8)	6 (5.7)	0.554
Liver disease	28 (5.1)	8 (7.5)	0.444
Chronic pulmonary disease	45 (8.2)	12 (11.3)	0.401
Solid tumor without metastasis	29 (5.3)	8 (7.5)	0.496
Valvular disease	34 (6.2)	8 (7.5)	0.771

Συμπεριλήφθηκαν τα αποτελέσματα των εργαστηριακών εξετάσεων και του ιατρικού ιστορικού που ήταν διαφορετικά (p-value < 0,1) μεταξύ ασθενών με ευνοϊκή και κακή έκβαση και τα δημογραφικά χαρακτηριστικά τους ως ανεξάρτητες μεταβλητές στο μοντέλο Lasso. Στον Πίνακα 3.11 (Tseng *et.al*, 2020) παρουσιάζονται οι συντελεστές του μοντέλου (lambda = 0,009326033). Μόνο οι σημαντικές μεταβλητές των οποίων η συνδιακύμανση δεν ήταν μηδενική παρουσιάζονται. Οι μεταβλητές που αναδείχθηκαν ως παράγοντες κινδύνου της κακής θεραπευτικής έκβασης για ασθενείς με οξύ ισχαιμικό εγκεφαλικό επεισόδιο περιελάμβαναν ασθενείς που ήταν γυναίκες και είχαν αναιμία, πιο σοβαρό SSI, υψηλότερη τιμή γλυκόζης, χαμηλότερη MCHC και χαμηλότερο αριθμό αιμοπεταλίων.

Πίνακας 3.11

Παράγοντες κινδύνου που εντοπίστηκαν από το μοντέλο Lasso και οι σχετικοί συντελεστές τους.

Selected Variable	Coefficient
Anemia	0.752
Sex: Male	−0.178
SSI [a]	0.887
MCHC	−0.042
Platelet Count	−0.142
Glucose	0.200

[a] Stroke severity index

Στην ανάλυση ευαισθησίας, επαναπροσδιορίστηκε η κακή έκβαση ω ενδονοσοκομειακός θάνατος ή παραμονή στη ΜΕΘ. Υπήρχαν 591 και 61 περιστατικά στις νέες ομάδες ευνοϊκής και κακής έκβασης, αντίστοιχα. Σε σύγκριση με την μονοπαραγοντική ανάλυση των χαρακτηριστικών των ασθενών, οι μεταβλητές που ήταν σημαντικά διαφοροποιημένες μεταξύ των δύο ομάδων ήταν οι ίδιες σε αυτή την ανάλυση ευαισθησίας (Πίνακας 3.2.2). Στη μονοπαραγοντική ανάλυση των αποτελεσμάτων των εργαστηριακών εξετάσεων, οι υψηλότερες τιμές γλυκόζης και καλίου συσχετίστηκαν σημαντικά με τον αυξημένο κίνδυνο κακής έκβασης της θεραπείας (Πίνακας 3.2.3). Σε αντίθεση με την πρωτογενή ανάλυση (Πίνακας 3.2.2), οι αριθμοί των αιμοπεταλίων δεν διέφεραν, αλλά οι αριθμοί των ερυθροκυττάρων και τα επίπεδα καλίου ήταν διαφορετικά μεταξύ των δύο ομάδων σε αυτή την ανάλυση ευαισθησίας. Τα αποτελέσματα του ιατρικού ιστορικού περιγράφονται στον Πίνακα 3.2.4, όπου μόνο η υπέρταση που επιπλέκεται από άλλες ασθένειες σχετιζόταν με κακή έκβαση. Ο παράγοντας κινδύνου που επιλέχθηκε από τη Lasso παλινδρόμηση σε αυτή την ανάλυση ήταν το SSI, του οποίου η συνδιακύμανση ήταν 0,146 και το λάμδα ήταν 0,1519911.

Στη δεύτερη ανάλυση ευαισθησίας, χωρίστηκαν τα αποτελέσματα των εργαστηριακών εξετάσεων σε κατηγορίες με βάση τα εύρη αναφοράς τους και τα αποτελέσματα της μονομεταβλητής ανάλυσης παρουσιάζονται στον πίνακα 5. Αυτός ο πίνακας δείχνει ότι οι κακές εκβάσεις σχετίζονται με μη φυσιολογικά επίπεδα αιμοσφαιρίνης, αιματοκρίτη, RBCs και γλυκόζης. Οι τιμές αιμοσφαιρίνης και γλυκόζης άνω του φυσιολογικού ήταν πιο πιθανό να οδηγήσουν σε κακή πρόγνωση. Όσον αφορά τους παράγοντες κινδύνου, μετά το μετασχηματισμό των αποτελεσμάτων των εργαστηριακών εξετάσεων, του φύλου, του SSI, η αιμοσφαιρίνη, ο αιματοκρίτης, το νάτριο, η γλυκόζη και το ιστορικό αναιμίας θεωρήθηκαν σχετικοί παράγοντες που επιλέχθηκαν με τη μέθοδο Lasso, όπως φαίνεται στον πίνακα 6. Παρόμοια με τα αποτελέσματα της πρωτογενούς ανάλυσης, στις γυναίκες,η ύπαρξη υψηλότερων επιπέδων SSI και γλυκόζης και η αναιμία, βρέθηκαν επίσης να είναι σημαντικά χαρακτηριστικά σε αυτήν την ανάλυση ευαισθησίας. Λόγω των ανεπιθύμητων παρενεργειών, όπως η αιμορραγία, και των χαμηλών ποσοστών χρήσης του rt-PA, είναι σημαντικό να προσδιοριστούν οι παράγοντες κινδύνου σε σχέση με τα αποτελέσματα της θεραπείας. Συμπεριλήφθηκαν δημογραφικές πληροφορίες, ιατρικά ιστορικά και αποτελέσματα εργαστηριακών εξετάσεων για να προσδιοριστούν οι μεταβλητές που σχετίζονται με την

έκβαση της θεραπείας με rt-PA. Οι γυναίκες ασθενείς, οι ασθενείς με αναιμία, οι ασθενείς με υψηλότερα επίπεδα SSI και γλυκόζης και τα χαμηλότερα MCHCs και ο αριθμός αιμοπεταλίων είχαν περισσότερες πιθανότητες να έχουν κακή έκβαση μετά τη λήψη rt-PA. Η σοβαρότητα του εγκεφαλικού επεισοδίου επιλέχθηκε ως σημαντικός παράγοντας κινδύνου στην πρωτογενή ανάλυση και στις δύο αναλύσεις ευαισθησίας. Η αναιμία, το γυναικείο φύλο και τα υψηλότερα επίπεδα γλυκόζης ορού επιλέχθηκαν στην πρωτογενή ανάλυση και την ανάλυση ευαισθησίας, η οποία ομαδοποίησε τα εργαστηριακά αποτελέσματα με βάση τα εύρη αναφοράς τους. Στην ανάλυση ευαισθησίας που χρησιμοποίησε έναν εναλλακτικό ορισμό έκβασης, μόνο η SSI διατηρήθηκε στο Lasso μοντέλο. Ο πιθανός λόγος είναι ότι ο αριθμός των περιστατικών στις εναλλακτικές ομάδες κακής έκβασης ήταν μόνο 61. Η εναλλακτική ομάδα κακής έκβασης ήταν πολύ μικρότερη από την ομάδα ευνοϊκής έκβασης. Προηγούμενες μελέτες έχουν δείξει ότι η σοβαρότητα του εγκεφαλικού επεισοδίου και τα επίπεδα γλυκόζης είχαν σημαντική επίδραση στην ενδοεγκεφαλική αιμορραγία μετά από ενδοφλέβια θεραπεία με ενεργοποιητή ιστικού πλασμινογόνου, η οποία συμφωνούσε με τα ευρήματά της παρούσας μελέτης, παρόλο που χρησιμοποιήθηκε το SSI ως υποκατάστατο της βαθμολογίας NIHSS ως μέτρο της σοβαρότητας του εγκεφαλικού επεισοδίου. Εκτός από το επίπεδο γλυκόζης, ο διαβήτης μπορεί να σχετίζεται με κακή έκβαση της θεραπείας με rt-PA. Μια μελέτη διαπίστωσε ότι στην υποομάδα των διαβητικών, ο δείκτης γλυκοζυλιωμένης αιμοσφαιρίνης συσχετίστηκε θετικά με τη συμπτωματική ενδοκρανιακή αιμορραγία. Η παρούσα μελέτη αντί να στοχέυσει σε μία συγκεκριμένη υποομάδα, εξετάστηκαν όλους οι ασθενείς με οξύ ισχαιμικό εγκεφαλικό επεισόδιο που έλαβαν θεραπεία με rt-PA ως περιπτώσεις μελέτης. Στον γενικό πληθυσμό, η εξέταση του δείκτη γλυκοζυλιωμένης αιμοσφαιρίνης πραγματοποιήθηκε σε μόνο σε λίγους ασθενείς, επομένως συμπεριλήφθηκε μόνο το επίπεδο γλυκόζης στις εργαστηριακές αξιολογήσεις. Μια πρόσφατη μελέτη χρησιμοποίησε τη λειτουργική περιπατητική κατάσταση ως έκβαση για να διερευνήσει τη συσχέτιση μεταξύ του κινδύνου παραγόντων και των εκβάσεων σε ασθενείς με ισχαιμικό εγκεφαλικό επεισόδιο που έλαβαν rt-PA και λάμβαναν αντιυπερτασικά φάρμακα. Στην παρούσα μελέτη, ορίστηκε ως κακή έκβαση, ο ενδονοσοκομειακός θάνατος, η παραμονή στη ΜΕΘ και η παρατεταμένη LOS, τα οποία είναι διαφορετικά από τη δημοσιευμένη μελέτη. Οι διαφορές μεταξύ των φύλων μεταξύ των ασθενών με οξύ ισχαιμικό εγκεφαλικό επεισόδιο που έλαβαν rt-PA έχουν συζητηθεί σε πολλές μελέτες. Μια συστηματική ανασκόπηση υπέδειξε ότι δεν υπήρχε διαφορά των δύο φύλων στην έκβαση μεταξύ των ασθενών που έλαβαν ενδοφλέβια rt-PA. Η κλινική έκβαση και ο αριθμός των ασθενών με ευνοϊκή έκβαση δεν διέφερε μεταξύ γυναικών και ανδρών. Λίγες μελέτες έχουν αναφέρει ότι δεν παρατηρήθηκε η συνήθης διαφοροποίηση των φύλων στην έκβαση υπέρ των ανδρών μεταξύ των ασθενών που έλαβαν θεραπεία με rt-PA. Ωστόσο, παρόμοια με τα ευρήματα της παρούσας μελέτης, λίγες μελέτες έδειξαν ότι οι γυναίκες έχουν υψηλότερη θνησιμότητα μετά το εγκεφαλικό επεισόδιο, ποσοστό αναπηρίας, κατάθλιψη και άνοια και χειρότερη βαθμολογία mRS (τροποποιημένη κλίμακα Rankin) κατά την έξοδο από το νοσοκομείο σε σύγκριση με τους άνδρες. Οι διαφορές μεταξύ των δύο φύλων σε συμπτώματα κατά την παρουσίαση μπορεί να δημιουργήσουν καθυστερήσεις στη θεραπεία για τις γυναίκες. Η διαφορά που διαπιστώθηκε σε αυτήν την μελέτη θα μπορούσε ενδεχομένως να εξηγηθεί από φυσιολογικές διαταραχές που δεν περιλαμβάνονται στο σύνολο δεδομένων. Παρόμοια με μια

προηγούμενη μελέτη, διαπιστώθηκε ότι οι γυναίκες μέσης ηλικίας έχουν καλύτερη έκβαση από ό,τι οι μεσήλικες άνδρες, ενώ σε πιο προχωρημένη ηλικία, οι άνδρες έχουν καλύτερη έκβαση από τις γυναίκες. Ωστόσο, οι διαφορές δεν είναι στατιστικά σημαντικές. Οι διαφορές μεταξύ των δύο φύλων στη θεραπεία με rt-PA εξακολουθούν να υπάρχουν μεταξύ των διάφορων μελετών και απαιτείται περαιτέρω έρευνα. Η παρούσα μελέτη χρησιμοποίησε την προσέγγιση Lasso για την επιλογή σημαντικών μεταβλητών στην πολυμεταβλητή ανάλυση επειδή υπήρχε υψηλή συσχέτιση μεταξύ του MCHC, των επιπέδων αιμοσφαιρίνης και των επιπέδων αιματοκρίτη. Ακόμη, ορισμένες μελέτες ανέφεραν ότι οι συντελεστές παλινδρόμησης σε ένα μοντέλο σταδιακής επιλογής μπορεί να έχουν σημαντική μεροληψία. Ομοίως, με την stepwise regression, η Lasso, το οποίο προσθέτει στην κανονικοποίηση penalty στον αριθμό των παραμέτρων στο μοντέλο, αποτρέπει το overfitting και την πολυσυγγραμμικότητα. Το κύριο χαρακτηριστικό της μελέτης είναι ότι εξετάστηκαν τα δημογραφικά χαρακτηριστικά, τα εργαστηριακά αποτελέσματα εξετάσεων και το ιατρικό ιστορικό για τη διερεύνηση των παραγόντων κινδύνου για κακή έκβαση μετά τη λήψη rt-PA. Μόνο λίγες μελέτες περιλαμβάνουν αποτελέσματα εργαστηριακών εξετάσεων για την ανάλυση των αποτελεσμάτων της θεραπείας με rt-PA και καμία μελέτη δεν έχει διερευνήσει τις συσχετίσεις μεταξύ των εργαστηριακών αποτελεσμάτων και της κακής έκβασης της θεραπείας στον γενικό πληθυσμό των ασθενών με οξύ ισχαιμικό εγκεφαλικό επεισόδιο, όπου διαπιστώθηκε ότι το επίπεδο γλυκόζης, τα MCHCs και τα αιμοπετάλια σχετίζονταν με την έκβαση της θεραπείας. Επιπλέον, επειδή τα αποτελέσματα των εργαστηριακών εξετάσεων μπορεί να περιγραφούν ως φυσιολογικά και μη φυσιολογικά, καθώς και ο ορισμός της έκβασης της θεραπείας μπορεί να επηρεάσει τα αποτελέσματα της ανάλυσης, αξιολογήθηκε επίσης η επίδραση της αλλαγής του ορισμού των μεταβλητών. Για την αντιμετώπιση των προβλημάτων που αναφέρθηκαν παραπάνω, πραγματοποιήθηκαν αναλύσεις ευαισθησίας για να αξιολογηθεί η επίδραση της εφαρμογής διαφορετικών αποτελεσμάτων και της ομαδοποίησης των αποτελεσμάτων των εργαστηριακών εξετάσεων με βάση τα εύρη αναφοράς τους. Τα αποτελέσματά της μελέτης μπορούν παρέχουν προγνωστικές πληροφορίες σχετικά με τη χρήση rt-PA για ισχαιμικό εγκεφαλικό επεισόδιο. Όμως η παρούσα μελέτη έχει αρκετούς περιορισμούς. Πρώτον, ο αριθμός των ασθενών που έλαβαν θεραπεία με rt-PA ήταν μόνο 643, λόγω της χαμηλής χρήσης του rt-PA, αν και συμπεριλήφθηκαν δεδομένα 11 ετών και 42.679 οξέα ισχαιμικά εγκεφαλικά επεισόδια. Δεύτερον, οι συννοσηρότητες Charlson έχουν συχνά χρησιμοποιηθεί για τη μέτρηση των συννοσηροτήτων. Ωστόσο, ο δείκτης συννοσηρότητας Charlson περιλαμβάνει μόνο 17 κατηγορίες συννοσηρότητας και ορισμένες σημαντικές ασθένειες που σχετίζονται με την αξιολόγηση της έκβασης του rt-PA. Για παράδειγμα, η αναιμία, δεν περιλαμβάνεται στον Charlson. Για τον εκτενή καθορισμό και την ανάλυση του ιατρικού ιστορικού των ασθενών με εγκεφαλικό επεισόδιο, επιλέχθηκε ο δείκτης συννοσηρότητας Elixhauser, ο οποίος περιλαμβάνει 30 κατηγορίες συννοσηρότητας, για την ομαδοποίηση των διαγνώσεων. Τρίτον, τα ιατρικά ιστορικά ενδέχεται να λείπουν εάν οι ασθενείς δεν έχουν σχετική διάγνωση στα EMR τους. Ωστόσο, στην ομάδα μελέτης, πάνω από το 70% των ασθενών είχαν επισκεφθεί το CGMH πριν να έχουν υποστεί εγκεφαλικό επεισόδιο. Επιπλέον, τα ενδονοσοκομειακά EMR θα πρέπει να χρησιμοποιούνται μόνο για τη δημιουργία ενός μοντέλου κινδύνου σε σενάρια χειρότερης περίπτωσης, όπου οι αναίσθητοι ασθενείς εισάγονται στο τμήμα επειγόντων

περιστατικών και δεν μπορούν να παρασχεθούν πρόσθετες πληροφορίες. Τέταρτον, οι ασθενείς που συμπεριλήφθηκαν στη μελέτη ενδέχεται να μην είναι επιλέξιμοι για θεραπεία με rt-PA βάσει των κατευθυντήριων οδηγιών. Στη μελέτη, συμπεριλήφθηκαν όλοι οι ασθενείς στους οποίους χορηγήθηκε rt-PA για ισχαιμικό αγγειακό εγκεφαλικό επεισόδιο, ώστε να εξασφαλιστεί επαρκής αριθμός περιπτώσεων και να αντικατοπτριστούν οι συνθήκες στον πραγματικό κόσμο. Ένας άλλος περιορισμός είναι ότι τα δεδομένα για τον χρόνο μέχρι τη θεραπεία, ο οποίος είναι γνωστός ως παράγοντας που σχετίζεται με τα αποτελέσματα, δεν ήταν διαθέσιμα στο σύνολο των δεδομένων. Μια προηγούμενη μελέτη κατέληξε στο συμπέρασμα ότι η θρομβολυτική θεραπεία πέραν του χρονικού παραθύρου των 4,5 ωρών φαίνεται να σχετίζεται με σημαντική αύξηση της θνησιμότητας σε κλινικές πρακτικές. Ωστόσο, σύμφωνα με προηγούμενες μελέτες, το τρίτο τεταρτημόριο του χρόνου έναρξης της θεραπείας μέχρι τη θεραπεία ήταν μικρότερο από 3 ώρες στην Ταϊβάν, δηλαδή μόνο ένα μικρό ποσοστό των περιπτώσεων υποβλήθηκε σε θεραπεία πέραν τις 4,5 ώρες. Τα άλλα κλινικά χαρακτηριστικά που είναι σημαντικά για την πρόβλεψη της κλινικής έκβασης, όπως ο υποτύπος του εγκεφαλικού επεισοδίου, η εγκεφαλική αρτηριακή ανακάθαρση, καθώς και η παρουσία και η θέση της απόφραξης δεν μπόρεσαν να εξαχθούν από το σύνολο των δεδομένων. Αυτά τα δεδομένα θα πρέπει να συλλεχθούν για περαιτέρω ανάλυση, μεταξύ άλλων για τη δημιουργία προγνωστικών μοντέλων. Για την τιμή NIHSS κατά την εισαγωγή, χρησιμοποιήθηκε το SSI, ένα έγκυρο υποκατάστατο της βαθμολογίας NIHSS, ως μέτρο της σοβαρότητας του εγκεφαλικού επεισοδίου.

Συνοψίζοντας, αναλύθηκαν οι παράγοντες κινδύνου που μπορεί να σχετίζονται με δυσμενή αποτελέσματα της θεραπείας με rt-PA. Τα ευρήματά της παρούσας μελέτης έδειξαν ότι το γυναικείο φύλο, τα υψηλότερα επίπεδα γλυκόζης, το χαμηλότερο MCHC, ο χαμηλότερος αριθμός αιμοπεταλίων, το ιστορικό αναιμίας και το σοβαρό εγκεφαλικό επεισόδιο ήταν οι παράγοντες κινδύνου που σχετίζονται με τη θεραπεία με rt-PA και τα ευρήματα αυτά θα μπορούσαν να βοηθήσουν να μελετηθεί η στρατηγική θεραπείας για το οξέο ισχαιμικό εγκεφαλικό επεισόδιο (Tseng *et.al,* 2020).

3.3 Πρόβλεψη κινδύνου για την πρώιμη χρόνια νεφρική νόσο

Στην έρευνα των Shih *et.al* (2020) αναφέρεται πως η χρόνια νεφρική νόσος (Chronic Kidney Disease - CKD) αποτελεί παγκόσμιο πρόβλημα δημόσιας υγείας και σχετίζεται με σοβαρή νοσηρότητα, θνησιμότητα και χρήση πόρων υγείας. Το 2017, ο αριθμός των περιπτώσεων παγκοσμίως ήταν 69,75 εκατομμύρια, και η CKD προκάλεσε 1,2 εκατομμύρια θανάτους. Ο παγκόσμιος επιπολασμός της CKD ήταν 9,1% το 2017. Σύμφωνα με την ετήσια έκθεση του Υπουργείου Υγείας και Πρόνοιας της Ταϊβάν, η CKD αντιπροσωπεύει το μεγαλύτερο αριθμό απαιτήσεων από την ασφάλιση υγείας, με 364.000 εισαγόμενους ασθενείς, με κόστος περίπου 51,3 δισεκατομμύρια δολάρια το 2018. Με τη γήρανση του πληθυσμού και τη συναφή αυξανόμενη επικράτηση της υπέρτασης, της υπερλιπιδαιμίας και της υπεργλυκαιμίας, ο αριθμός των ασθενών με CKD αυξάνεται συνεχώς. Η πρώιμη CKD δεν έχει εμφανή συμπτώματα. Η νεφρική λειτουργία ενός ασθενούς με CKD μειώνεται σταδιακά και

αναπτύσσεται ουραιμία, όπου σε αυτό το στάδιο ο ασθενής πρέπει να υποβληθεί σε αιμοκάθαρση ή μεταμόσχευση νεφρού. Δύο πρότυπα καθορίζουν την CKD: (1) το νεφρό έχει υποστεί βλάβη για πάνω από τρεις μήνες, συμπεριλαμβανομένων των δομικών και λειτουργικών ανωμαλιών, κάποιος άλλος τρόπος αντιμετώπισης αυτών που φαίνεται να είναι παθολογικές ανωμαλίες, το αίμα, τα ούρα ή απεικονιστικές ανωμαλίες, και (2) ο ρυθμός σπειραματικής διήθησης (glomerular filtration rate - GFR) < 60 ml/min/1,73 m2 για πάνω από τρεις μήνες. Σε γενικές γραμμές, η CKD χωρίζεται σε πέντε στάδια με βάση τον εκτιμώμενο GFR (eGFR) (Πίνακας 3.12 - Shih *et.al*, 2020).

Πίνακας 3.12
Τα στάδια της χρόνιας νεφρικής νόσου (CKD)

Stage	Description	Estimated GFR
1	Kidney damage with normal or increased GFR	≥90 mL/min/1.73 m^2
2	Kidney damage with small decrease in GFR	60–89.9 mL/min/1.73 m^2
	Kidney damage with moderate decrease in GFR	30–59.9 mL/min/1.73 m^2
3	3a	45–59.9 mL/min/1.73 m^2
	3b	30–44.9 mL/min/1.73 m^2
4	Kidney damage with large decrease in GFR	15–29.9 mL/min/1.73 m^2
5	Kidney failure with need for dialysis (end-stage renal disease)	<15 mL/min/1.73 m^2

GFR: Glomerular Filtration Rate; 3a: Stage 3a of kidney disease; 3b: Stage 3b of kidney disease.

Οι τρέχουσες διαδικασίες για την ανίχνευση της πρώιμης CKD ανεπαρκείς. Στην Ταϊβάν, υπάρχουν τουλάχιστον 2 εκατομμύρια ασθενείς με CKD, αλλά μόνο το 3,5% αυτών έχουν διαγνωστεί και είναι ενημερωμένοι. Η ανίχνευση της χρόνιας νεφρικής ανεπάρκειας είναι δύσκολη έως ότου έχει ήδη χαθεί το 25% της νεφρικής λειτουργίας. Η έγκαιρη διάγνωση μπορεί ενδεχομένως να αποτρέψει ή να μετριάσει την εξέλιξη της CKD σε νεφρική νόσο τελικού σταδίου. Η παρούσα μελέτη σχεδιάστηκε για τον εντοπισμό των παραγόντων κινδύνου CKD μέσω της προληπτικής υγείας των ενηλίκων της Ταϊβάν για την έγκαιρη πρόβλεψη της μειωμένης νεφρικής λειτουργίας. Από το 2012, η Ταϊβάν έχει εφαρμόσει το "πενταετές σχέδιο για την πρόληψη της χρόνιας νεφρικής νόσου και τη βελτίωση της ποιότητας της περίθαλψης, 2012-2016". Τα αποτελέσματα του προγράμματος περιλάμβαναν μειωμένες περιπτώσεις αιμοκάθαρσης και αυξημένο ποσοστό πενταετούς επιβίωσης των ασθενών μετά από μεταμόσχευση νεφρού. Ωστόσο, το 2017, η Ταϊβάν ανέφερε 275.000 περιπτώσεις με CKD και 6743 θανάτους από CKD. Λαμβάνοντας υπόψη την ετερογένεια της επιδείνωσης της CKD, είναι κρίσιμη η διενέργεια εκτίμησης κινδύνου, παρακολούθησης και πρόγνωσης από μια τεκμηριωμένη ιατρική άποψη. Μια πρόσφατη έρευνα έδειξε υψηλό επιπολασμό της CKD στον πληθυσμό της Ταϊβάν, με ανησυχητικά χαμηλό ποσοστό ευαισθητοποίησης. Επιπλέον, οι παράγοντες κινδύνου CKD, όπως η υψηλή αρτηριακή πίεση, η χαμηλή κοινωνικοοικονομική κατάσταση και η φυτική φαρμακευτική αγωγή, είναι κοινοί στην Ταϊβάν. Οι παράγοντες πρόβλεψης για τη CKD έχουν εξεταστεί εκτενώς τα τελευταία χρόνια, αλλά παραμένουν αμφιλεγόμενοι. Με βάση μια έκθεση από την US Preventive Services Task Force (USPSTF) και του Αμερικανικού Κολλεγίου Ιατρών (ACP), ο έλεγχος της CKD σε ασυμπτωματικά άτομα είναι ανεπαρκής και δεν υπάρχουν έγκυρα εργαλεία για τον έλεγχο της. Το American Society of Nephrology συνιστά ανεπιφύλακτα τον τακτικό έλεγχο για τη CKD, ανεξάρτητα από τους παράγοντες κινδύνου. Είναι γνωστό ότι η αμφίδρομη λειτουργία παίζει κρίσιμο ρόλο στη

δυσλιπιδαιμία και την πρωτεϊνουρία και επηρεάζει επίσης τον μεταβολισμό των λιποπρωτεϊνών. Οι μέσες τιμές της HDL-L (λιποπρωτεΐνη υψηλής πυκνότητας χοληστερόλης) και της LDL-C (χοληστερόλη λιποπρωτεϊνών χαμηλής πυκνότητας) είναι χαμηλότερες στο στάδιο 3 έως 5 της CKD από ό,τι σε υγιή άτομα. Η χρόνια νεφρική ανεπάρκεια σχετίζεται με πολλούς παράγοντες, συμπεριλαμβανομένης της υπέρτασης και της πρωτεϊνουρίας. Για παράδειγμα, είναι γνωστό ότι το μέγεθος της μείωσης της αρτηριακής πίεσης (υπέρτασης) εμφανίστηκε μεγαλύτερο με την εξέλιξη της CKD. Σε αντίθεση με την πρώιμη CKD, έχει αναφερθεί από πολλές μελέτες ότι η υπέρταση αποτελεί συννοσηρότητα της CKD, αλλά έχει μελετηθεί ελάχιστα στην πρώιμη CKD. Λόγω της ετερογένειας της CKD, οι απαντήσεις στον έλεγχο και την κλινική πρακτική δεν είναι σαφείς. Ωστόσο, απαιτείται επειγόντως ένα ακριβές εργαλείο για την πρόβλεψη της CKD. Η πρώιμη ευαισθητοποίηση για την CKD είναι απαραίτητη για τους δυνητικούς ασθενείς ώστε να συμμετέχουν και να συμμορφώνονται με τα προγράμματα προληπτικής εξέτασης των ενηλίκων. Πράγματι, η εξόρυξη δεδομένων έχει χρησιμοποιηθεί με επιτυχία για τη δημιουργία ενός προγνωστικού μοντέλου στον τομέα της υγειονομικής περίθαλψης. Έτσι, στην παρούσα μελέτη, τέσσερις αλγόριθμοι εξόρυξης δεδομένων, συμπεριλαμβανομένου ενός δέντρο ταξινόμησης και παλινδρόμησης (classification and regression tree - CART), ένα δέντρο απόφασης C4.5, μια γραμμική διαχωριστική ανάλυση (Linear Discriminant Analysis - LDA) και ένα extreme learning machine (ELM) μοντέλο χρησιμοποιούνται για την πρόβλεψη της πρώιμης CKD. Ειδικότερα, η παρούσα μελέτη είχε ως στόχο να χρησιμοποιήσει τέσσερις μεθόδους εξόρυξης δεδομένων. Επιπλέον, αυτές οι μέθοδοι έχουν τη δυνατότητα να διερευνήσουν σημαντικούς παράγοντες κινδύνου της πρώιμης CKD και την ερμηνεία της συσχέτισης μεταξύ τους. Όλα τα δείγματα ελήφθησαν από ένα σύνολο δεδομένων εξετάσεων υγείας ενηλίκων, τα οποία συλλέγονται από 32 αλυσίδες κλινικών και τρία ειδικά κέντρα φυσικής εξέτασης. Τα δεδομένα από την 1η Ιανουαρίου 2015 έως τις 31 Δεκεμβρίου 2019 συμπεριελήφθησαν, δίνοντας συνολικά 19.270 πραγματικές εγγραφές, συμπεριλαμβανομένων 5101 ασθενών με CKD και 14.169 ασθενείς που δεν πάσχουν από CKD. Οι προσωπικές πληροφορίες, τα δεδομένα της φυσικής εξέτασης και τα αποτελέσματα των αιματολογικών εξετάσεων από την βάση δεδομένων φυσικής εξέτασης συμπεριλήφθηκαν και προσδιορίστηκαν συνολικά 11 ανεξάρτητες μεταβλητές. Η εξαρτημένη μεταβλητή ήταν το GFR (Πίνακας 3.13 - Shih *et.al*, 2020).

Πίνακας 3.13

Σημαντικές μεταβλητές και κωδικοποίηση στην παρούσα μελέτη

Variable	Name	Definition of Normal Test Data
X1	Gender	Male/Female
X2	Age	Age greater than 40 years
X3	Red blood cells (RBC)	0–5
X4	Glucose Fasting (GLU)	70–100
X5	Triglycerides (TG)	50–150
X6	Total Cholesterol (T-CHO)	50–200
X7	High-Density Lipoprotein Cholesterol (HDL-C)	>40
X8	Low-Density Lipoprotein Cholesterol (LDL-C)	<130
X9	Albumin (ALB)	3.5–5.0
X10	Proteinuria (PRO)	+/−
X11	Urine protein and creatinine ratio (UPCR)	<150
Y	Glomerular filtration rate (GFR)	≥90 mL/min/1.73 m^2

Αυτή η μελέτη είχε ως στόχο να χρησιμοποιήσει τέσσερις μεθόδους εξόρυξης δεδομένων που περιλαμβάνουν τις μεθόδους CART, C4.5, LDA και ELM για την πρόβλεψη της πρώιμης CKD. Το CART είναι ένα σύστημα δέντρων απόφασης το οποίο χρησιμοποιεί μια δυαδική αναδρομική διαδικασία για την κατάτμηση των δεδομένων σε ομοιογενή υποσύνολα με βάση το δείκτη Gini. Η κατάτμηση επαναλαμβάνεται έως ότου οι κόμβοι είναι αρκετά ομοιογενείς ώστε να είναι τερματικοί. Το πρώτο βήμα της ανάλυσης CART είναι η δημιουργία του μέγιστου δέντρου με δυαδική διαδικασία διαχωρισμού, το οποίο περιγράφει τα δεδομένα. Το δεύτερο βήμα είναι το pruning του δέντρου και η εξαγωγή μιας σειράς λιγότερο σύνθετων δέντρων από το μέγιστο δέντρο. Το τελευταίο βήμα είναι η επιλογή ενός βέλτιστου δέντρου χρησιμοποιώντας μια διαδικασία cross-validation. Ο C4.5 είναι επίσης ένας αλγόριθμος δέντρων απόφασης ο οποίος επιλέγει τα χαρακτηριστικά του δέντρου απόφασης σε κάθε κόμβο με βάση την έννοια της εντροπίας. Υιοθετεί μια προσέγγιση κατά την οποία τα δέντρα απόφασης κατασκευάζονται με έναν αναδρομικό τρόπο "διαίρει και βασίλευε" από πάνω προς τα κάτω. Σε κάθε κόμβο του δέντρου, το C4.5 επιλέγει ένα χαρακτηριστικό με βάση την μέγιστη πληροφορία που διαχωρίζει αποτελεσματικότερα τα δείγματα του τρέχοντος κόμβου σε υποσύνολα της μιας ή της άλλης κατηγορίας. Στη συνέχεια, ο αλγόριθμος C4.5 προχωρά αναδρομικά έως ότου ικανοποιήσει κάποια κοινώς χρησιμοποιούμενα κριτήρια διακοπής, όπως ο ελάχιστος αριθμός δειγμάτων σε ένα τερματικό κόμβο. Η LDA είναι μια γνωστή γενική μέθοδος που χρησιμοποιείται για τη μείωση της διαστασιμότητας και την ταξινόμηση. Η LDA προσπαθεί να βρει ένα χώρο χαμηλής διαστασιμότητας για διάφορες κατηγορίες. Σε αυτόν τον χώρο, οι αποστάσει μεταξύ των δειγμάτων από διαφορετικές κατηγορίες είναι μεγάλες, αλλά οι αποστάσεις μεταξύ των δειγμάτων της ίδιας κατηγορίας είναι μικρές. Κατά τη διαδικασία εκμάθησης, η LDA μπορεί να λάβει μια συνάρτηση για την προβολή των δειγμάτων από διαφορετικές κατηγορίες στο χώρο χαμηλής διάστασης. Εφαρμόζει μια eigendecomposition στους πίνακες διασποράς για τον υπολογισμό της βέλτιστης προβολής. Με βάση την προβολή, η LDA μπορεί να εξάγει ένα μοντέλο ταξινόμησης που επικεντρώνεται στη συσχέτιση μεταξύ πολλαπλών ανεξάρτητων μεταβλητών και μιας κατηγορικής εξαρτημένης μεταβλητής σχηματίζοντας ένα μία σύνθεση των ανεξάρτητων μεταβλητών. Ένα ELM είναι ένα υπολογιστικά ευφυές μοντέλο νευρωνικού δικτύου με μη επαναληπτική μάθηση. Επιλέγει τυχαία τα βάρη εισόδου και καθορίζει αναλυτικά τα βάρη εξόδου του νευρωνικού δικτύου. Ο χρόνος μοντελοποίησης του ELM είναι ταχύτερος από τους παραδοσιακούς αλγορίθμους μάθησης δικτύων, όπως το γνωστό back-propagation νευρωνικό δίκτυο. Μειώνει επίσης πολλές από τις δυσκολίες στη ρύθμιση των παραμέτρων, συμπεριλαμβανομένων των κριτηρίων διακοπής, του learning rate, των learning epochs,κ.λπ. Το CART μοντέλο πρόβλεψης κατασκευάστηκε χρησιμοποιώντας το πακέτο raprt R της έκδοσης 4.1.15 (R core team, Vienna, Austria). Για την αναζήτηση του καλύτερου συνόλου παραμέτρων για τη δημιουργία ενός υποσχόμενου μοντέλου CART, το πακέτο OptimClassifier R της έκδοση 0.1.5 (R core team, Βιέννη, Αυστρία) εφαρμόστηκε για τις παραμέτρους του βάθους του δέντρου, τον αριθμό των παρατηρήσεων σε κάθε τερματικό κόμβο και το pruning του δέντρου. Για τη δημιουργία του μοντέλου C4.5, χρησιμοποιήθηκε το πακέτο RWeka R πακέτο της έκδοσης 0.4-42 (R core team, Βιέννη, Αυστρία). Για την εύρεση της καλύτερης παραμέτρου για το κόστος για την κατασκευή ενός αποτελεσματικού μοντέλου C4.5, χρησιμοποιήθηκε το πακέτο caret R της

έκδοσης 6.0-84 (R core team, Βιέννη, Αυστρία). Η LDA υλοποιήθηκε με τη χρήση του πακέτου MASS R της έκδοσης 7.3-51.5 (R core team, Βιέννη, Αυστρία). Για τη δημιουργία ενός μοντέλου LDA χρησιμοποιήθηκαν οι προεπιλεγμένες ρυθμίσεις. Το μοντέλο ELM κατασκευάστηκε με την εφαρμογή του πακέτου elmNN R της έκδοσης 1.0 (R core team, Vienna, Austria). Η προεπιλεγμένη συνάρτηση ενεργοποίησης σε αυτό το πακέτο είναι η radial basis. Για την αναζήτηση του καλύτερου αριθμό κρυφών νευρώνων που θα δημιουργούσε υποσχόμενα μοντέλα ELM, το πακέτο caret R της έκδοσης 6.0-84 (R core team, Vienna, Austria) χρησιμοποιήθηκε για τη ρύθμιση σημαντικών υπερπαραμέτρων. Η ακρίβεια της ταξινόμησης αξιολογήθηκε με τη χρήση της καμπύλης ROC για την εκτίμηση της περιοχής κάτω από την καμπύλη (AUC). Στην παρούσα μελέτη εξετάστηκαν η ακρίβεια, η ευαισθησία και η ειδικότητα. Στην παρούσα μελέτη, εφαρμόστηκαν προσεγγίσεις μηχανικής μάθησης σε ένα σύνολο δεδομένων εξετάσεων υγείας ενηλίκων για να προβλεφθούν ασθενείς με υψηλό κίνδυνο CKD με βάση τα δεδομένα κάθε μεταβλητής για κάθε ασθενή. Στόχος ήταν να συγκριθούν διαφορετικά μοντέλα ταξινόμησης και να εντοπιστεί το πιο αποτελεσματικό. Τα δημογραφικά στοιχεία των ατόμων περιγράφονται στον Πίνακα 3.14 (Shih *et.al,* 2020). Οι ανεξάρτητες μεταβλητές στην ανάλυση ήταν το φύλο, η ηλικία, ο αριθμός των ερυθρών αιμοσφαιρίων (RBC), το επίπεδο γλυκόζης (GLU), τα τριγλυκερίδια (TG), η ολική χοληστερόλη(T-CHO), η χοληστερόλη λιποπρωτεϊνών υψηλής πυκνότητας (HDL-C), η χοληστερόλη λιποπρωτεϊνών χαμηλής πυκνότητας (LDL-C), η λευκωματίνη (ALB), η πρωτεϊνουρία (PRO) και ο λόγος πρωτεΐνης ούρων προς κρεατινίνη (UPCR). Το t-test χρησιμοποιήθηκε για τη σύγκριση των μέσων όρων της ηλικίας για CKD και μη CKD και χρησιμοποιήθηκε το τεστ chi-square για να αξιολογηθούν οι συσχετίσεις μεταξύ της εξαρτημένης μεταβλητής και όλων των ανεξάρτητων μεταβλητών εκτός από την ηλικία. Διαπιστώθηκε ότι η ηλικία ($p < 0,001$), η διαφορά φύλου ($p < 0,001$), οι φυσιολογικές ή μη φυσιολογικές επιδόσεις του RBC ($p < 0,001$), GLU ($p = 0,004$), TG ($p = 0,011$), HDL ($p = 0,029$), PRO ($p < 0,001$) και UPRC ($p < 0,01$) συσχετίστηκαν σημαντικά με τον επιπολασμό της CKD. Τα αποτελέσματα του t-test έδειξαν ότι η ομάδα με CKD στην μέση ηλικία ήταν σημαντικά διαφορετική από την ομάδα χωρίς CKD. Η ανάλυση της chi-square test έδειξε ότι τα δύο φύλα είχαν ανόμοια παρέμβαση στον επιπολασμό της CKD, και η ζευγαρωμένη σύγκριση αποκάλυψε ότι το ποσοστό των ανδρών στην ομάδα με CKD ήταν υψηλότερο (48,3% έναντι 39,6%) από ό,τι σε σχέση με το αντίστοιχο ποσοστό στην ομάδα των ατόμων χωρίς CKD.

Πίνακας 3.14

Δημογραφικά χαρακτηριστικά των ατόμων

Characteristic	Non-CKD	CKD	*p*-Value
N (%)	14,169 (73.5%)	5101 (26.5%)	
Gender			
Male	5608 (39.6%)	2465 (48.3%)	<0.001 **
Female	8561 (60.4%)	2636 (51.7%)	
Age			
Mean (±SD)	63.37 ± 11.56	69.19 ± 10.74	<0.001 *
RBC			
Normal	11,460 (80.9%)	3917 (76.8%)	<0.001 **
Abnormal	2709 (19.1%)	1184 (23.2%)	
GLU			
Normal	11,502 (81.2%)	1055 (20.7%)	0.004 **
Abnormal	2667 (18.8%)	4046 (79.3%)	
TG			
Normal	5878 (41.5%)	2012 (39.4%)	0.011 *
Abnormal	8291 (58.5%)	3089 (60.6%)	
T-CHO			
Normal	9198 (64.9%)	3284 (64.4%)	0.491
Abnormal	4971 (35.1%)	1817 (35.6%)	
HDL-C			
Normal	11,954 (84.4%)	4369 (85.6%)	0.029 *
Abnormal	2215 (15.6%)	732 (14.4%)	
LDL-C			
Normal	11,400 (80.5%)	4095 (80.3%)	0.782
Abnormal	2769 (19.5%)	1006 (19.7%)	
ALB			
Normal	14,162 (100.0%)	5097 (99.9%)	0.457
Abnormal	7 (0.0%)	4 (0.1%)	
PRO			
Normal	9203 (65.0%)	915 (17.9%)	<0.001 *
Abnormal	4966 (35.0%)	4186 (82.1%)	
UPCR			
Normal	12,364 (87.3%)	1639 (32.1%)	<0.001 *
Abnormal	1805 (12.7%)	3462 (67.9%)	

** *p*-value < 0.01; * *p*-value < 0.05.

Ένα υψηλότερο ποσοστό ατόμων με μη φυσιολογικό RBC βρέθηκε στην ομάδα CKD από ό,τι στην ομάδα ομάδα χωρίς CKD (23,2% έναντι 19,1%), και ένα υψηλότερο ποσοστό φυσιολογικού GLU βρέθηκε στην ομάδα με CKD από ό,τι στην ομάδα χωρίς CKD (20,7% έναντι 18,8%). Η ομάδα CKD περιείχε υψηλότερο ποσοστό ατόμων με μη φυσιολογικά τριγλυκερίδια (TG) σε σχέση με την ομάδα χωρίς CKD (60,6% έναντι 58,5%), καθώς και υψηλότερο ποσοστό ατόμων με φυσιολογικές λιποπρωτεΐνες υψηλής πυκνότητας (HDL) (85,6% έναντι 84,4%). Η ομάδα με CKD περιείχε υψηλότερο ποσοστό ατόμων με μη φυσιολογική πρωτεϊνουρία (PRO) σε σχέση με την ομάδα χωρίς CKD (82,1% έναντι 35,0%) και ένα υψηλότερο ποσοστό ατόμων με παθολογική UPRC σε σύγκριση με την ομάδα χωρίς CKD (67,9% έναντι 12.7%). Δεν βρέθηκαν σημαντικές διαφορές μεταξύ των φυσιολογικών και μη φυσιολογικών επιδόσεων της T-CHO (p = 0,491), των λιποπρωτεϊνών χαμηλής πυκνότητας (p = 0,782) ή της ALB (p = 0,457). Επιλέχθηκαν τυχαία 15.416 ασθενείς (80% του συνόλου των ασθενών) ως training set , ενώ οι υπόλοιποι 3854 ασθενείς (20% του συνόλου των ασθενών) χρησιμοποιήθηκαν ως test set για τη μέτρηση της προβλεπτικής ικανότητας των

τεσσάρων μεθόδων εκτός δείγματος. Επιπλέον, εφαρμόστηκε μια μέθοδος 10-fold cross validation που χρησιμοποιήθηκε για την εκπαίδευση των μοντέλων ταξινόμησης των τεσσάρων μεθόδων. Στον Πίνακα 3.15 (Shih *et.al,* 2020) παρουσιάζονται τα αποτελέσματα ταξινόμησης των μεθόδων CART, ELM, C4.5 και LDA. Δείχνει ότι οι τιμές AUC των μοντέλων CART, ELM, C4.5 και LDA ήταν 0,779, 0,692, 0,788 και 0,773, αντίστοιχα. Το μοντέλο C4.5 παρείχε την υψηλότερη τιμή AUC, ακολουθούμενο από το CART, το LDA και το το μοντέλο ELM, αντίστοιχα. Οι τιμές ακρίβειας, ευαισθησίας και ειδικότητας του μοντέλου C4.5 είναι όλες μεγαλύτερες από τα υπόλοιπα τρία μοντέλα. Στο Σχήμα 3.4 (Shih *et.al,* 2020) παρουσιάζονται οι καμπύλες ROC των τεσσάρων μοντέλων ταξινόμησης για την εμφάνιση πρώιμης CKD. Ακόμη, το σχήμα αυτό απεικονίζει ότι η μέθοδος C4.5 παρουσίασε την καλύτερη προβλεπτική ικανότητα σε σύγκριση με τα τρία συγκριτικά μοντέλα και αποτελεί μια πολλά υποσχόμενη μέθοδο για την πρώιμη πρόβλεψη της CKD.

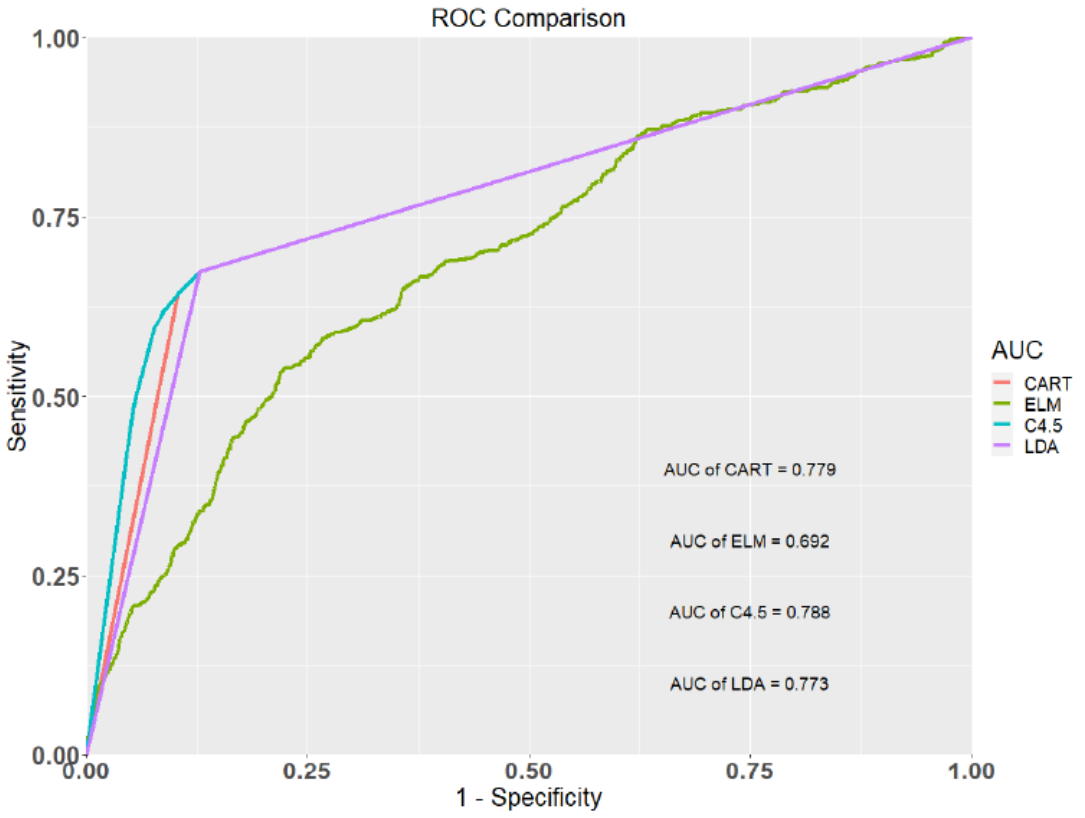
Πίνακας 3.15

Αποτελέσματα ταξινόμησης των τεσσάρων μεθόδων

Methods	Accuracy	Sensitivity	Specificity	AUC
Classification and Regression Tree (CART)	0.819	0.670	0.871	0.779
Extreme Learning Machine (ELM)	0.715	0.539	0.777	0.692
C4.5	0.820	0.673	0.872	0.788
Linear Discriminant Analysis (LDA)	0.818	0.669	0.868	0.773

Σχήμα 3.4

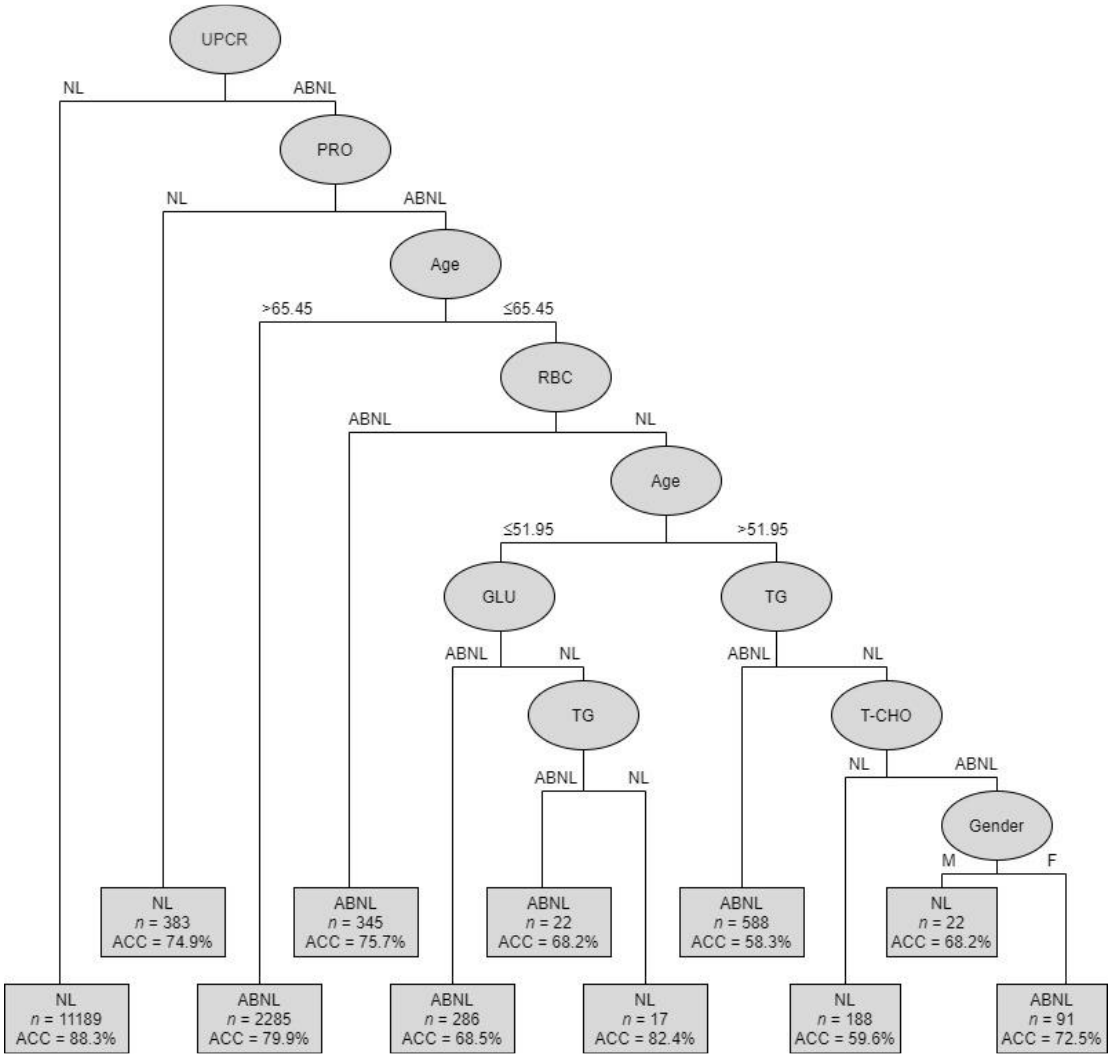
Καμπύλες ROC των τεσσάρων μεθόδων



Στόχος της ανάλυσης ήταν να εντοπιστούν οι σημαντικότεροι παράγοντες κινδύνου από δέκα πιθανούς παράγοντες: φύλο, ηλικία, RBC, GLU, TG, T-CHO, HDLC, LDLC, ALB, PRO και UPCR. Τα αποτελέσματά έδειξαν ότι η μέθοδος C4.5 μπορεί να παράγει την καλύτερη ταξινόμηση και τα πιο υποσχόμενα αποτελέσματα για την πρόβλεψη της CKD. Η μέθοδος C4.5 αυτοματοποιεί την ανίχνευση των συσχετίσεων μεταξύ των προγνωστικών παραγόντων και των αποτελεσμάτων και τις αλληλεπιδράσεις μεταξύ των προγνωστικών παραγόντων και παρέχει μέτρα για τους προγνωστικούς παράγοντες. Στο Σχήμα 3.5 (Shih *et.al*, 2020) παρουσιάζεται το δέντρο ταξινόμησης των προγνωστικών παραγόντων CKD με τη μέθοδο C4.5. Ο Πίνακας 3.16 (Shih *et.al*, 2020) παρουσιάζει συνοπτικά την ακρίβεια του μοντέλου με την προσθαφαίρεση των μεταβλητών.

Σχήμα 3.5

Δέντρο ταξινόμησης που απεικονίζει τους προγνωστικούς παράγοντες CKD της μεθόδου C4.5.



Πίνακας 3.16

Ακρίβεια του μοντέλου με την προσθαφαίρεση των μεταβλητών.

Rules No.	Combinations of Condition Variables	Cases of (Ab)normal		Accuracy
1	UPCR (NL)	9879	NL	88.3%
2	UPCR (ABNL) + PRO (NL)	287	NL	74.9%
3	UPCR (ABNL)+PRO (ABNL) + Age (>65.45)	1826	ABNL	79.9%
4	UPCR (ABNL) + PRO (ABNL) + Age (≤65.45) + RBC (ABNL)	261	ABNL	75.7%
5	UPCR (ABNL) + PRO (ABNL) + Age (≤65.45) + RBC (NL) + Age (≤51.95) + GLU (ABNL)	196	ABNL	68.5%
6	UPCR (ABNL) + PRO (ABNL) + Age (≤65.45) + RBC(NL) + Age (>51.95)|+ TG(ABNL)	343	ABNL	58.3%
7	UPCR (ABNL) + PRO (ABNL) + Age (≤65.45) + RBC (NL) + Age (≤51.95) + GLU (ABNL) + TG (ABNL)	15	ABNL	68.2%
8	UPCR (ABNL) + PRO (ABNL) + Age (≤65.45) + RBC(NL) + Age (≤51.95) + GLU (ABNL) + TG(NL)	14	NL	82.4%
9	UPCR (ABNL) + PRO (ABNL) + Age (≤65.45) + RBC (NL) + Age (≤51.95) + GLU (ABNL) + TG (NL) + T-CHO (NL)	112	NL	59.6%
10	UPCR (ABNL) + PRO (ABNL) + Age (≤65.45) + RBC (NL) + Age (≤51.95) + GLU (ABNL) + TG (NL) + T-CHO (ABNL) + Gender (M)	15	NL	68.2%
11	UPCR (ABNL) + PRO (ABNL) + Age (≤65.45) + RBC (NL) + Age (≤51.95) + GLU (ABNL) + TG (NL) + T-CHO (ABNL) + Gender (F)	66	ABNL	72.5%

UPCR: Urine protein and creatinine ratio, PRO: Proteinuria, RBC: Red blood cells, GLU: Glucose Fasting, TG: Triglycerides, T-CHO: Total Cholesterol.

Τα άτομα χωρίστηκαν σε 11 υποομάδες, από τον κόμβο της ρίζας του δέντρου προς τους κόμβους των φύλλων, μέσω διαφορετικών κλάδων. Όπως εξηγήθηκε προηγουμένως, η μεταβλητή UPCR έχει μεγάλη επίδραση στην ερμηνεία του eGFR και, ως εκ τούτου, προσδιορίστηκε ως ο κόμβος ρίζας του ταξινομημένου δέντρου απόφασης. Το πρώτο επίπεδο του δέντρου απόφασης προέκυψε από τον καθοριστικό παράγοντα: UPCR. Η ακρίβεια (ACC) που επιτεύχθηκε ήταν 88,3% σε όλα τα 11.189 δείγματα. Αυτό σημαίνει ότι από τους 11.189 ασθενείς χωρίς CKD, 9879 ασθενείς ταξινομήθηκαν σωστά με τη χρήση της μεταβλητής UPCR. Το δεύτερο επίπεδο του δέντρου απόφασης δημιουργήθηκε με τον ακόλουθο τρόπο παράγοντες: PRO, ηλικία, RBC, GLU, TG, T-CHO και φύλο. Ως εκ τούτου, το δέντρο απόφασης μπορεί να χωριστεί σε μη φυσιολογικές (ABNL, CKD) ή φυσιολογικές (NL, μη CKD) καταστάσεις. Η ακρίβεια κυμάνθηκε από 58,3% έως 88,3%. Το δεύτερο επιπέδο του δέντρου απόφασης προέκυψε από τους ακόλουθους προσδιοριστικούς παράγοντες: UPCR (με ABNL)+PRO (με NL) και η ακρίβεια που προέκυψε ήταν 0,749 σε 383 δείγματα. Το τρίτο επίπεδο του δέντρου απόφασης προέκυψε από τους ακόλουθους προσδιοριστικούς παράγοντες: UPCR (ABNL)+PRO (ABNL)+ηλικία (>65,45), και η ακρίβεια που προέκυψε ήταν 0,799 σε 2285 δείγματα. Το τέταρτο επίπεδο του δέντρου απόφασης προέκυψε από τους ακόλουθους προσδιοριστικούς παράγοντες: UPCR (ABNL) + PRO (ABNL) + ηλικία (<65,45) + RBC (ABNL), και η ακρίβεια που προέκυψε ήταν 0,757 σε 345 δείγματα. Το αριστερό πέμπτο επίπεδο του δέντρου απόφασης προέκυψε από τους ακόλουθους προσδιοριστικούς παράγοντες: UPCR (ABNL) + PRO (ABNL) + ηλικία (<65,45) + RBC (NL) + ηλικία (<51,95) + GLU (ABNL), και η ακρίβεια που προέκυψε ήταν 0,685 σε όλα τα 286 δείγματα. Εντω μεταξύ, για τους ακόλουθους προσδιοριστικούς παράγοντες: UPCR (ABNL) +PRO (ABNL) + ηλικία (<65,45) + RBC(NL) + ηλικία (<51,95) + GLU (NL) + TG (ABNL), η ακρίβεια που επιτεύχθηκε ήταν 0,682 σε όλα τα 22 δείγματα. Το πέμπτο επίπεδο του δεξιού δέντρου απόφασης προέκυψε από τους ακόλουθους καθοριστικούς παράγοντες: UPCR (ABNL) + PRO (ABNL) + ηλικία (<65,45) + RBC (NL) + ηλικία (<51,95) + TG (NL)+T-CHO (ABNL) + φύλο

(άνδρας), και η ακρίβεια που επιτεύχθηκε ήταν 0,682 σε 22 δείγματα. Εντωμεταξύ, για τους ακόλουθους καθοριστικούς παράγοντες: UPCR (ABNL) + PRO (ABNL) + ηλικία (<65,45) + RBC (NL) + ηλικία (<51,95) + TG (NL) + T-CHO (ABNL) + φύλο (γυναίκα), η ακρίβεια που επιτεύχθηκε ήταν 0,725 σε 91 δείγματα. Με τη χρήση αυτώ των διαφορετικών μοντέλων, οι κλινικοί γιατροί μπορούν να προσδιορίσουν τους συνδυασμούς παραγόντων για μια πάθησηπου ενδιαφέρονται. Τα ευρήματα της παρούσας μελέτης ήταν συνεπή με εκείνα προηγούμενων εκθέσεων, συμπεριλαμβανομένων της πιο πρόσφατης έκθεσης του Εθνικού Ινστιτούτου Ερευνών Υγείας για τις νεφρικές παθήσεις στην αναλογία πρωτεΐνης ούρων προς κρεατινίνη (UPCR) και τον αριθμό των ερυθρών αιμοσφαιρίων (RBC), καθώς και την έκθεση των Xiao *et.al* (2019) σχετικά με την πρόβλεψη της χρόνιας νεφρικής νόσου στην πρωτεϊνουρία (PRO). Τα ευρήματα της αλβουμίνης (ALB) και των επιπέδων γλυκόζης (GLU) συνάδουν με προηγούμενες μελέτες που ακολουθούν τους Korbut *et.al* (2019) και Kshirsagar *et.al* (2008). Ομοίως, οι Xue *et.al* (2019), οι Mahmood *et.al* (2017) και οι Kshirsagar *et.al* (2008) ανέφεραν ότι τα τριγλυκερίδια (TG), η ηλικία και το φύλο είναι κρίσιμα για την πρόβλεψη της χρόνιας νεφρικής νόσου. Όπως υποδηλώνεται από τα αποτελέσματά της παρούσας μελέτης, το κύριο ζήτημα είναι πώς να προβλεφθεί η CKD των ατόμων που είναι ασυμπτωματικοί και που υποβάλλονται μόνο σε ένα πρόγραμμα υγειονομικής εξέτασης ρουτίνας για ενήλικες. Μια ολοκληρωμένη, κλινική προσέγγιση της πρόληψης που λαμβάνει υπόψη όλους αυτούς τους παράγοντες απαιτείται για την επιτυχή των εκθέσεων υψηλού κινδύνου στον ενήλικο πληθυσμό. Συνεπώς, η βέλτιστη προκλινική διαχείριση της πρώιμης CKD θα ωφεληθεί από την καλύτερη κατανόηση της φύσης. Πολλοί από τους παράγοντες κινδύνου που ενδεχομένως σχετίζονται με την πρώιμη συνειδητοποίηση της CKD, δηλ, διαχείριση της υπέρτασης, είναι ενδιαφέροντες και χρήζουν περαιτέρω διερεύνησης. Τα εμπειρικά αποτελέσματα έδειξαν ότι η C4.5 υπερτερεί ελαφρώς έναντι των μεθόδων CART και LDA. Αλλά, καθώς η παρούσα μελέτη είχε ως στόχο τη διερεύνηση σημαντικών παραγόντων κινδύνου της πρώιμης CKD και τη συσχέτισης μεταξύ τους, τα αποτελέσματα της καλύτερης μεθόδου με υποσχόμενες επιδόσεις είναι τα καταλληλότερα για περαιτέρω συζήτηση. Έτσι, το δέντρο ταξινόμησης που απεικονίζει τους παράγοντες πρόβλεψης της CKD του C4.5 συζητείται στην παρούσα μελέτη. Χρησιμοποιώντας διαφορετικά είδη δεδομένων CKD για τη σύγκριση της αποτελεσματικότητας των C4.5, CART και LDA για την πρόβλεψη της πρώιμης CKD μπορεί να θεωρηθεί ως μία από τις μελλοντικές ερευνητικές κατευθύνσεις. Στην παρούσα μελέτη χρησιμοποιήθηκαν τα ELM, C4.5, CART και LDA για την πρόβλεψη της πρώιμης CKD. Η LDA είναι μια στατιστική μέθοδος και το σημαντικό χαρακτηριστικό της είναι ότι προβάλλει τα δεδομένα σε ένα διάνυσμα χαμηλότερης διάστασης στο χώρο, ο οποίος είναι ένας υποχώρος με μεγαλύτερη διακριτική ικανότητα, δεδομένου ότι ο λόγος της απόστασης μεταξύ των κλάσεων προς την απόσταση εντός της κλάσης μεγιστοποιείται. Η ELM είναι μια μέθοδος νευρωνικών δικτύων και το κύριο χαρακτηριστικό της είναι ότι οι παράμετροι των κρυφών στρωμάτων παράγονται τυχαία και ανεξάρτητα από τα δείγματα εκπαίδευσης με συνέπεια να έχει μεγαλύτερη ταχύτητα εκμάθησης σε σύγκριση με τους αλγορίθμους με την παραδοσιακή εκμάθηση νευρωνικών δικτύων. Η CART και η C4.5 είναι και οι δύο μέθοδοι δέντρων απόφασης. Το χαρακτηριστικό ενός δέντρου αποφάσεων είναι η χρήση ενός συνόλου συνθηκών "αν-τότε" για την ταξινόμηση των περιπτώσεων. Το κύριο χαρακτηριστικό του

CART είναι ότι παράγει μόνο δυαδικά δέντρα με βάση το δείκτη Gini. Το C4.5 τον λόγο κέρδους για τη δημιουργία δέντρων απόφασης τα οποία περιλαμβάνουν διαχωρισμό πολλών κλάδων (δηλαδή όχι μόνο δυαδικό) σε κάθε κόμβο. Ωστόσο, το μοντέλο ELM είχε κακές επιδόσεις στην παρούσα μελέτη. Το ELM είναι ένας αλγόριθμος νευρωνικών δικτύων και ο μηχανισμός μοντελοποίησής του είναι διαφορετικός από αυτόν των άλλων τριών μεθόδων, C4.5, CART, και LDA. Οι αλγόριθμοι νευρωνικών δικτύων είναι ισχυρά εργαλεία για την ανάλυση κλινικών δεδομένων για την πρόβλεψη ενός αποτελέσματος. Στην πραγματικότητα, μπορεί να είναι χρήσιμοι αλλά το νευρωνικό δίκτυο δεν θα μας δώσει καμία πληροφορία διορατικότητας σε αυτή τη μελέτη, καθώς δεν μπορεί να χρησιμοποιηθεί για την επιλογή σημαντικών μεταβλητών.

Η χρόνια νεφρική νόσος (CKD) αποτελεί μείζον παγκόσμιο πρόβλημα δημόσιας υγείας, αλλά η διάγνωση σε πρώιμο στάδιο είναι προβληματική λόγω της ασυμπτωματικής εμφάνισης. Επί του παρόντος, δεν υπάρχουν ευρέως αποδεκτές προγνωστικά εργαλεία. Ως εκ τούτου, οι γιατροί πρέπει να λαμβάνουν κλινικές αποφάσεις σχετικά με το ποιοι ασθενείς θα θεραπευθούν. Στην παρούσα μελέτη, στόχος ήταν να διερευνηθούν σημαντικοί παράγοντες κινδύνου της πρώιμης CKD και να συζητηθούν οι συσχετίσεις μεταξύ τους. Είναι σημαντικό ότι η ενημέρωση για την πρώιμη CKD είναι απαραίτητη για τους δυνητικούς ασθενείς να συμμετάσχουν και να συμμορφωθούν με τα προγράμματα υγειονομικής εξέτασης, κάτι που έχει μεγάλη κλινική και οικονομική σημασία. Επιπλέον, από όσα γνωρίζουν οι συγγραφείς, δεν υπάρχουν μελέτες που να χρησιμοποιούν τεχνικες ταξινόμησης εξόρυξης δεδομένων για τη δημιουργία προγνωστικών μοντέλων για διασικασίες πρόβλεψης της πρώιμης CKD. Στην παρούσα μελέτη, εφαρμόστηκαν εννέα μεταβλητές φυσικής εξέτασης και δύο δημογραφικές παραμέτρους για τον προσδιορισμό των παραγόντων κινδύνου CKD, χρησιμοποιώντας τέσσερις αλγορίθμους εξόρυξης δεδομένων. Ο αλγόριθμος C4.5 απέδωσε ένα output οκτώ μεταβλητών που ήταν σημαντικές για την πρόβλεψη της πρώιμης CKD. Διαπιστώθηκε ότι η μείωση του αριθμού των μεταβλητών αύξησε την ακρίβεια των αποτελεσμάτων. Ένα άλλο σημαντικό εύρημα αυτής της μελέτης ήταν ότι η μέθοδος C4.5 είχε την καλύτερη προβλεπτική ικανότητα σε σύγκριση με τα άλλα τρία μοντέλα σύγκρισης. Το C4.5 αποκάλυψε επίσης ότι οι διαφορετικοί συνδυασμοί μεταβλητών του συνόλου δεδομένων οδήγησαν σε διαφορετικά ποσοστά ακρίβειας που κυμαίνονταν από 59,6% έως 88,3%. Επίσης το UPCR, το PRO, η ηλικία, το RBC, το GLU, το TG, το T-CHO και το φύλο είχαν σημαντικές επιπτώσεις στην προβλεψιμότητα των μοντέλων, ενώ άλλοι προγνωστικοί παράγοντες, όπως η HDL, η LDL και η ALB, ήταν λιγότερο σημαντικοί. Με την αργή εξέλιξη της CKD, η έγκαιρη ανίχνευση και η αποτελεσματική θεραπεία είναι οι μόνοι τρόποι για τη μείωση της θνησιμότητας. Η έγκαιρη εκτίμηση του κινδύνου της CKD και η κατάλληλη παρακολούθηση στην κοινότητα είναι σημαντικές για την πρόληψη περαιτέρω νεφρικής βλάβης σε ασθενείς με πρώιμη CKD. Συμπερασματικά, η παρούσα μελέτη παρουσιάζει στοιχεία της δυνατότητας εφαρμογής ενός συνόλου δεδομένων των εξετάσεων υγείας ενηλίκων και της ευρωστίας των τεσσάρων μοντέλων για την κλινική εκτίμηση του κινδύνου της πρώιμης CKD (Shih *et.al,* 2020).

Κ Ε Φ Α Λ Α Ι Ο 4

Πρόβλεψη για την επίτευξη της ανοσίας της αγέλης για τη νόσο COVID-19

Η παρούσα εφαρμογή πραγματοποιήθηκε με σκοπό να παρουσιάσουμε την εφαρμογή μεθόδων της Στατιστικής Μηχανικής Μάθησης σε δεδομένα που σχετίζονται με την δημόσια υγεία. Μάλιστα, η συγκεκριμένη εφαρμογή αφορά την πανδημία του κορονοϊού (COVID-19). Ο στόχος αυτής της εφαρμογής ήταν να εξάγουμε πληροφορίες ώστε να απαντήσουμε ερωτήματα σχετικά με το ποια χώρα χρησιμοποιεί ποιο εμβόλιο, σε ποια χώρα το πρόγραμμα εμβολιασμού είναι πιο προχωρημένο, πού εμβολιάζονται περισσότεροι άνθρωποι ανά ημέρα, αλλά σε ποσοστό επί τοις εκατό από το σύνολο του πληθυσμού, ποιο εμβόλιο είναι πιο δημοφιλή και σε ποια χώρα. Ακόμη, θέλαμε να προβλέψουμε με την βοήθεια στατιστικών μοντέλων το πότε μπορούμε να περιμένουμε ότι μία χώρα θα χορηγήσει αρκετά εμβόλια για να επιτύχει ανοσία της αγέλης. Με βάση τις μελέτες που γίνονται, ο Π.Ο.Υ. (WHO newsletter, 2020) δεν είναι σε θέση να προσδιορίσει το ποσοστό εμβολιασμού που απαιτείται για την ανοσία αγέλης του Covid-19. Από την άλλη πλευρά, το τμήμα Δημόσιας Υγείας John Hopkins αναφέρει ότι το απαιτούμενο ποσοστό εμβολιασμού είναι περίπου 70%. Συνεπώς, θα το χρησιμοποιήσουμε αυτό ως σημείο αναφοράς (D'Souza & Dowdy, 2021). Για την εξαγωγή των αποτελεσμάτων χρησιμοποιήθηκε η γλώσσα προγραμματισμού Python.

4.1 Πανδημία του κορονοϊού (COVID-19)

Η πανδημία του κορονοϊού (COVID-19) είναι μια τρέχουσα πανδημία η οποία προκλήθηκε από τον κορονοϊό SARS-CoV-2 και αναγνωρίστηκε για πρώτη φορά στην πόλη Ουχάν, που είναι πρωτεύουσα της επαρχίας Χουπέι της Κίνας, τον Δεκέμβριο του 2019. Οι απόπειρες περιορισμού της νόσου απέτυχαν, με αποτέλεσμα ο Παγκόσμιος Οργανισμός Υγείας να ανακοινώσει ότι αποτελεί Έκτακτη Ανάγκη Δημόσιας Υγείας Διεθνούς Ενδιαφέροντος στις 30 Ιανουαρίου 2020 και στις 11 Μαρτίου ότι αποτελεί πανδημία. Η συγκεκριμένη πανδημία έχει προκαλέσει πάνω από 6 εκατομμύρια θανάτους σε πάνω από 570 εκατομμύρια μολύνσεις, γεγονός που την καθιστά μία από τις πιο θανατηφόρες στην ιστορία. Ο ιός μεταδίδεται μεταξύ των ανθρώπων μέσω των σταγονιδίων που παράγονται όταν οι άνθρωποι φτερνίζονται ή βήχουν. Ο χρόνος μεταξύ της έκθεσης και της εμφάνισης συμπτωμάτων είναι συνήθως από 2 έως 14 ημέρες. Τα συμπτώματα μπορεί να περιλαμβάνουν πυρετό, βήχα και δυσκολίες στην αναπνοή, αλλά μπορεί να είναι μεταδοτική

κατά τη διάρκεια αυτής της περιόδου και μετά την αποκατάσταση, ενώ επιστημονικές έρευνες υποστηρίζουν ότι πιθανή απώλεια γεύσης και όσφρησης αποτελούν συμπληρωματικές ενδείξεις μόλυνσης από τον ιό. Οι επιπλοκές μπορούν να περιλαμβάνουν πνευμονία και σύνδρομο οξείας αναπνευστικής δυσχέρειας. Άλλες επιπλοκές περιλαμβάνουν θρομβοεμβολικά επεισόδια και το σύνδρομο μακρού COVID. Ακόμη, έχουν εμφανιστεί διάφορες παραλλαγές του ιού με διαφορές στη νοσογόνο δύναμη και τη μεταδοτικότητα.

Κατά τη διάρκεια της πανδημίας αναπτύχθηκαν αποτελεσματικά και ασφαλή εμβόλια, εγκεκριμένα από επίσημες ρυθμιστικές αρχές. Στις 2 Δεκεμβρίου του 2020, η αρμόδια επιτροπή MHRA στο Ηνωμένο Βασίλειο χορήγησε την πρώτη άδεια εμβολίου έναντι του COVID-19, εγκρίνοντας τη σχετική αίτηση των εταιρειών Pfizer & BioNTech, αρχίζοντας τους εμβολιασμούς την Τρίτη 8-12-2020. Επίσης, άλλα μέτρα πρόληψης περιλαμβάνουν το πλύσιμο των χεριών, η διατήρηση της απόστασης άνω των 2 μέτρων από ανθρώπους που βήχουν, τη χρήση προστατευτικής μάσκας, τη βελτίωση του αερισμού των χώρων και την καραντίνα όσων ήρθαν σε επαφή με επιβεβαιωμένο κρούσμα. Οι κυβερνήσεις πολλών χωρών έλαβαν επιπλέον μέτρα, όπως περιορισμός μετακινήσεων και δραστηριοτήτων και έλεγχο των πολιτών για COVID-19.

Η πανδημία προκάλεσε εκτεταμένη αναστάτωση στην οικονομία και στις κοινωνικές δραστηριότητες, με αποτέλεσμα τη μεγαλύτερη ύφεση μετά το 2008 και ελλείψεις στις εφοδιαστικές αλυσίδες. Πολλά πανεπιστήμια και σχολεία έμειναν κλειστά, ενώ εκδηλώσεις ακυρώθηκαν. Υπήρξε εκτεταμένη παραπληροφόρηση από κοινωνικά δίκτυα και πολιτική ένταση. Η πανδημία επίσης ανέδειξε προβλήματα ρατσισμού και άνισης πρόσβασης σε υγειονομικές υπηρεσίες και συζήτηση για την προστασία της δημόσιας υγείας έναντι των ατομικών ελευθεριών (https://en.wikipedia.org/wiki/COVID-19).

4.2 Σύνολο Δεδομένων

Το σύνολο δεδομένων που χρησιμοποιήθηκε για την παρούσα έρευνα, συλλέχθηκε από τον αποθηκευτικό χώρο του Our World in Data GitHub για τον covid-19, όπου συγχωνεύθηκαν και μεταφορτώθηκαν. Πρόκειται για δεδομένα που αφορούν όλες τις χώρες και περιέχουν τις εξής μεταβλητές:

- country (Χώρα): Πρόκειται για τη χώρα για την οποία παρέχονται οι πληροφορίες εμβολιασμού
- iso_code: Κωδικός ISO (κώδικες για τα ονόματα των χωρών) για την κάθε χώρα
- date (Ημερομηνία): Ημερομηνία καταχώρησης των δεδομένων- για ορισμένες από τις ημερομηνίες έχουμε μόνο τους ημερήσιους εμβολιασμούς, για άλλες μόνο το (αθροιστικό) σύνολο

- total_vaccinations (Συνολικός αριθμός εμβολιασμών): Πρόκειται για τον απόλυτο αριθμό των συνολικών εμβολιασμών στη χώρα
- people_vaccinated (Συνολικός αριθμός εμβολιασμένων ατόμων): Ένα άτομο, ανάλογα με το πρόγραμμα εμβολιασμού, θα λάβει ένα ή περισσότερα (συνήθως 2) εμβόλια- σε μια συγκεκριμένη στιγμή, ο αριθμός των εμβολιασμών μπορεί να είναι μεγαλύτερος από τον αριθμό των ατόμων
- people_fully_vaccinated (Συνολικός αριθμός ατόμων που έχουν εμβολιαστεί πλήρως): Πρόκειται για τον αριθμό των ατόμων που έλαβαν ολόκληρο το σύνολο των εμβολιασμών σύμφωνα με τον σχεδιασμό εμβολιασμού (συνήθως 2)- σε μια συγκεκριμένη χρονική στιγμή, μπορεί να υπάρχει ένας συγκεκριμένος αριθμός ατόμων που έλαβαν ένα εμβόλιο και ένας άλλος αριθμός (μικρότερος) ατόμων που έλαβαν όλα τα εμβόλια του σχεδιασμού
- daily_vaccinations (Ημερήσιοι εμβολιασμοί): Για μια συγκεκριμένη καταχώρηση δεδομένων, ο αριθμός των εμβολιασμών για τη συγκεκριμένη ημερομηνία/χώρα
- total_vaccinations_per_hundred (Συνολικοί εμβολιασμοί ανά εκατό): Αναλογία (σε ποσοστό) μεταξύ του αριθμού των εμβολιασμών και του συνολικού πληθυσμού μέχρι τη συγκεκριμένη ημερομηνία στη χώρα
- people_vaccinated_per_hundred (Συνολικός αριθμός εμβολιασθέντων ανά εκατό): Αναλογία (σε ποσοστό) μεταξύ του πληθυσμού που εμβολιάστηκε και του συνολικού πληθυσμού μέχρι την ημερομηνία στη χώρα
- people__fully_vaccinated_per_hundred (Συνολικός αριθμός πλήρως εμβολιασμένων ατόμων ανά εκατό: Αναλογία (σε ποσοστό) μεταξύ του πλήρως εμβολιασμένου πληθυσμού και του συνολικού πληθυσμού μέχρι την ημερομηνία στη χώρα
- daily_vaccinations_per_million (Ημερήσιοι εμβολιασμοί ανά εκατομμύριο): Αναλογία μεταξύ του αριθμού εμβολιασμών και του συνολικού πληθυσμού για την τρέχουσα ημερομηνία στη χώρα
- vaccines (Εμβόλια): Εμβόλια που χρησιμοποιήθηκαν στη χώρα
- source_name (Όνομα πηγής): Πηγή των πληροφοριών (εθνική αρχή, διεθνής οργανισμός, τοπικός οργανισμός κ.λπ.),
- source_website (Δικτυακός τόπος της πηγής): Δικτυακός τόπος της πηγής των πληροφοριών

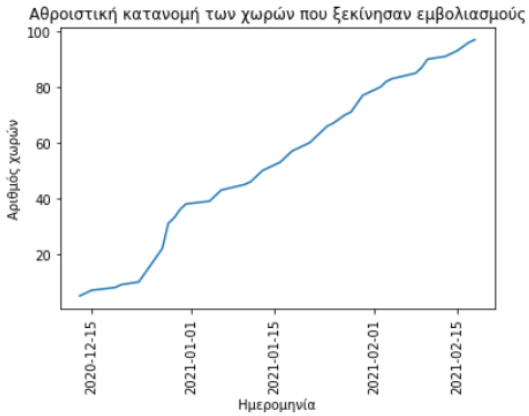
Το σύνολο δεδομένων που χρησιμοποιήθηκε περιείχε 3555 εγγραφές (γραμμές) και 15 μεταβλητές (στήλες). Ακόμη, τα δεδομένα μας αφορούν τους εμβολιασμούς που έγιναν σε κάθε χώρα μεταξύ των ημερομηνιών 13-12-2020 και 19-02-2021. Για την διευκόλυνση στους υπολογισμούς μας, αφαιρέθηκαν οι στήλες source_name και source_website. Ακόμα, για την αντιμετώπιση των ελλειπών τιμών στην μεταβλητή total_vaccinations, αφαιρέσαμε τις εγγραφές στις οποίες υπήρχαν ελλειπείς τιμές. Συνεπώς, εκτελώντας τις παραπάνω διαδικασίες, το τελικό σύνολο δεδομένων περιείχε 2341 εγγραφές και 13 μεταβλητές.

4.3 Περιγραφική Ανάλυση

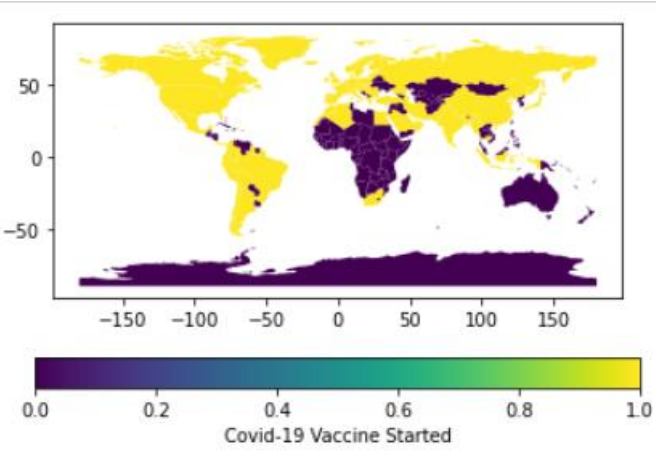
Αρχικά, εκτελέσαμε περιγραφική ανάλυση του συνόλου των δεδομένων μας με σκοπό να εξάγουμε ορισμένα χρήσιμα συμπεράσματα που θα μας βοηθήσουν στην καλύτερη κατανόηση των δεδομένων μας αλλά και μετέπειτα στην εκτέλεση των προβλέψεων μας.

Από την περιγραφική ανάλυση που εκτελέσαμε, παρατηρήσαμε ότι 97 χώρες ξεκίνησαν, σε εκείνη την περίοδο που αναφέρονται τα δεδομένα, να χορηγούν εμβόλια στους πολίτες και μάλιστα οι τέσσερις πρώτες χώρες που χορήγησαν εμβόλια στις 13-12-2020 ήταν η Αγγλία (55437 εμβόλια), η Σκωτία (18993 εμβόλια), η Ουαλία (8212 εμβόλια) και η Βόρεια Ιρλανδία (3623 εμβόλια). Ακόμη, στο παρακάτω γράφημα (Σχήμα 4.1), μπορούμε να δούμε την συνεχή άυξηση των χωρών που χορηγούν εμβόλια με την πάροδο των ημερών και στο Σχήμα 4.2 έχουμε κατασκευάσει έναν γεωγραφικό χάρτη με τις χώρες που ξεκίνησαν τον εμβολιασμό.

Σχήμα 4.1
Αθροιστική κατανομή των χωρών που ξεκίνησαν εμβολιασμούς



Σχήμα 4.2
Γεωγραφικός χάρτης των χωρών που ξεκίνησαν τον εμβολιασμό

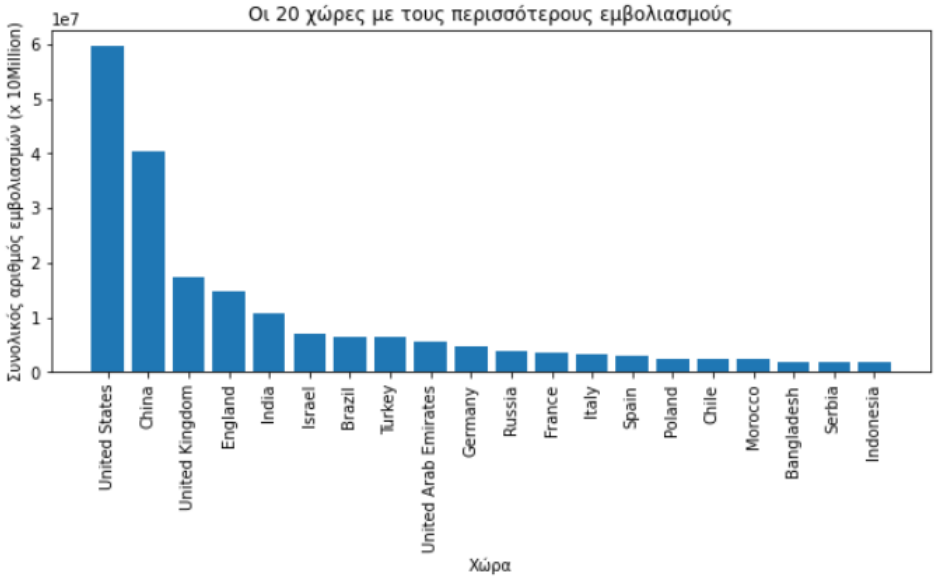


Ακόμη, παρατηρήσαμε ότι ο συνολικός αριθμός των εμβολιασμών που έγιναν σε όλες τις χώρες έως 19-02-2021 ήταν 218.115.733 εμβόλια, ενώ η χώρα με τους περισσότερους

εμβολιασμούς μέχρι εκείνη την ημέρα ήταν οι Η.Π.Α. με 59.585.043 εμβόλια. Μάλιστα, παρακάτω μπορούμε να δούμε ένα γράφημα (Σχήμα 4.3) με τις 20 πρώτες χώρες με τους περισσότερους εμβολιασμούς. Είναι εμφανές ότι οι Η.Π.Α. έχουν τους περισσότερους εμβολιασμούς, στη συνέχεια ακολουθεί η Κίνα, ενώ οι υπόλοιπες χώρες όπως η Αγγλία, η Ινδία, το Ισραήλ κ.τ.λ φαίνεται να έχουν πολύ λιγότερους εμβολιασμούς σε σχέση με τις δύο προαναφερθέντες χώρες. Επιπλέον, στον παρακάτω γεωγραφικό χάρτη (Σχήμα 4.4) παρατηρούμε τις χώρες που ξεκίνησαν τον εμβολιασμό για covid-19, σε σχέση με τον συνολικό αριθμό των εμβολιασμών που πραγματοποιήθηκαν.

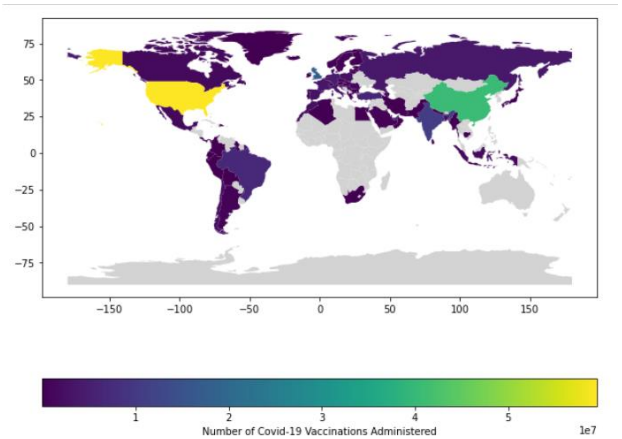
Σχήμα 4.3

Οι 20 πρώτες χώρες με τους περισσότερους εμβολιασμούς



Σχήμα 4.4

Γεωγραφικός χάρτης των 20 πρώτων χωρών που ξεκίνησαν τον εμβολιασμό για Covid-19

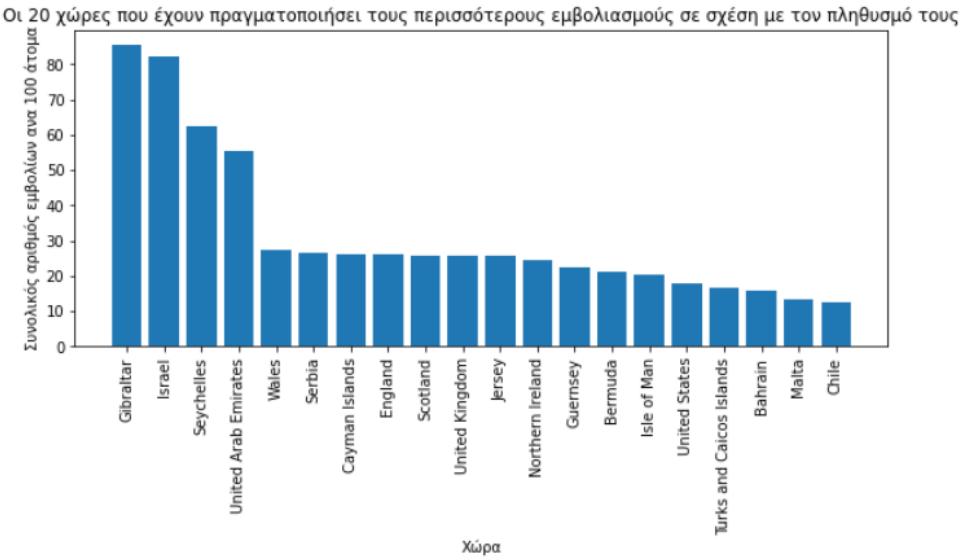


Παραπάνω είδαμε ότι οι Η.Π.Α. έχουν χορηγήσει τα περισσότερα εμβόλια μέχρι στιγμής, ωστόσο ο πληθυσμός τους είναι μεγάλος με συνέπεια να χρειαστεί λίγος χρόνος μέχρι να εμβολιαστεί το μεγαλύτερο μέρος του πληθυσμού. Για να μελετήσουμε τους εμβολιασμούς από μια διαφορετική οπτική γωνία, αναρωτιόμαστε επίσης ποιες χώρες έχουν εμβολιάσει το μεγαλύτερο μέρος του πληθυσμού τους. Έτσι, θα μελετήσουμε ποιες χώρες έχουν χορηγήσει τον μεγαλύτερο αριθμό εμβολίων ανά κάτοικο. Να σημειώσουμε ότι, δεδομένου ότι κάθε άτομο απαιτεί κατά μέσο όρο 2 δόσεις, αναμένουμε ότι μια πλήρως εμβολιασμένη χώρα θα έχει 200 εμβολιασμούς ανά 100 κατοίκους.

Τα αποτελέσματα της μελέτης έδειξαν ότι το Γιβραλτάρ ήταν η χώρα η οποία εμβολίασε το μεγαλύτερο μέρος του πληθυσμού της, αφού χορήγησε περίπου 86 εμβόλια ανά 100 κατοίκους. Στο παρακάτω γράφημα (Σχήμα 4.5) βλέπουμε τις 20 πρώτες χώρες που εμβολίασαν το μεγαλύτερο μέρος του πληθυσμού τους, όπου το Γιβραλτάρ είναι πρώτο, ακολουθεί το Ισραήλ με ελάχιστη διαφορά, ενώ λίγο παρακάτω βρίσκονται οι Σεϋχέλλες και τα Ηνωμένα Αραβικά Εμιράτα. Τέλος, παρακάτω βλέπουμε και έναν γεωγραφικό χάρτη (Σχήμα 4.6) που κατασκευάστηκε για το ζητούμενο της παραπάνω μελέτης.

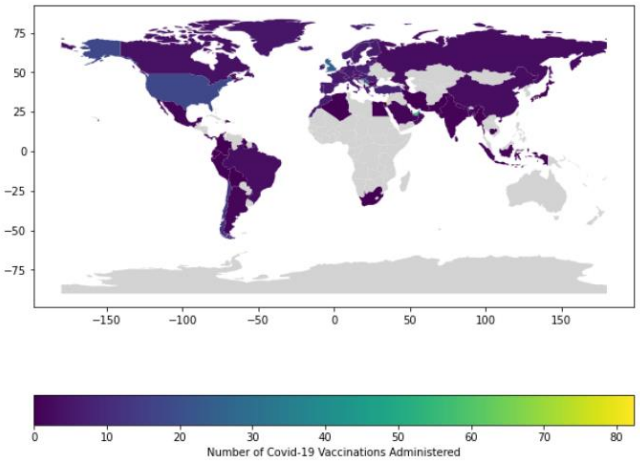
Σχήμα 4.5

Οι 20 πρώτες χώρες που έχουν πραγματοποιήσει τους περισσότερους εμβολιασμούς σε σχέση με τον πληθυσμό τους



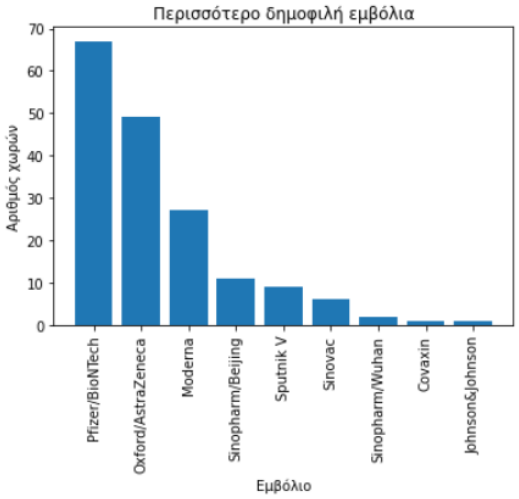
Σχήμα 4.6

Γεωγραφικός χάρτης με τις 20 πρώτες χώρες που εμβολίασαν το μεγαλύτερο μέρος του πληθυσμού τους

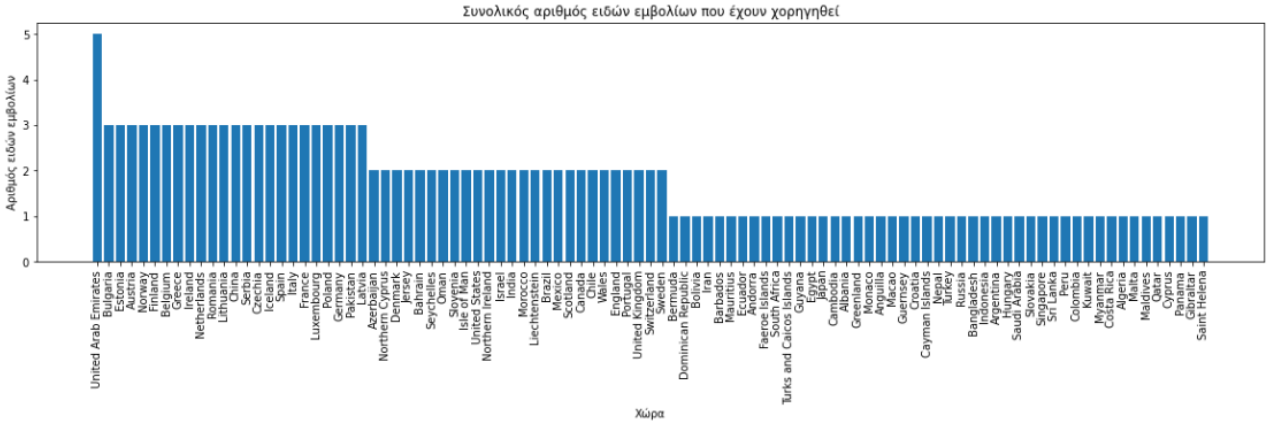


Ένα ακόμη συμπέρασμα που έχουμε εξάγει από την παρούσα μελέτη είναι ότι το εμβόλιο COVID-19 των Pfizer – BioNTech ήταν το πιο δημοφιλές εμβόλιο ανάμεσα σε όλα τα άλλα που κατασκευάστηκαν, αφού χορηγήθηκε σε 67 χώρες. Μάλιστα, φαίνεται να είναι λογικό αφού το συγκεκριμένο εμβόλιο ήταν το πρώτο εμβόλιο COVID-19 που εγκρίθηκε από μία αυστηρή ρυθμιστική αρχή για χρήση έκτακτης ανάγκης και το πρώτο που εγκρίθηκε για τακτική χρήση. Επιπλέον, στο παρακάτω γράφημα (Σχήμα 4.7) παρατηρούμε την εξέλιξη όλων των εμβολίων, που κατασκευάστηκαν για την πανδημία COVID-19, όσον αφορά την δημοτικότητα τους σε όλο τον κόσμο. Φαίνεται, λοιπόν, ότι το εμβόλιο των Pfizer – BioNTech να είναι το πιο δημοφιλή με μικρή διαφορά από το εμβόλιο των Oxford – AstraZeneca, ενώ όλα τα υπόλοιπα εμβόλια φαίνεται να είναι λιγότερο δημοφιλή στον κόσμο. Είναι προφανές ότι ο αριθμός των χωρών που χρησιμοποιούν κάθε εμβόλιο δεν αντιστοιχεί στον αριθμό των χωρών που έχουν ξεκινήσει εμβολιασμούς. Αυτό οφείλεται στο γεγονός ότι ορισμένες χώρες χρησιμοποιούν περισσότερους από έναν τύπους εμβολίων. Έτσι, από το Σχήμα 4.8 παρατηρούμε πόσα είδη εμβολίων (διαφορετικές εταιρείες που κατασκευάστηκαν) έχουν χορηγηθεί σε κάθε χώρα, όπου φαίνεται ότι στα Ηνωμένα Αραβικά Εμιράτα να έχουν χορηγηθεί 5 είδη εμβολίων που είναι και τα περισσότερα σε σχέση με τις υπόλοιπες χώρες, οι οποίες φαίνεται να χρησιμοποιούν από 1 έως 3 είδη εμβολίων.

Σχήμα 4.7
Δημοφιλή εμβόλια για τον COVID-19

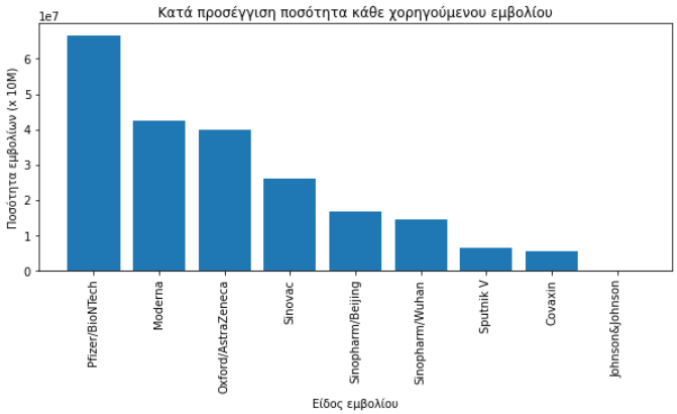


Σχήμα 4.8
Αριθμός ειδών εμβολίων που έχουν χορηγηθεί σε κάθε χώρα



Τέλος, θέλαμε να δούμε από την άποψη των πωλήσεων των εμβολίων και του απόλυτου αριθμού των εμβολίων που χορηγούνται, ποιες μάρκες είναι οι πιο δημοφιλείς. Να σημειώσουμε ότι για τις χώρες που αγοράζουν περισσότερους από έναν τύπους εμβολίων, οι συνολικές αγορές υποτίθεται ότι κατανέμονται ισομερώς μεταξύ των τύπων εμβολίων που χρησιμοποιούνται. Έτσι, από το παρακάτω γράφημα (Σχήμα 4.9) βλέπουμε ότι οι περισσότερες πωλήσεις εμβολίων ανήκουν και πάλι στο εμβόλιο των Pfizer – BioNTech. Ενώ πιο κάτω με μεγάλη διαφορά από το παραπάνω εμβόλιο βρίσκονται τα εμβόλια Moderna και Oxford – AstraZeneca.

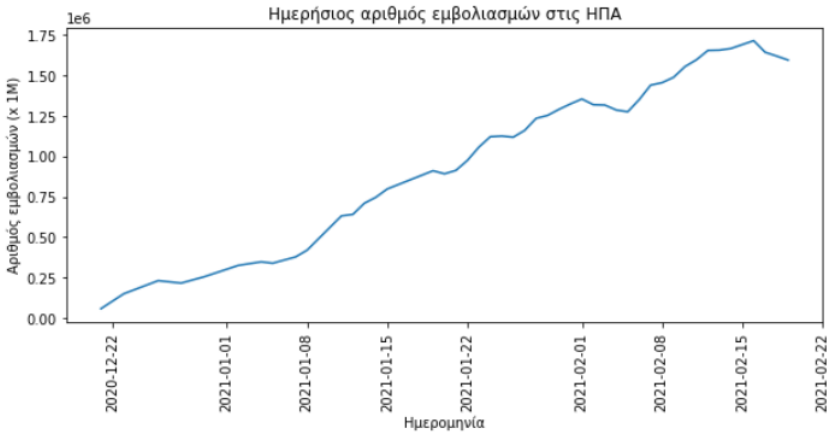
Σχήμα 4.9
Ποσότητα κάθε χορηγούμενου εμβολίου



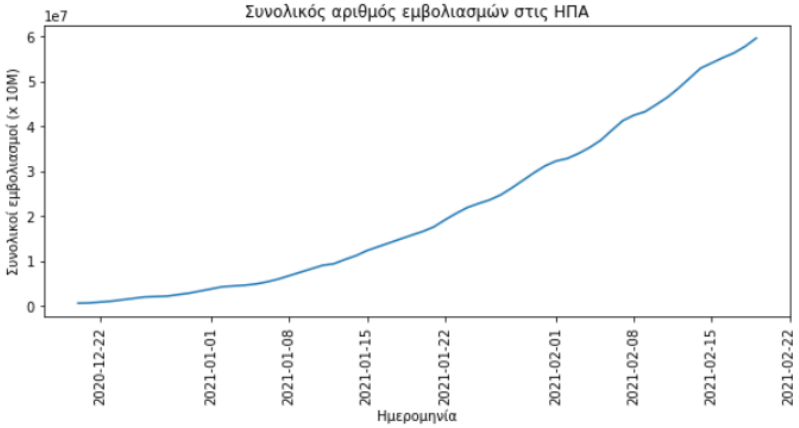
4.4 Πρόβλεψη για την επίτευξη ανοσίας της αγέλης

Στην παρούσα ενότητα θα επικεντρωθούμε στις Ηνωμένες Πολιτείες της Αμερικής, καθώς διαθέτουν τα πληρέστερα στοιχεία στο σύνολο δεδομένων. Ωστόσο, το μοντέλο που θα κατασκευάσουμε μπορεί να επαναληφθεί για οποιαδήποτε άλλη χώρα. Σε αυτή την ενότητα, ο κύριος στόχος μας είναι να προβλέψουμε πότε μπορούμε να περιμένουμε ότι οι Η.Π.Α. θα χορηγήσουν αρκετά εμβόλια για να επιτύχουν ανοσία της αγέλης. Αρχικά, με το παρακάτω γράφημα (Σχήμα 4.10) παρατηρούμε τον ημερήσιο αριθμό εμβολιασμών στις Η.Π.Α., όπου είναι εμφανές πως με την πάροδο των ημερών, ο αριθμός των εμβολιασμών αυξανόταν ραγδαία. Επίσης, από την κατασκευή ενός ακόμη γραφήματος (Σχήμα 4.11) σχετικά με τους αθροιστικούς εμβολιασμούς στις Η.Π.Α., φαίνεται να υπάρχει μία εκθετική αύξηση των εμβολιασμών.

Σχήμα 4.10
Ημερήσιος αριθμός εμβολιασμών στις Η.Π.Α.



Σχήμα 4.11
Συνολικός αριθμός εμβολιασμών στις Η.Π.Α.



Στη συνέχεια για να προχωρήσουμε στην κατασκευή του μοντέλου, για λόγους διευκόλυνσης στους υπολογισμούς δημιουργήσαμε ένα καινούργιο σύνολο δεδομένων το οποίο περιέχει τις εγγραφές που αναφέρονται στις Η.Π.Α. και τις στήλες «date» και «total_vaccinations». Το συγκεκριμένο σύνολο δεδομένων περιέχει 48 εγγραφές και 2 μεταβλητές. Έπειτα για να ξεκινήσουμε με την προετοιμασία των δεδομένων για αυτό το πρόβλημα, μετατρέψαμε όλες τις ημερομηνίες στην τιμή "t", που αντιπροσωπεύει τις ημέρες από την έναρξη των εμβολιασμών στις Η.Π.Α. και χωρίσαμε το σύνολο των δεδομένων μας σε training set και test set με το training set να περιέχει 38 εγγραφές και το test set να περιέχει 10 εγγραφές.

Στην προσπάθεια μας να βρούμε το καταλληλότερο μοντέλο κατασκευάσαμε και εκπαιδεύσαμε διάφορα μοντέλα για την προσαρμογή των δεδομένων. Για κάθε μοντέλο καταγράψαμε τις τιμές RMSE (Root-Mean-Square Error). Τα μοντέλα που εκπαιδεύτηκαν είναι τα εξής:

- Linear Model (Γραμμικό μοντέλο Παλινδρόμησης)
- Exponential
- Quadratic
- Log
- Sqrt

Στα μοντέλα που κατασκευάστηκαν η μεταβλητή απόκρισης ήταν ο Συνολικός αριθμός εμβολιασμών (total_vaccinations) και η επεξηγηματική μεταβλητή ήταν η μεταβλητή «t» που ορίσαμε παραπάνω. Στον παρακάτω πίνακα (Πίνακας 4.1) φαίνονται οι τιμές του RMSE, που όσο πιο μικρό είναι αυτό σημαίνει ότι το μοντέλο μας προσαρμόζεται καλύτερα στα δεδομένα μας, για κάθε μοντέλο. Με βάση τις τιμές RMSE, το καλύτερο μοντέλο για την πρόβλεψη των συνολικών εμβολιασμών στις Η.Π.Α. είναι το τετραγωνικό μοντέλο (quadratic model). Οι πληροφορίες αυτού του μοντέλου που κατασκευάστηκε συνοψίζονται στον παρακάτω πίνακα (Πίνακας 4.2).

Πίνακας 4.1
RMSE scores

	Model	RMSE_Values
0	rmse_linear	1.028151e+07
1	rmse_exponential	5.372120e+07
2	rmse_quadratic	1.699031e+06
3	rmse_log	2.380259e+07
4	rmse_sqrt	1.686654e+07

Πίνακας 4.2
Πληροφορίες του quadratic model

OLS Regression Results

Dep. Variable:	total_vaccinations	**R-squared:**	0.999
Model:	OLS	**Adj. R-squared:**	0.999
Method:	Least Squares	**F-statistic:**	1.773e+04
Date:	Sat, 07 Jan 2023	**Prob (F-statistic):**	2.47e-53
Time:	22:42:38	**Log-Likelihood:**	-544.82
No. Observations:	38	**AIC:**	1096.
Df Residuals:	35	**BIC:**	1101.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	7.841e+05	2.36e+05	3.317	0.002	3.04e+05	1.26e+06
t	-2.364e+04	1.95e+04	-1.215	0.233	-6.32e+04	1.59e+04
t_squared	1.646e+04	350.449	46.973	0.000	1.58e+04	1.72e+04

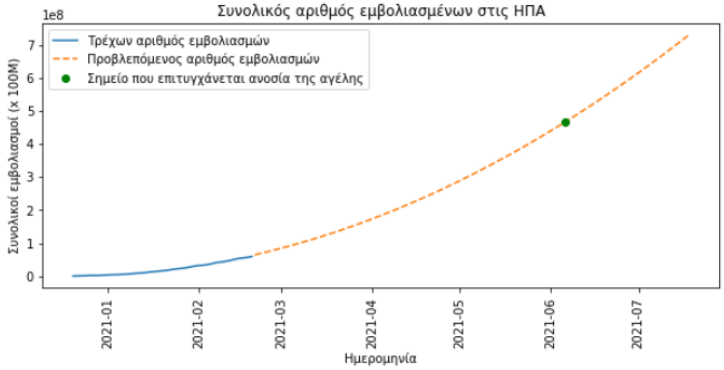
Omnibus:	0.497	**Durbin-Watson:**	0.686
Prob(Omnibus):	0.780	**Jarque-Bera (JB):**	0.629
Skew:	0.213	**Prob(JB):**	0.730
Kurtosis:	2.536	**Cond. No.**	4.66e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.66e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

Από τον παραπάνω πίνακα (Πίνακας 4.2), η τιμή του R - τετραγώνου (R-squared) για το μοντέλο είναι πολύ υψηλή (0.999), κάτι που δείχνει ότι είναι καλή προσαρμογή του μοντέλου που κατασκευάσαμε στα δεδομένα. Εκτός αυτού, η τιμή p-value του t_squared είναι μικρότερη του 0.05, δηλαδή είναι στατιστικά σημαντική η μεταβλητή για την πρόβλεψη του συνολικού αριθμού των εμβολιασμών. Επομένως, χρησιμοποιώντας αυτό το μοντέλο, προσπαθήσαμε να προβλέψουμε πότε οι Η.Π.Α. θα είναι σε θέση να επιτύχουν ανοσία αγέλης. Ο σημερινός πληθυσμός των Η.Π.Α. είναι 332 εκατομμύρια άνθρωποι (U.S. Department of Commerce, 2022). Άρα, για να επιτευχθεί ποσοστό εμβολιασμού 70% για ανοσία της αγέλης, θα πρέπει να εμβολιαστούν 70% * 332 εκατ. = 232 εκατ. του πληθυσμού. Δεδομένου ότι κάθε άτομο απαιτεί

2 δόσεις, ο συνολικός αριθμός εμβολιασμών θα πρέπει να είναι μεγαλύτερος 464 εκατομμύρια άτομα. Χρησιμοποιώντας το test set προβλέψαμε ότι η ελάχιστη ημερομηνία κατά την οποία οι Η.Π.Α. θα ήταν σε θέση να επιτύχουν ανοσία της αγέλης ήταν στις 06-06-2021. Από το παρακάτω γράφημα (Σχήμα 4.12) που κατασκευάσαμε για τον συνολικό αριθμό εμβολιασμένων στις Η.Π.Α. φαίνεται με την μπλε γραμμή ο τρέχων αριθμός εμβολιασμένων, με την πορτοκαλί διακεκομμένη γραμμή ο προβλεπόμενος αριθμός εμβολιασμένων, ενώ με την πράσινη τελεία δείχνουμε το σημείο στο οποίο θα επιτευχθεί ανοσία της αγέλης στις Η.Π.Α.

Σχήμα 4.12
Συνολικός αριθμός εμβολιασμένων στις Η.Π.Α



Επειδή τα δεδομένα μας βρίσκονται σε χρονική σειρά, παραβιάζεται η υπόθεση της ανεξαρτησίας. Όμως, τα νευρωνικά δίκτυα αν και είναι πολύπλοκα, είναι ευέλικτα και μπορούν να επιλέγουν τον καλύτερο τύπο παλινδρόμησης (How Neural Networks are used for Regression in R Programming (2020)). Συνεπώς, για να προβλέψουμε το πότε σε μία χώρα με την χορήγηση των εμβολίων στους πολίτες της θα επιτευχθεί η ανοσία της αγέλης, καυασκευάσαμε ένα νευρωνικό δίκτυο (neural network) και συγκεκριμένα ένα Feed-Forward neural network.

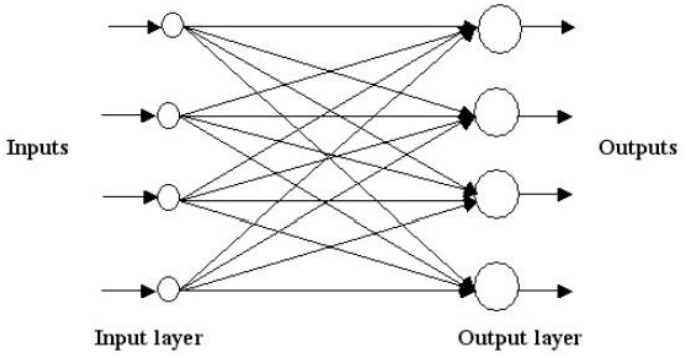
4.5 Feed-Forward neural networks

Ο Sazli (2006) αναφέρει πως τα τεχνητά νευρωνικά δίκτυα έχουν εμπνευστεί το ονομά τους από τον βιολογικό εγκέφαλο και το νευρικό σύστημα. Ο βιολογικός εγκέφαλος είναι εντελώς διαφορετικός από τον συμβατικό ψηφιακό υπολογιστή όσον αφορά το δομή και τον τρόπο με τον οποίο επεξεργάζεται τις πληροφορίες. Με πολλούς τρόπους, ο βιολογικός εγκέφαλος είναι πολύ πιο προηγμένος και ανώτερος από τον από τους συμβατικούς υπολογιστές. Το σημαντικότερο διακριτικό χαρακτηριστικό ενός βιολογικού εγκεφάλου είναι η ικανότητά του να "μαθαίνει" και να "προσαρμόζεται", ενώ ένας συμβατικός υπολογιστής δεν

έχει τέτοιες ικανότητες. Οι υπολογιστές επιτελούν συγκεκριμένες εργασίες με βάση τις οδηγίες που τους έχουν φορτωθεί, τα λεγόμενα "προγράμματα" ή "λογισμικό". Βασικό δομικό στοιχείο των νευρωνικών δικτύων είναι ο "νευρώνας". Ένας νευρώνας μπορεί να είναι αντιληπτός ως μονάδα επεξεργασίας. Σε ένα νευρωνικό δίκτυο, οι νευρώνες συνδέονται μεταξύ τους μεταξύ τους μέσω "βαρών". Κάθε νευρώνας σε ένα δίκτυο λαμβάνει "σταθμισμένες" πληροφορίες μέσω αυτών των συνδέσεων από τους νευρώνες με τους οποίους είναι συνδεδεμένος και παράγει μια έξοδο περνώντας το σταθμισμένο άθροισμα αυτών των σημάτων εισόδου (είτε εξωτερικές εισόδους από το περιβάλλον είτε τις εξόδους άλλων νευρώνων) μέσω μιας "συνάρτησης ενεργοποίησης" (activation function). Υπάρχουν δύο κύριες κατηγορίες νευρωνικών δικτύων ανάλογα με τον τύπο των συνδέσεων μεταξύ των νευρώνων, τα "νευρωνικά δίκτυα τροφοδότησης προς τα εμπρός" (feed-forward neural networks) και τα "επαναλαμβανόμενα νευρωνικά δίκτυα" (recurrent neural networks). Εάν δεν υπάρχει "ανατροφοδότηση" από τις εξόδους των νευρώνων προς τις εισόδους σε όλο το δίκτυο, τότε το δίκτυο αναφέρεται ως "feed-forward neural network". Διαφορετικά, εάν υπάρχει μια τέτοια ανατροφοδότηση, δηλαδή μία σύνδεση από τις εξόδους προς τις εισόδους (είτε τις δικές τους εισόδους είτε τις εισόδους άλλων νευρώνων), τότε το δίκτυο καλείται "recurrent neural networks". Συνήθως, τα νευρωνικά δίκτυα διατάσσονται με τη μορφή "στρωμάτων" (layers). Τα feed-forward neural networks διακρίνονται σε δύο κατηγορίες ανάλογα με τον αριθμό των στρωμάτων, είτε "ένα στρώμα" (single layer) ή "πολλαπλά στρώματα" (multi-layer).

Στο Σχήμα 4.13 (Sazli, 2006) παρουσιάζεται ένα single layer feed-forward neural network. Συμπεριλαμβανομένου του στρώματος εισόδου (input layer), υπάρχουν δύο στρώματα σε αυτή τη δομή. Ωστόσο, το input layer δεν προσμετράται επειδή δεν εκτελείται κανένας υπολογισμός σε αυτό το στρώμα. Τα σήματα εισόδου (input signals) διαβιβάζονται στο στρώμα εξόδου μέσω των βαρών και των νευρώνων στο στρώμα εξόδου (output layer) ώστε να υπολογιστούν τα σήματα εξόδου (output signals).

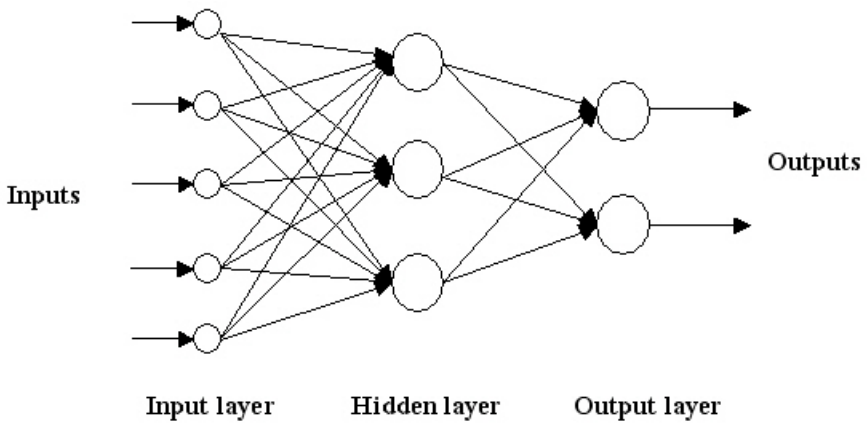
Σχήμα 4.13
Single layer feed-forward neural network



Στο Σχήμα 4.14 (Sazli, 2006) παρουσιάζεται ένα feed-forward neural network πολλαπλών στρωμάτων (multi-layer feed-forward neural network) με ένα "κρυφό στρώμα" (hidden layer). Σε αντίθεση με ένα νευρωνικό δίκτυο ενός στρώματος, υπάρχει (τουλάχιστον) ένα στρώμα "κρυφών νευρώνων" μεταξύ των στρωμάτων εισόδου και εξόδου. Σύμφωνα με τον Haykin (1999), η λειτουργία των κρυφών νευρώνων (hidden neurons) είναι να παρεμβαίνουν

μεταξύ της εξωτερικής εισόδου και του εξόδου του νευρωνικού δικτύου με κάποιο χρήσιμο τρόπο. Η ύπαρξη ενός ή περισσότερων κρυφών στρωμάτων (hidden layers) επιτρέπει στο δίκτυο να εξάγει καλύτερα στατιστικά αποτελέσματα. Στο Σχήμα 4.14 (Sazli, 2006), υπάρχει μόνο ένα κρυφό στρώμα και το δίκτυο αναφέρεται ως νευρωνικό δίκτυο 5-3-2. επειδή υπάρχουν 5 νευρώνες εισόδου, 3 κρυφοί νευρώνες και 2 νευρώνες εξόδου. Τόσο στο Σχήμα 4.13 όσο και στο Σχήμα 4.14, τα δίκτυα είναι "πλήρως συνδεδεμένα"(fully connected) επειδή κάθε νευρώνας σε κάθε στρώμα συνδέεται με κάθε άλλο νευρώνα στο επόμενο στρώμα. Εάν κάποιες από τις συνδέσεις έλειπαν, το δίκτυο θα ονομαζόταν "μερικώς συνδεδεμένο" (partially connected). Το σημαντικότερο χαρακτηριστικό ενός νευρωνικού δικτύου που το διακρίνει είναι η ικανότητα "μάθησης" του. Ένα νευρωνικό δίκτυο μπορεί να μάθει από το περιβάλλον του και να βελτιώνει την απόδοσή του μέσω της μάθησης (Sazli, 2006).

Σχήμα 4.14
Multi-layer feed-forward neural network



4.6 Εφαρμογή του Feed-Forward neural network

Θέλαμε να κατασκευάσουμε Feed-Forward neural network στα δεδομένα που είχαμε στη διάθεση μας, καθώς όπως προείπαμε είναι πιο ευέλικτο στην υπόθεση της ανεξαρτησίας. Έτσι, κατασκευάσαμε αυτό το νευρωνικό δίκτυο το οποίο αποτελούνταν από 1 νευρώνα στο input layer, 25 νευρώνες στο hidden layer και 1 νευρώνα στο output layer. Το input layer είναι η μεταβλητή «t» και το output layer είναι η μεταβλητή «total vaccinations». Πρόκειται για ένα fully connected δίκτυο και αναφέρεται ως 1-25-1. Χρησιμοποιήθηκε για το μοντέλο η συνάρτηση «Sequential», καθώς το μοντέλο αποτελείται από διαδοχικά στρώματα και ως activation function χρησιμοποιήθηκε η relu (rectified linear unit) η οποία είναι μια συνάρτηση που εισάγει την ιδιότητα της μη γραμμικότητας σε ένα μοντέλο βαθιάς μάθησης και είναι μία από τις πιο δημοφιλείς συναρτήσεις στη βαθιά μάθηση (Krishnamurthy, 2022). Κατασκευάζοντας το Feed-Forward neural network μοντέλο στα δεδομένα μας, παρατηρήσαμε πως η τιμή R^2 ήταν πολύ μικρή και συγκεκριμένα ήταν περίπου ίση με $3.52*10^{-6}$. Ένας λόγος

για τον οποίο δεν είχαμε καλό αποτελέσμα σχετίζεται με το ότι δεν είχαμε πολλά δεδομένα στην διάθεση μας για τις Η.Π.Α.

4.7 Συμπεράσματα

Στην παρούσα μελέτη που διενεργήθηκε θέλαμε να αναδείξουμε την σημασία και τη συνεισφορά της Στατιστικής Μηχανικής Μάθησης σε θέματα που αφορούν την Δημόσια Υγεία και μάλιστα σε ένα τόσο μείζον θέμα όπως η πανδημία του κορονοϊού (COVID-19) που προκάλεσε πολλούς θανάτους αλλά και εκτεταμένη αναστάτωση στην οικονομία και στις κοινωνικές δραστηριότητες. Στόχος αυτής της μελέτης ήταν με την βοήθεια στατιστικών μοντέλων να μπορέσουμε να προβλέψουμε με τα δεδομένα που είχαμε στην διάθεσή μας το πότε σε μία χώρα με την χορήγηση των εμβολίων στους πολίτες της θα επιτευχθεί η ανοσία της αγέλης. Είδαμε πως για τις Η.Π.Α. χρησιμοποιώντας το τετραγωνικό μοντέλο παλινδρόμησης (quadratic regression model), το οποίο προσαρμοζόταν καλύτερα στα δεδομένα μας σε σχέση με τα υπόλοιπα μοντέλα που εκπαιδεύσαμε, παρατηρήσαμε πως η ελάχιστη ημερομηνία κατά την οποία οι Η.Π.Α. θα ήταν σε θέση να επιτύχουν ανοσία της αγέλης ήταν στις 06-06-2021. Όμως, τα δεδομένα μας βρισκόντουσαν σε χρονική σειρά κάτι το οποίο παραβιάζει την υπόθεση της ανεξαρτησίας και με τα μοντέλα που χρησιμοποιήσαμε δεν θα είχαμε ασφαλείς εκετιμήσεις. Συνεπώς, κατασκευάσαμε ένα Feed-Forward neural network στο οποίο η τιμή R^2 ήταν πολύ μικρή, διότι δεν είχαμε πολλά δεδομένα στη διάθεση μας.

Επιπλέον, από την περιγραφική ανάλυση που διενεργήσαμε είχαμε την δυνατότητα να εξάγουμε ορισμένα χρήσιμα συμπεράσματα για τη εξέλιξη του εμβολιασμού σε ολόκληρο τον κόσμο. Πιο συέβαιαγκεκριμένα, είδαμε ότι 97 χώρες ξεκίνησαν, σε εκείνη την περίοδο που αναφέρονται τα δεδομένα, να χορηγούν εμβόλια στους πολίτες και μάλιστα οι τέσσερις πρώτες χώρες που χορήγησαν εμβόλια στις 13-12-2020 ήταν η Αγγλία (55437 εμβόλια), η Σκωτία (18993 εμβόλια), η Ουαλία (8212 εμβόλια) και η Βόρεια Ιρλανδία (3623 εμβόλια) και με την βοήθεια ενός γραφήματος είδαμε την συνεχή άυξηση των χωρών που χορηγούσαν εμβόλια με την πάροδο των ημερών. Ακόμη, παρατηρήσαμε ότι ο συνολικός αριθμός των εμβολιασμών που έγιναν σε όλες τις χώρες έως 19-02-2021 ήταν 218.115.733 εμβόλια, ενώ η χώρα με τους περισσότερους εμβολιασμούς μέχρι εκείνη την ημέρα ήταν οι Η.Π.Α με 59.585.043 εμβόλια. Μάλιστα κάποιες από τις 20 πρώτες χώρες που χορήγησαν εμβόλια ήταν οι Η.Π.Α η Κίνα, η Αγγλία, η Ινδία, το Ισραήλ κ.τ.λ. Στη συνέχεια, αναρωτηθήκαμε ποιες χώρες έχουν εμβολιάσει το μεγαλύτερο μέρος του πληθυσμού τους. Έτσι, μελετήσαμε ποιες χώρες έχουν χορηγήσει τον μεγαλύτερο αριθμό εμβολίων ανά κάτοικο. Τα αποτελέσματα της μελέτης έδειξαν ότι το Γιβραλτάρ ήταν η χώρα η οποία εμβολίασε το μεγαλύτερο μέρος του πληθυσμού της, αφού χορήγησε περίπου 86 εμβόλια ανά 100 κατοίκους ενώ κάποιες ακόμα χώρες το Ισραήλ, οι Σεϋχέλλες, τα Ηνωμένα Αραβικά Εμιράτα κ.τ.λ.

Ένα ακόμη συμπέρασμα που έχουμε εξάγει από την παρούσα μελέτη ήταν ότι το εμβόλιο COVID-19 των Pfizer – BioNTech ήταν το πιο δημοφιλές εμβόλιο ανάμεσα σε όλα τα άλλα που κατασκευάστηκαν, αφού χορηγήθηκε σε 67 χώρες, ενώ με μικρή διαφορά ήταν το εμβόλιο των Oxford – AstraZeneca, ενώ όλα τα υπόλοιπα εμβόλια φάνηκε να είναι λιγότερο δημοφιλή

στον κόσμο. Βέβαια είναι προφανές ότι ο αριθμός των χωρών που χρησιμοποιούν κάθε εμβόλιο δεν αντιστοιχεί στον αριθμό των χωρών που έχουν ξεκινήσει εμβολιασμούς. Αυτό οφείλεται στο γεγονός ότι ορισμένες χώρες χρησιμοποιούν περισσότερους από έναν τύπους εμβολίων. Έτσι, αναζητήσαμε να ελέγξουμε πόσα είδη εμβολίων (διαφορετικές εταιρείες που κατασκευάστηκαν) έχουν χορηγηθεί σε κάθε χώρα και συμπεράναμε ότι στα Ηνωμένα Αραβικά Εμιράτα είχαν χορηγηθεί 5 είδη εμβολίων που είναι και τα περισσότερα σε σχέση με τις υπόλοιπες χώρες, οι οποίες φαίνεται να χρησιμοποιούσαν από 1 έως 3 είδη εμβολίων.

Βέβαια, ένα μειονέκτημα αυτής της μελέτης ήταν ότι τα δεδομένα αναφερόντουσαν σε μία συγκεκριμένη περίοδο (13-12-2020 έως 19-02-2021), δεν είχαμε στην διάθεση μας περισσότερα δεδομένα ώστε να έχουμε πιο ασφαλείς εκτιμήσεις και στατιστικά συμπεράσματα, κάτι που φάνηκε από τα μοντέλα που κατασκευάσαμε για την πρόβλεψη της επίτευξης της ανοσίας της αγέλης, όπου τα εκπαιδεύσαμε με λίγα δεδομένα. Όμως, θεωρήσαμε σημαντικό να εξάγουμε κάποια σταστιστικά συμπεράσματα για αυτήν την χρονική περίοδο, διότι σε εκείνη την περίοδο η πανδημία βρισκόταν σε έξαρση και αυτό θα βοηθούσε να μπορέσουμε να παρατηρήσουμε την κοινή γνώμη των ατόμων σε όλα τα κράτη σχετικά με τα εμβόλια, που όπως είδαμε η χορήγηση των εμβολίων βρέθηκε να έχει μία αυξανόμενη άνοδο. Φυσικά είναι ενθαρρυντικό να γίνονται τέτοιου είδους μελέτες για ένα τόσο σημαντικό θέμα όπως αυτό, που όπως φαίνεται έχουν διεξαχθεί πολλές μελέτες και ελπίζουμε να διεξαχθούν ακόμα περισσότερες ώστε να υπάρχει η κατάλληλη πρόληψη και αντιμετώπιση αυτής της πανδημίας.

Παράρτημα

Πηγαίος κώδικας σε Python για την εφαρμογή

```
#Εισαγωγή βιβλιοθηκών
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
from matplotlib import pyplot as plt
import geopandas as gpd
import statsmodels.formula.api as smf
import datetime
import os
for dirname, _, filenames in os.walk('country_vaccinations.csv'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
#Εισαγωγή δεδομένων
data = pd.read_csv("country_vaccinations.csv")
data.shape
data.head()
# Αφαίρεση των στηλών source_name & source_website
data.drop(['source_name','source_website'], inplace=True, axis=1)
# Αφαίρεση των γραμμών όπου στην στήλη "total vaccinations" δεν υπάρχει τιμή(NaN values)
data = data.drop(data[data.total_vaccinations.isna()].index)
data.shape
data.head()
# Έυρεση των χωρών που ξεκίνησαν εμβολιασμούς
print(str(len(data[data.total_vaccinations > 0].country.value_counts())) + " countries have started vaccinations")
#Ευρεση της 1ης ημερομηνίας των δεδομένων μας
data['date'].min()
# Έυρεση της τελευταίας ημερομηνίας των δεδομένων μας
data['date'].max()
# Έυρεση των 5 πρώτων χωρών που ξεκίνησαν εμβολιασμούς στις 13/12/2020
data['date'] = pd.to_datetime(data['date'], utc=True)
b = data.loc[data[data.total_vaccinations > 0].groupby('country')['date'].idxmin()].sort_values('date')
b.head(5)
f = data[data.date == '2020-12-13 00:00:00+00:00']
# Δημιουργία νέου dataframe που περιέχει μόνο τις στήλες "date", "total_vaccinations" και "country"
data1 = f.loc[:,['total_vaccinations','date','country']]
```

```
data1
data1.groupby(by="country")["total_vaccinations"].nlargest(5)
# Αθροιστική κατανομή των χωρών που ξεκίνησαν εμβολιασμούς
c = pd.Series(b.date.value_counts())
c.index = pd.to_datetime(c.index)
c.sort_index(inplace=True)
plt.plot(c.cumsum())
plt.xticks(rotation=90)
plt.title('Αθροιστική κατανομή των χωρών που ξεκίνησαν εμβολιασμούς')
plt.xlabel('Ημερομηνία')
plt.ylabel('Αριθμός χωρών')
plt.show()
# Δημιουργία γραφήματος χάρτη με τις χώρες που ξεκίνησαν τον εμβολιασμό
countries = gpd.read_file(gpd.datasets.get_path("naturalearth_lowres"))
countries['score'] = 0
countries['score'] = np.where(countries.iso_a3.isin(b.iso_code), 1, 0)
countries.plot(column='score',legend=True, legend_kwds={'label': "Covid-19 Vaccine Started",'orientation': "horizontal"})
plt.show()
# Συνολικός αριθμός των εμβολιασμών που έχουν γίνει μέχρι 19/02/2021
total = data.loc[data.groupby('country')['total_vaccinations'].idxmax()].sort_values('total_vaccinations',ascending=False)
print("Ο συνολικός αριθμός των εμβολιασμών μέχρι 19/02/2021 είναι ", total.total_vaccinations.sum())
# Η χώρα με τους περισσότερους εμβολιασμούς μέχρι 19/02/2021
print(total.iloc[0].country, " έχει τους περισσότερους εμβλιασμούς μέχρι 19/02/2021 με αριθμό ", total.iloc[0].total_vaccinations)
# Γράφημα με τις 20 χώρες που έχουν τους περισσότερους εμβολιασμούς
plt.figure(figsize=(10, 4))
plt.bar(total.country[0:20], total.total_vaccinations[0:20])
plt.xticks(rotation=90)
plt.title('Οι 20 χώρες με τους περισσότερους εμβολιασμούς')
plt.xlabel('Χώρα')
plt.ylabel('Συνολικός αριθμός εμβολιασμών (x 10Million)')
plt.show()
# Γεωγραφικός χάρτης των χωρών που ξεκίνησαν τον εμβολιασμό για covid-19, σε σχέση με τον συνολικό αριθμό των εμβολιασμών που πραγματοποιήθηκαν
countries = gpd.read_file(gpd.datasets.get_path("naturalearth_lowres"))

countries = countries.merge(total[['iso_code', 'total_vaccinations']], how = 'left',
            left_on = 'iso_a3', right_on = 'iso_code').drop('iso_code', axis=1)
```

```
fig, ax = plt.subplots(1, 1,figsize=(10,10))
countries.plot(column='total_vaccinations',legend=True,ax=ax,
        legend_kwds={'label': "Number of Covid-19 Vaccinations Administered",'orientation': "horizontal"},
        missing_kwds={"color": "lightgrey"})
plt.show()
# Η χώρα που έχει χορηγήσει τον μεγαλύτερο αριθμό εμβολίων ανά κάτοικο(Σημειώστε ότι, δεδομένου ότι κάθε άτομο απαιτεί κατά μέσο όρο 2 δόσεις, αναμένουμε ότι μια πλήρως εμβολιασμένη χώρα θα έχει 200 εμβολιασμούς ανά 100 κατοίκους)
total = data.loc[data.groupby('country')['total_vaccinations_per_hundred'].idxmax()].sort_values('total_vaccinations_per_hundred',ascending=False)
print(total.iloc[0].country, " με ", total.iloc[0].total_vaccinations_per_hundred, " εμβόλια ανα 100 κατοίκους, μέχρι 19/02/2021 ")
# Γράφημα με τις 20 χώρες που έχουν πραγματοποιήσει τους περισσότερους εμβολιασμούς σε σχέση με τον πληθυσμό τους
plt.figure(figsize=(10, 4))
plt.bar(total.country[0:20], total.total_vaccinations_per_hundred[0:20])
plt.xticks(rotation=90)
plt.title('Οι 20 χώρες που έχουν πραγματοποιήσει τους περισσότερους εμβολιασμούς σε σχέση με τον πληθυσμό τους')
plt.xlabel('Χώρα')
plt.ylabel('Συνολικός αριθμός εμβολίων ανα 100 άτομα')
plt.show()
#Γεωγραφικός χάρτης των χωρών που έχουν πραγματοποιήσει τους περισσότερους εμβολιασμούς σε σχέση με τον πληθυσμό τους
countries = gpd.read_file(gpd.datasets.get_path("naturalearth_lowres"))

countries = countries.merge(total[['iso_code', 'total_vaccinations_per_hundred']], how = 'left',
        left_on = 'iso_a3', right_on = 'iso_code').drop('iso_code', axis=1)

fig, ax = plt.subplots(1, 1,figsize=(10,10))

countries.plot(column='total_vaccinations_per_hundred',legend=True,ax=ax,
legend_kwds={'label': "Number of Covid-19 Vaccinations Administered",'orientation': "horizontal"},missing_kwds={"color": "lightgrey"})
plt.show()
# Το πιο δημοφιλές εμβόλιο στον κόσμο. Αυτό προσδιορίζεται ποσοτικά με βάση τον αριθμό των χωρών στις οποίες χρησιμοποιείται
```

```
total = data.loc[data.groupby('country')['total_vaccinations'].idxmax()].sort_values('total_vaccinations',ascending=False)
total = pd.concat((total,total["vaccines"].str.split(", ", expand = True)),axis=1)

types = total.iloc[:,13:].apply(pd.Series.value_counts).sum(axis=1).sort_values(ascending=False)

print("Το πιο δημοφιλές εμβόλιο είναι το ", types.index[0], " με ", int(types[0]), "χώρες να το έχουν χορηγήσει")
# Πώς εξελίσσονται τα υπόλοιπα εμβόλια όσον αφορά τη δημοτικότητά τους σε όλο τον κόσμο;
plt.bar(types.index, types)
plt.xticks(rotation=90)
plt.title('Περισσότερο δημοφιλή εμβόλια')
plt.xlabel('Εμβόλιο')
plt.ylabel('Αριθμός χωρών')
plt.show()
# Πόσα είδη εμβολίων έχουν χορηγηθεί σε κάθε χώρα
total['vacc_brands'] = total.iloc[:,-5:].apply(lambda x: (5 - x.isnull().sum()), axis='columns')
total = total.sort_values('vacc_brands',ascending=False)
plt.figure(figsize=(20, 4))
plt.bar(total.country, total.vacc_brands)
plt.xticks(rotation=90)
plt.title('Συνολικός αριθμός ειδών εμβολίων που έχουν χορηγηθεί')
plt.xlabel('Χώρα')
plt.ylabel('Αριθμός ειδών εμβολίων')
plt.show()
# Από την άποψη των πωλήσεων εμβολίων και του αριθμού των εμβολίων που χορηγούνται, ποια είδη εμβολίων είναι πιο δημοφιλείς;
# Σημειώστε ότι για τις χώρες που αγοράζουν περισσότερους από 1 τύπους εμβολίων, οι συνολικές αγορές υποτίθεται ότι κατανέμονται ισομερώς μεταξύ των τύπων εμβολίων που χορηγούνται.
total = data.loc[data.groupby('country')['total_vaccinations'].idxmax()].sort_values('total_vaccinations',ascending=False)
total = pd.concat((total,total["vaccines"].str.split(", ", expand = True)),axis=1)

types = total.iloc[:,13:].apply(pd.Series.value_counts).sum(axis=1).sort_values(ascending=False)

frame = { 'number_of_countries': types }
result = pd.DataFrame(frame)
```

```
result['number_sold'] = 0


for index, row in total.iterrows():
   if(row.iloc[-1] is None):
      if(row.iloc[-2] is None):
         if(row.iloc[-3] is None):
            if(row.iloc[-4] is None):
               result.loc[row.iloc[-5],'number_sold'] = result.loc[row.iloc[-5],'number_sold'] +
row.total_vaccinations
            else:
               result.loc[row.iloc[-4],'number_sold'] = result.loc[row.iloc[-4],'number_sold'] +
(1/2*row.total_vaccinations)
               result.loc[row.iloc[-5],'number_sold'] = result.loc[row.iloc[-5],'number_sold'] +
(1/2*row.total_vaccinations)
         else:
            result.loc[row.iloc[-3],'number_sold'] = result.loc[row.iloc[-3],'number_sold'] +
(1/3*row.total_vaccinations)
            result.loc[row.iloc[-4],'number_sold'] = result.loc[row.iloc[-4],'number_sold'] +
(1/3*row.total_vaccinations)
            result.loc[row.iloc[-5],'number_sold'] = result.loc[row.iloc[-5],'number_sold'] +
(1/3*row.total_vaccinations)
      else:
         result.loc[row.iloc[-2],'number_sold'] = result.loc[row.iloc[-2],'number_sold'] +
(1/4*row.total_vaccinations)
         result.loc[row.iloc[-3],'number_sold'] = result.loc[row.iloc[-3],'number_sold'] +
(1/4*row.total_vaccinations)
         result.loc[row.iloc[-4],'number_sold'] = result.loc[row.iloc[-4],'number_sold'] +
(1/4*row.total_vaccinations)
         result.loc[row.iloc[-5],'number_sold'] = result.loc[row.iloc[-5],'number_sold'] +
(1/4*row.total_vaccinations)
   else:
      result.loc[row.iloc[-1],'number_sold'] = result.loc[row.iloc[-1],'number_sold'] +
(1/5*row.total_vaccinations)
      result.loc[row.iloc[-2],'number_sold'] = result.loc[row.iloc[-2],'number_sold'] +
(1/5*row.total_vaccinations)
      result.loc[row.iloc[-3],'number_sold'] = result.loc[row.iloc[-3],'number_sold'] +
(1/5*row.total_vaccinations)
      result.loc[row.iloc[-4],'number_sold'] = result.loc[row.iloc[-4],'number_sold'] +
(1/5*row.total_vaccinations)
      result.loc[row.iloc[-5],'number_sold'] = result.loc[row.iloc[-5],'number_sold'] +
(1/5*row.total_vaccinations)
```

```
result = result.sort_values('number_sold',ascending=False)
plt.figure(figsize=(10, 4))
plt.bar(result.index, result.number_sold)
plt.xticks(rotation=90)
plt.title('Κατά προσέγγιση ποσότητα κάθε χορηγούμενου εμβολίου')
plt.xlabel('Είδος εμβολίου')
plt.ylabel('Ποσότητα εμβολίων (x 10M)')
plt.show()
# Ημερήσιοι εμβολιασμοί στις ΗΠΑ
USA = data[data.country == 'United States']
plt.figure(figsize=(10, 4))
plt.plot(USA.date, USA.daily_vaccinations)
plt.xticks(rotation=90)
plt.title('Ημερήσιος αριθμός εμβολιασμών στις ΗΠΑ')
plt.xlabel('Ημερομηνία')
plt.ylabel('Αριθμός εμβολιασμών (x 1M)')
plt.show()
# Αθροιστικοί εμβολιασμοί στις ΗΠΑ
USA = data[data.country == 'United States']
plt.figure(figsize=(10, 4))
plt.plot(USA.date, USA.total_vaccinations)
plt.xticks(rotation=90)
plt.title('Συνολικός αριθμός εμβολιασμών στις ΗΠΑ')
plt.xlabel('Ημερομηνία')
plt.ylabel('Συνολικοί εμβολιασμοί (x 10M)')
plt.show()
# Δημιουργία νέου dataframe που περιέχει μόνο τις στήλες "date" και "total_vaccinations"
USA_data = USA.loc[:,['total_vaccinations','date']]
USA_data.head(5)
USA_data.shape
# Μετατροπή όλων των ημερομηνιών στην τιμή "t", που αντιπροσωπεύει τις ημέρες από την
έναρξη των εμβολιασμών στις ΗΠΑ.
USA_data["t"] = (USA_data['date'] - USA_data.date.iloc[0]).dt.days + 1
USA_data["t_squared"] = USA_data["t"]*USA_data["t"]
USA_data["log"] = np.log(USA_data["t"])
USA_data["exp"] = np.log(USA_data["total_vaccinations"])
USA_data["sqrt"] = np.sqrt(USA_data["t"])
# Χωρισμός του dataset σε training set(80%) και test set(20%)
train_set = USA_data.iloc[0:int(np.floor(0.8*len(USA_data))),:]
test_set = USA_data.iloc[int(np.floor(0.8*len(USA_data))):,:]
train_set.shape
```

```
test_set.shape
# Γραμμικό Μοντέλο
linear_model = smf.ols('total_vaccinations ~ t', data=train_set).fit()
pred_linear =  pd.Series(linear_model.predict(pd.DataFrame(test_set['t'])))
rmse_linear = np.sqrt(np.mean((np.array(test_set['total_vaccinations'])-
np.array(pred_linear))**2))
# Εκθετικό μοντέλο
exponential_model = smf.ols('exp ~ t', data=train_set).fit()
pred_exponential = pd.Series(exponential_model.predict(pd.DataFrame(test_set['t'])))
rmse_exponential = np.sqrt(np.mean((np.array(test_set['total_vaccinations'])-
np.array(np.exp(pred_exponential)))**2))

# Quadratic μοντέλο
quadratic_model = smf.ols('total_vaccinations ~ t + t_squared',data=train_set).fit()
pred_quadratic = pd.Series(quadratic_model.predict(test_set[["t","t_squared"]]))
rmse_quadratic = np.sqrt(np.mean((np.array(test_set['total_vaccinations'])-
np.array(pred_quadratic))**2))

# Log model
log = smf.ols('total_vaccinations ~ log',data=train_set).fit()
pred_log = pd.Series(log.predict(pd.DataFrame(test_set[["log"]])))
rmse_log = np.sqrt(np.mean((np.array(test_set['total_vaccinations'])-np.array(pred_log))**2))

# Sqrt model
sqrt = smf.ols('total_vaccinations ~ sqrt',data=train_set).fit()
pred_sqrt = pd.Series(sqrt.predict(pd.DataFrame(test_set[["sqrt"]])))
rmse_sqrt = np.sqrt(np.mean((np.array(test_set['total_vaccinations'])-
np.array(pred_sqrt))**2))
#Εξαγωγή του RMSE για κάθε μοντέλο
m =
{"Model":pd.Series(["rmse_linear","rmse_exponential","rmse_quadratic","rmse_log","rmse_s
qrt"]),"RMSE_Values":pd.Series([rmse_linear,rmse_exponential,rmse_quadratic,rmse_log,rm
se_sqrt])}
table=pd.DataFrame(m)
table
quadratic_model.summary()
e = pd.Series(list(range(USA_data.t.iloc[-1] + 1, USA_data.t.iloc[-1] + 150)))
e_squared = e*e
pred = pd.DataFrame({'t':e, 't_squared':e_squared})
# Πρόβλεψη του πότε οι ΗΠΑ θα επιτύχουν ανοσία της αγέλης, όταν ο συνολικός αριθμός των
εμβολιασμένων θα ξεπεράσουν τα 464 εκατομμύρια άτομα
pred_y = quadratic_model.predict(pred)
```

```
days_after = (pred_y > 464000000).idxmax() + 1 # to offset the first index of 0
herd_date = USA.iloc[-1].date + datetime.timedelta(days=int(days_after))
print("Η μικρότερη ημερομηνία που οι ΗΠΑ είναι σε θέση να πετύχουν ανοσία της αγέλης είναι ", herd_date)
pred_y.index = pd.to_datetime(pred_y.index + 1, unit='D',origin=pd.Timestamp(USA.iloc[-1].date.tz_localize(None)))
# Γράφημα για την απεικόνιση των προβλέψεων μας πάνω από τα αρχικά δεδομένα
plt.figure(figsize=(10, 4))
plt.plot(USA_data.date, USA_data.total_vaccinations, label="Τρέχων αριθμός εμβολιασμών")
plt.plot(pred_y.index, pred_y,'--', label="Προβλεπόμενος αριθμός εμβολιασμών")
plt.plot(herd_date,pred_y[herd_date.tz_localize(None)], 'go', label="Σημείο που επιτυγχάνεται ανοσία της αγέλης")
plt.legend(loc="upper left")
plt.xticks(rotation=90)
plt.title('Συνολικός αριθμός εμβολιασμένων στις ΗΠΑ')
plt.xlabel('Ημερομηνία')
plt.ylabel('Συνολικοί εμβολιασμοί (x 100M)')
plt.show()
#Διαχωρισμός της εξαρτημένης μεταβλητής και της ανεξάρτητης μεταβλητής
X=USA_data[["t"]]
Y=USA_data[["total_vaccinations"]]
#Εισαγωγή βιβλιοθηκών
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense
input_dim       = X.shape[1] #Αριθμός των νευρώνων στο input layer
n_neurons       =  25       #Αριθμός των νευρώνων στο hidden layer
epochs          = 150       # Αριθμός των επαναλήψεων
model = Sequential()
# input layer
model.add(Dense(n_neurons, input_dim=input_dim,
                kernel_initializer='normal',
                activation='relu'))
# output layer
model.add(Dense(1, kernel_initializer='normal'))
# Εφαρμογή του μοντέλου
model.compile(loss='mean_squared_error', optimizer='adam')
model.fit(X, Y, epochs=epochs, verbose=0)
#Πληροφορίες του μοντέλου
model.summary()
predictions = model.predict(X)
#Εκτίμηση του R^2
metrics.explained_variance_score(Y,predictions)
```

Βιβλιογραφία

Ελληνική

Γ.Κ.Τούντας (2005) Αγωγή και Προαγωγή Υγείας. Διαθέσιμο στον δικτυακό τόπο:http://asclepieion.mpl.uoa.gr/pubaspis/%CE%91%CE%B3%CF%89%CE%B3%CE%AE_%CE%BA%CE%B1%CE%B9_%CE%A0%CF%81%CE%BF%CE%B1%CE%B3%CF%89%CE%B3%CE%AE.htm (19/8/2022)

Ξένη

Airoldi, C., Ferrante, D., Miligi, L., Piro, S., Stoppa, G., Migliore, E., Chellini, E., Romanelli, A., Sciacchitano, C., Mensi, C. and Cavone, D., 2020. Estimation of Occupational Exposure to Asbestos in Italy by the Linkage of Mesothelioma Registry (ReNaM) and National Insurance Archives. Methodology and Results. International Journal of Environmental Research and Public Health, 17(3), p.1020.

Angwin, J., Larson, J., Mattu, S. and Kirchner, L., 2016. Machine bias. In Ethics of Data and Analytics (pp. 254-264). Auerbach Publications.

Arora, P., Kumar, H. and Panigrahi, B.K., 2020. Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India. Chaos, Solitons & Fractals, 139, p.110017.

Ayyoubzadeh, S.M., Ayyoubzadeh, S.M., Zahedi, H., Ahmadi, M. and Kalhori, S.R.N., 2020. Predicting COVID-19 incidence through analysis of google trends data in Iran: data mining and deep learning pilot study. JMIR public health and surveillance, 6(2), p.e18828.

B.Krishnamurthy (2022) A rectified linear unit (ReLU) is an activation function that introduces the property of nonlinearity to a deep learning model and solves the vanishing gradients issue. Here's why it's so popular. Διαθέσιμο στον δικτυακό τόπο: https://builtin.com/machine-learning/relu-activation-function (20/3/2023)

Barret, M.A., Humblet, O., Hiatt, R.A. and Adler, N.E., Big data and disease prevention: From quantified self to quantified communities, in «Big Data», vol. 1, no. 3, 2013. DOI, 10, pp.168-175.

Chan, C.L. and Chang, C.C., 2020.Big Data, Decision Models, and Public Health. International Journal of Environmental Research and Public Health, 17(18), p.6723.(σελ.1)

Chimmula, V.K.R. and Zhang, L., 2020. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos, Solitons & Fractals, 135, p.109864.

COVID-19 Διαθέσιμο στον δικτυακό τόπο: https://en.wikipedia.org/wiki/COVID-19 (06/01/2023)

Crosignani, P., Massari, S., Audisio, R., Amendola, P., Cavuto, S., Scaburri, A., Zambon, P., Nedoclan, G., Stracci, F., Pannelli, F. and Vercelli, M., 2006. The Italian surveillance system for occupational cancers: characteristics, initial results, and future prospects. American journal of industrial medicine, 49(9), pp.791-798.

D'Souza & Dowdy (2021) Rethinking Herd Immunity and the Covid-19 Response End Game. Διαθέσιμο στον δικτυακό τόπο: https://publichealth.jhu.edu/2021/what-is-herd-immunity-and-how-can-we-achieve-it-with-covid-19 (05/01/2023)

Deiner, M.S., Lietman, T.M., McLeod, S.D., Chodosh, J. and Porco, T.C., 2016. Surveillance tools emerging from search engines and social media data for determining eye disease patterns. JAMA ophthalmology, 134(9), pp.1024-1030.

Dhillon, A.S., Majumdar, S., St-Hilaire, M. and El-Haraki, A., 2018, July. A mobile complex event processing system for remote patient monitoring. In 2018 IEEE International Congress on Internet of Things (ICIOT) (pp. 180-183). IEEE.

Dolley, S., 2018. Big data's role in precision public health. Frontiers in public health, p.68.

Eysenbach, G., 2002. Infodemiology: The epidemiology of (mis) information. The American journal of medicine, 113(9), pp.763-765.

Ginsberg, J., 2009. mohebbi, mH. Patel, Rs, Brammer, l., smolinski, ms, & Brilliant, l, pp.1012-1014.

Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J. and Kim, R., 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. Jama, 316(22), pp.2402-2410.

How Neural Networks are used for Regression in R Programming (2020). Διαθέσιμο στον δικτυακό τόπο: https://www.geeksforgeeks.org/how-neural-networks-are-used-for-regression-in-r-programming/ (20/3/2023)

Huang, M.Y., Wang, M.Y., Lin, Y.S., Lin, C.J., Lo, K., Chang, I.J., Cheng, T.Y., Tsai, S.Y., Chen, H.H., Lin, C.Y. and Liu, S.J., 2020. The association between metabolically healthy obesity, cardiovascular disease, and all-cause

mortality risk in Asia: a systematic review and meta-analysis. International journal of environmental research and public health, 17(4), p.1320.

Jain, S.H., Powers, B.W., Hawkins, J.B. and Brownstein, J.S., 2015. The digital phenotype. Nature biotechnology, 33(5), pp.462-463.

Jain, S.H., Powers, B.W., Hawkins, J.B. and Brownstein, J.S., 2015. The digital phenotype. Nature biotechnology, 33(5), pp.462-463.

Khoury, M.J., 2015. Precision public health and precision medicine: two peas in a pod. Blogs. CDC. Available online at: https://blogs. cdc. gov/genomics/2015/03/02/precision-public/(accessed August 29, 2019).

Knight, W., 2017. Biased algorithms are everywhere, and no one seems to care. Technology Review, p.2018.

Korbut, A.I., Klimontov, V.V., Vinogradov, I.V. and Romanov, V.V., 2019. Risk factors and urinary biomarkers of non-albuminuric and albuminuric chronic kidney disease in patients with type 2 diabetes. World Journal of Diabetes, 10(11), p.517.

Kshirsagar, A.V., Bang, H., Bomback, A.S., Vupputuri, S., Shoham, D.A., Kern, L.M., Klemmer, P.J., Mazumdar, M. and August, P.A., 2008. A simple algorithm to predict incident kidney disease. Archives of internal medicine, 168(22), pp.2466-2473.

Mahmood, U., Healy, H.G., Kark, A., Cameron, A., Wang, Z., Abeysekera, R. and Hoy, W.E., 2017. Spectrum (characteristics) of patients with chronic kidney disease (CKD) with increasing age in a major metropolitan renal service. BMC nephrology, 18(1), pp.1-10.

Morgenstern, J.D., Rosella, L.C., Costa, A.P., de Souza, R.J. and Anderson, L.N., 2021. Perspective: big data and machine learning could help advance nutritional epidemiology. Advances in Nutrition, 12(3), pp.621-631.

Oddone, E., Edefonti, V., Scaburri, A., Vai, T., Crosignani, P. and Imbriani, M., 2013. Female breast cancer in Lombardy, Italy (2002–2009): A case–control study on occupational risks. American journal of industrial medicine, 56(9), pp.1051-1062.

Odlum, M. and Yoon, S., 2015. What can we learn about the Ebola outbreak from tweets?. American journal of infection control, 43(6), pp.563-571.

Prosperi, M., Min, J.S., Bian, J. and Modave, F., 2018. Big data hurdles in precision medicine and precision public health. BMC medical informatics and decision making, 18(1), pp.1-15.

Qin, L., Sun, Q., Wang, Y., Wu, K.F., Chen, M., Shia, B.C. and Wu, S.Y., 2020. Prediction of number of cases of 2019 novel coronavirus (COVID-19) using

social media search index. International journal of environmental research and public health, 17(7), p.2365.

Ram, S., Zhang, W., Williams, M. and Pengetnze, Y., 2015. Predicting asthma-related emergency department visits using big data. IEEE journal of biomedical and health informatics, 19(4), pp.1216-1223.

Reece, A.G. and Danforth, C.M., 2017. Erratum to: Instagram photos reveal predictive markers of depression (EPJ Data Science,(2017), 6, 1,(15), 10.1140/epjds/s13688-017-0110-z).

Reps, J.M., Schuemie, M.J., Suchard, M.A., Ryan, P.B. and Rijnbeek, P.R., 2018. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. Journal of the American Medical Informatics Association, 25(8), pp.969-975.

S. Haykin. "Neural Networks, A Comprehensive foundation", 2nd edition. Prentice Hall, 1999

Savage, N., 2017. Calculating disease. Nature, 550(7676), pp.S115-S117.

Sazli, M.H., 2006. A brief review of feed-forward neural networks. Communications Faculty of Sciences University of Ankara Series A2-A3 Physical Sciences and Engineering, 50(01).

Shih, C.C., Lu, C.J., Chen, G.D. and Chang, C.C., 2020. Risk prediction for early chronic kidney disease: results from an adult health examination program of 19,270 individuals. International Journal of Environmental Research and Public Health, 17(14), p.4973.

Tseng, Y.J., Hu, R.F., Lee, S.T., Lin, Y.L., Hsu, C.L., Lin, S.W., Liou, C.W., Lee, J.D., Peng, T.I. and Lee, T.H., 2020. Risk factors associated with outcomes of recombinant tissue plasminogen activator therapy in patients with acute ischemic stroke. International journal of environmental research and public health, 17(2), p.618.

U.S. Department of Commerce (2022) U.S. Population Estimated at 332,403,650 on Jan. 1, 2022. Διαθέσιμο στον δικτυακό τόπο: https://www.commerce.gov/news/blog/2022/01/us-population-estimated-332403650-jan-1-2022#:~:text=As%20our%20nation%20prepares%20to,since%20New%20Year's%20Day%202021. (06/01/2023)

Velmovitsky, P.E., Bevilacqua, T., Alencar, P., Cowan, D. and Morita, P.P., 2021. Convergence of precision medicine and public health into precision public health: toward a big data perspective. Frontiers in Public Health, 9, p.561873.

WHO newsletter (2020) Coronavirus disease (COVID-19): Herd immunity, lockdowns and COVID-19. Διαθέσιμο στον δικτυακό τόπο: https://www.who.int/news-room/questions-and-answers/item/herd-immunity-lockdowns-and-covid-19#:~:text=The%20percentage%20of%20people%20who,among%20those%20who%20are%20vaccinated. (06/01/2023)

Wiemken, T.L. and Kelley, R.R., 2019. Machine Learning in Epidemiology and Health Outcomes Research. Annual review of public health, 41, pp.21-36.

Xiao, J., Ding, R., Xu, X., Guan, H., Feng, X., Sun, T., Zhu, S. and Ye, Z., 2019. Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. Journal of translational medicine, 17(1), pp.1-13.

Xue, N., Fang, Y., Ding, X., Wang, L., Xu, L., Jiang, X. and Zhang, X., 2019. Serum triglycerides are related to chronic kidney disease (CKD) stage 2 in young and middle-aged chinese individuals during routine health examination. Medical Science Monitor: International Medical Journal of Experimental and Clinical Research, 25, p.2445.

Yu, X., Stuart, A.L., Liu, Y., Ivey, C.E., Russell, A.G., Kan, H., Henneman, L.R., Sarnat, S.E., Hasan, S., Sadmani, A. and Yang, X., 2019. On the accuracy and potential of Google Maps location history data to characterize individual mobility for air pollution health studies. Environmental pollution, 252, pp.924-930.