Πρόγραμμα Μεταπτυχιακών Σπουδών

στην Αναλυτική των Επιχειρήσεων και Επιστήμη των Δεδομένων

Τμήμα Οργάνωσης και Διοίκησης Επιχειρήσεων

Διπλωματική Εργασία

Διερεύνηση και ανάλυση ανοικτών βϊοιατρικών δεδομένων μεγάλου όγκου με χρήση μηχανικής μάθησης: Οι περιπτώσεις των βάσεων UK Biobank και National Inpatient Sample.

Exploring and analyzing open and big biomedical data using machine learning: The cases of the UK Biobank and National Inpatient Sample databases.

Του Βασίλειου Βασιλείου του Αναστασίου

Υποβλήθηκε ως απαιτούμενο για την απόκτηση του μεταπτυχιακού διπλώματος στην Αναλυτική των Επιχειρήσεων και Επιστήμη των Δεδομένων

Φεβρουάριος 2024

Περίληψη

Στην παρούσα μελέτη μας σκοπός μας είναι, πέρα από την κλασσική στατιστική ανάλυση με SPSS, η δημιουργία με Machine Learning Models ενός προβλεπτικού μοντέλου σχετικά με το κόστος και το ποσοστό θνησιμότητας ενδονοσοκομειακών ασθενών. Για την δημιουργία του μοντέλου αντλήθηκαν δεδομένα από το US National Inpatient Sample (NIS) του έτους 2013. Τα δεδομένα αφορούν πάνω από $7*10^6$ ασθενείς για τους οποίους έχουν καταγραφεί 144 ξεχωριστά στοιχεία, δημογραφικά, δεδομένα υγείας, διαχειριστικά δεδομένα και τα οποία αποτελούν τις μεταβλητές – variables που χρησιμοποιήσαμε. Στηριχθηκαμε σε αντίστοιχες μελέτες που έχουν δημοσιευτεί στο παρελθόν [11] [12], εμπλουτίσαμε όμως την κλασσική στατιστική ανάλυση με Machine Learning (ML) Models τα οποία δίνουν καλύτερο προβλεπτικό αποτέλεσμα, καθώς στην περίπτωση των ML Models δεν είμαστε υποχρεωμένοι να διαλέξουμε εξ αρχής της μεταβλητές, οι οποίες πιθανολογούμε ότι μπορούν να συσχετίζονται και να επεξηγούν το κόστος της θεραπείας και το ποσοστό θνησιμότητας. Ταυτόχρονα χρησιμοποιώντας τον classifier XGBoost, μπορέσαμε να διατηρήσουμε την επεξηγηματικότητα (expainability) του μοντέλου, ώστε να συγχρονίζεται με την κλινική πρακτική.

Πριν την στατιστική ανάλυση και την δημιουργία του προγνωστικού μοντέλου, στοχεύοντας στην καλύτερη κατανόηση του περιεχομένου της έρευνας μας, θα αναλύσουμε δύο open source βάσεις βιοιατρικών και ιατρικών δεδομένων, τη UK Biobank και το US National Inpatient Sample (NIS).

Abstract

The purpose of this study is to develop an explainable ML model regarding total charges and morbidity, from a sample of inpatient participants. In order to achieve this objective, data from the 2013 National Inpatient Sample, were used. The cohort comprised $7*10^6$ patients, described by 144 predictor variables, combining demographic, health and administrative data. Our study is based on scientific papers [11] [12], properly elaborated, so as to enhance the conventional statistical analysis with a Machine Learning (ML) Model, able to predict, in an explainable way, the end result. One of the main advantages of ML models is that there is no

prerequisite to sort out the variables, which could correlate with the output (Total Charges & Morbidity). The XGBoost classifier which was applied, excibited a robust predictive ability, that forms part of an explainable AI model, in conformity with clinical and administrative practice.

Prior to our analysis, an introdustory documentation of the UK Biobank and the US National Inpatient Sample (NIS) will be presented.

Περιεχόμενα

Εισαγωγή

Σκοπός της παρούσας μελέτης είναι να μελετήσουμε και να προβλέψουμε τους παράγοντες, οι οποίοι είναι σε θέση να επηρεάσουν το συνολικό κόστος της ενδονοσοκομειακής θεραπείας καθώς και το ποσοστό θνησιμότητας των ασθενών.

Η παρούσα μελέτη είναι χωρισμένη σε τρία τμήματα. Πριν την ανάλυση των δεδομένων μας, έχοντας ως στόχο την καλύτερη κατανόηση της έρευνας μας, εξετάζουμε τις δύο μεγαλύτερες open source βάσεις βιοιατρικών και ιατρικών δεδομένων παγκοσμίως της UK Biobank και του US National Inpatient Sample (NIS). Θα μελετήσουμε τον τρόπο με τον οποίων συστήνεται η βάση, τα δεδομένα που περιέχει καθώς και τεχνικά θέματα, όπως τους τύπους των αρχείων τους και τους δυνατούς τρόπους επεξεργασίας τους.

Η πρώτη παρουσίαση αφορά την UK Biobank, τη μεγαλύτερης ανοιχτή στους ερευνητές βιοιατρική βάση δεδομένων παγκοσμίως. Θα παρουσιάσουμε κατ' αρχάς τα Datasets τα οποία οποία συλλέγονται στην βάση. Στην συνέχεια θα επεκταθούμε στην κατηγοριοποίηση των δεδομένων, αλλά και στην κωδικοποίησης τους. Η κωδικοποίηση όπως θα δούμε αναλυτικότερα αναφέρεται τόσο στο είδος, στον τύπο των δεδομένων των βάσεων Datasets, όσο και στις μεταβλητές Variables – Data Fields που περιέχονται στα Datasets. Πέρα από την κωδικοποίηση των Data Fields, ενα σημαντικό τμήμα της παρουσίασης αναλύει με συγκεκριμένα παραδείγματα την ιεράρχιση των Data fields σε Categories & Subcategories, καθώς τα Data fields είναι interrelated.

Επιπροσθέτως παρουσιάζονται αναλυτικές οδηγίες σχετικά με την αναζήτηση και την πρόσβαση στα αρχεία του UKB, ταυτόχρονα με μια πλήρη ανάπτυξη των φακέλων που περιέχονται καθώς και του τρόπου που μπορούμε ως ερευνητές να έχουμε πρόσβαση στα αρχεία.

Στην συνέχεια εξετάζεται το US National Inpatient Sample (NIS), το οποίο έχει δημιουργηθεί ως τμήμα του Healthcare Cost and Utilization Project (HCUP) και περιέχει

διαπολιτειακά δεδομένα υγείας (multistate health data system). Το NIS δημιουργήθηκε αρχικά από τις βάσεις δεδομένων των νοσοκομείων, που είχαν ως σκοπό την καταγραφή των διαχειριστικών διαδικασιών και του κόστους, εν τούτοις παρέχει σημαντικές πληροφορίες για τα αποτελέσματα της περίθαλψης των ασθενών σε εθνικό επίπεδο. Αναλύεται όπως και στο UK Biobank ο τύπος των αρχείων, καθώς και ο τρόπος αναγωγής σε εθνικό επίπεδο του δείγματος των δεδομένων, που συλλέγονται από στρωματοποιημένα δεδομένα των πολιτείων της Αμερικής. Κλείνοντας το documentation των δύο διαφορετικών αυτών βάσεων, εξετάζουμε τις διαφορές στον τρόπο συλλογής και το είδος των στοιχείων- μεταβλητών που συλλέγονται.

Το τρίτο μέρος της έρευνας μας χωρίζεται εκ νέου σε δύο σκέλη. Το πρώτο είναι μελέτη των δεδομένων που προέρχονται από το NIS για το έτος 2013, όπου εξετάζουμε τις πιθανές συσχετίσεις διαφόρων μεταβλητών με τον συνολικό κόστος νοσηλείας καθώς και την συσχέτιση των ίδων μεταβλητών με το ποσοστό θνησιμότητας. Στο πρώτο αυτό σκέλος η μέλετη γίνεται με κλασσική μέθοδο στατιστικής αναλυσης με την βοήθεια του SPSS.

Στο δεύτερο σκέλος της ανάλυσης μας χρησιμόποιούμε το ίδιο Dataset από το NIS για το έτος 2013, αυτήν τη φορά όμως η επεξεργασία πραγματοποιείται με την χρήση Machine Learning Model (ML) και συγκεκριμένα με τον classifier XGBoost. Πέρα από την συσχέτιση εξετάζεται αν μπορεί να δημιουργηθεί ένα προβλεπτικό μοντέλο, το οποίο να εξετάζει το ποσοστό θνησιμότητας και το τελικό κόστος, λαμβάνοντας ως μεταβλητές κάποια από τα δεδομένα του Dataset.

Κεφάλαιο 1

UK Biobank Documentation

1.1 Εισαγωγή στο UK Biobank

Η UK Biobank είναι μια μεγάλης κλίμακας βιοϊατρική βάση δεδομένων και πηγή ερευνών που περιέχει αποπροσωποποιημένες γενετικές πληροφορίες, πληροφορίες για τον τρόπο ζωής και την υγεία και βιολογικά δείγματα από 500000 συμμετέχοντες στο Ηνωμένο Βασίλειο [2]. Είναι το πιο ολοκληρωμένο και ευρέως χρησιμοποιούμενο σύνολο δεδομένων του είδους του και είναι παγκοσμίως προσβάσιμο σε εγκεκριμένους ερευνητές, οι οποίοι διεξάγουν έρευνες σχετικά με την υγεία που είναι προς το κοινό όφελος συμφέρον, και μπορούν να προέρχονται από ακαδημαϊκό, εμπορικό, κυβερνητικό ή φιλανθρωπικό περιβάλλον [1].

Οι συμμετέχοντες στη Biobank του Ηνωμένου Βασιλείου έχουν παράσχει ένα ευρύ φάσμα πληροφοριών σχετικά με την υγεία και την ευεξία τους από την έναρξη της βιβλιοθήκης το 2006. Δεδομένα έχουν προστεθεί κατά την διάρκεια της λειτουργίας της και περιλαμβάνουν τα ακόλουθα είδη πληροφοριών:

- Imaging: Απεικόνιση εγκεφάλου, καρδιάς και μαγνητική τομογραφία (MRI) πλήρους σώματος, καθώς και σάρωση DEXA όλου του σώματος των οστών και των αρθρώσεων και υπερηχογράφημα των καρωτιδικών αρτηριών. Ο στόχος είναι η απεικόνιση 100.000 συμμετεχόντων και η πρόσκληση των συμμετεχόντων πίσω για επαναληπτική σάρωση μερικά χρόνια αργότερα.
- Genetics: Αλληλουχία πλήρους γονιδιώματος και για τους 500.000 συμμετέχοντες, αλληλουχία ολικού εξώματος - exome για 470.000 συμμετέχοντες, γονότυπος (800.000 παραλλαγές σε όλο το γονιδίωμα και μεταλλάξεις σε 90 εκατομμύρια παραλλαγές).
- Διασύνδεση με λοιπά δεδομένα υγείας: Σύνδεση με ένα ευρύ φάσμα ηλεκτρονικών αρχείων που σχετίζονται με την υγεία, συμπεριλαμβανομένων των

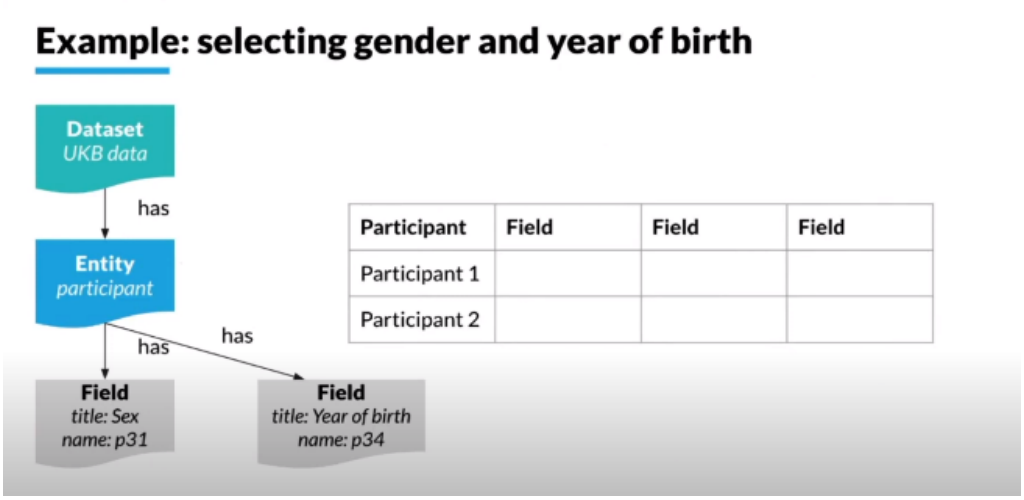
αρχείων θανάτου, καρκίνου, εισαγωγών στο νοσοκομείο και πρωτοβάθμιας περίθαλψης.

- Biomarkers/ Βιοδείκτες: Δεδομένα για περισσότερους από 30 βασικούς βιοχημικούς δείκτες από όλους τους συμμετέχοντες, που ελήφθησαν από δείγματα που συλλέχθηκαν κατά την πρόσληψη και την πρώτη επαναληπτική αξιολόγηση.
- Activity monitor/ Παρακολούθηση δραστηριότητας & Δεδομένα σωματικής δραστηριότητας: για μια περίοδο 7 ημερών που συλλέγονται μέσω συσκευής παρακολούθησης δραστηριότητας καρπού για 100.000 συμμετέχοντες, συν ένα εποχιακό follow up σε ένα υποσύνολο αυτού του πληθυσμού.
- Διαδικτυακά ερωτηματολόγια: Δεδομένα για μια σειρά στοιχείων και αποτελεσμάτων υγείας, που είναι δύσκολο να αξιολογηθούν μέσω αρχείων καταγραφής ρουτίνας του ιστορικού υγείας. Τέτοια δεδομένα αφορούν στη διατροφής, τις διατροφικές προτιμήσεις, του εργασιακού ιστορικού, του πόνου, της γνωστικής λειτουργίας, της πεπτικής υγείας και της ψυχικής υγείας.
- Επαναλαμβανόμενες αξιολογήσεις υγείας: Εκτελείται μια πλήρης βασική αξιολόγηση κατά τη διάρκεια της απεικονιστικής αξιολόγησης (imaging) των 100.000 συμμετεχόντων.
- Δείγματα: Συλλέχθηκαν αίμα και ούρα από όλους τους συμμετέχοντες και σάλιο για 100.000.

Τα δεδομένα της UK Biobank έχουν συγκεκριμένη μορφοποίηση, είναι διαταγμένα σε κατηγορίες και για την πρόσβαση σε αυτά απαιτούνται συγκεκριμένες διαδικασίες απο την ερευνητική ομάδα.

1.2 Σχετικά με Data Fields του UK Biobank

Τα δεδομένα στο UKB αποτελούν ένα Dataset. Εντός του Dataset περιέχονται ξεχωριστά Data Entities (participants), που στην ουσία είναι όλοι οι ξεχωριστοί συμμετέχοντες στις βιβλιοθήκη του UKB.



Εικόνα 1

Κάθε ξεχωριστό Entity είναι ένας πίνακας που περιέχει ένα ή περισσότερα Data Fields (Variables), ως μεταβλητές. Αυτά τα Data Fields ομαδοποιούνται σε διάφορες κατηγορίες ανάλογα με τις βασικές τους ιδιότητες και αναφέρονται σε μια συγκεκριμένη ερώτηση, μέτρηση ή αποτέλεσμα μέτρησης [9]. Οι βασικές ταξινομίσεις είναι οι ακόλουθες:

- Participants – Ο αριθμός των συμμετεχόντων στο data field [9].
- Item count - Ο αριθμός των συμμετεχόντων στο data field επί των αριθμό των μετρήσεων που πραγματοποιήθηκαν [9].
- Stability – Υποδεινύει αν η συλλογή των δεδομένων έχει ολοκληρωθεί, αν είναι σε εξέλιξη, αν οι τιμές τους ενδέχεται να αλλάξουν στην πάροδο του χρόνου [10]
- Value type – τον τύπο της μεταβλητής που αντιπροσοπεύει το κάθε data-field (κατηγορική, συνεχής, κείμενο, αρχείο, ημερομηνία κοκ) [11]
- Item type – Περιγράφει την φύση των δεδομένων. Μπορεί να ανήκει σε Data, Samples, Bulk & Records ανάλογα με τον τύπο των αρχείων και την φύση της πληροφορίας (εικόνα, κείμενο, μεγάλο αρχείο κοκ) [12].
- Strata – σε ποιο domain αναφέρεται το data-field. Στην ουσία εξετάζει αν τα δεδομένα είναι πρωτογενής μέτρηση (primary), συμπερασματική (derived) από

ένα ή περισσότερα διαφορετικά data fields, βοηθητικά δεδομένα (auxiliary) τα οποία υπερκαλύπτονται από τα primary kai derived [13].

- Sexed – σε ποιο φύλλο αναφέρονται τα δεδομένα (Αντρες, γυναίκες και τα δύο) [14].
- Instances – αν υπάρχουν πολλάπλές παρατηρήσεις για κάθε συμμετέχοντα. Λεπτομέρειες παρουσιάζονται στην συνέχεια. Οι παρατηρήσεις μπορεί να είναι μοναδικές (Singular), όπως το ετος γέννησης ή το φύλλο, καθορισμένες (defined) όπου είναι περισσότερες από μια αλλά κάθε μια παρουσιάζει ένα συγκεκριμένο αναγνωρίσιμο σύνολο αποτελεσμάτων μεταξύ όλων των συμμετεχόντων. Για παράδειγμα μπορεί η Τρίτη επίσκεψη όλων των ασθενών στο νοσοκομείο να καταγράφη την αρτηριακή τους πίεση. Μπορεί ακόμα να είναι μεταβαλόμενες (variable), όπου για παράδειγμα η τρίτη παρατήση για έναν ασθενή δεν σχετίζεται με την τρίτη παρατήρηση ενός άλλου ασθενή [15].
- Array – μπορεί να υπάρχουν μια ή περισσότερες παρατηρήσεις σε κάθε επίσκεψη/ instance. Το array μπορεί να είναι singular αν η παρατήρηση είναι μία, παραδείγματος χάριν η ερώτηση είναι πιο είναι το αγαπημένο σου χρώμα. Μπορεί να είναι multiple αν έχει περισσότερες από μια καταγραφές όπως η απάντηση στην ερώτηση «αναφέρετε τα αγαπημένα σας χρώματα» [16].
- Debut – πότε ενσωματώθηκε το field στην βάση δεδομένων [9].
- Version – Πότε έγινε η τελευταία ενσωμάτωση δεδομένων για αυτό το field στην βάση [9].

1.2.1 Κωδικοποίηση των Data Fields

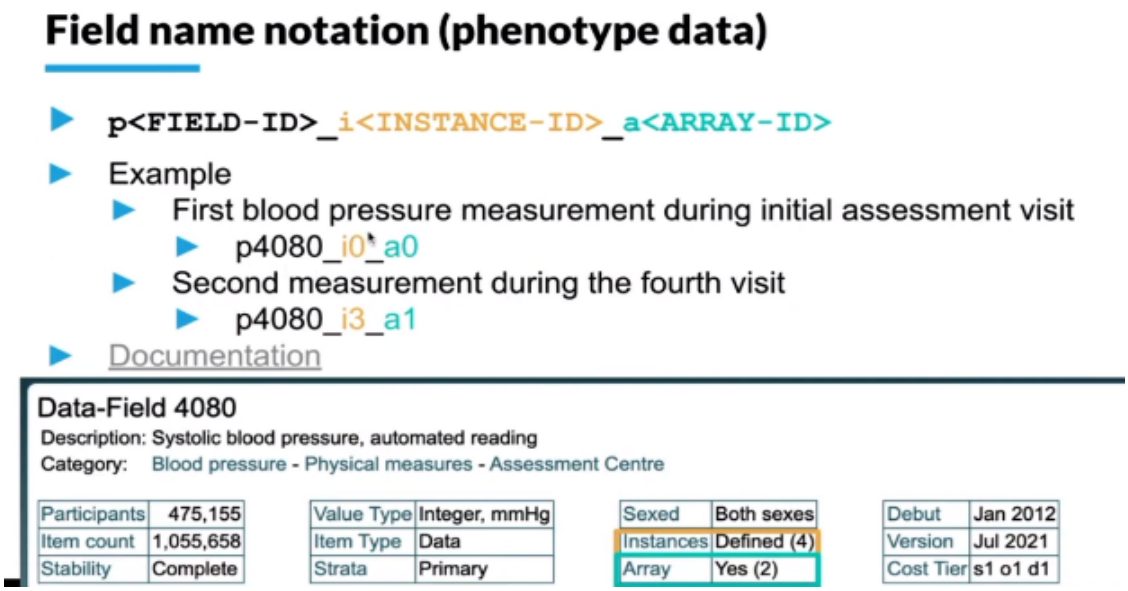
Τα Data Fields κωδικοποιούνται ως UDI (Unique Data Identifier) με την μορφή F-I.A.

F: Ο αριθμός που αντιστοιχεί στην κωδικοποίηση τους στο UKB Showcase. Πχ κωδικός Data Field 31 αντιστοιχεί στο φύλο ενώ ο κωδικός Data Field 34 στο έτος γέννησης, το Data Field 53 αντιστοιχεί στο Date of attending assessment center ενώ το Data Field 20002 στο Non-cancer illness code, self-reported, το Data Field 4080 Systolic blood pressure [17][5].

I: Ο δεύτερος αριθμός (I) αναφέρεται στις διαφορετικές στιγμές (instances) που έγινε κάθε καταχώρηση. 0 για την αρχική επίσκεψη, 1 για την δεύτερη κοκ.

A: Array Index που δείχνει την κάθε ξεχωριστή παρατήρηση (array) για κάθε instance. 0 για την πρώτη παρατήρηση, 1 για την δεύτερη κοκ.

Η μορφή δηλαδή που θα έχει κάθε Data Field που στην ουσία είναι μια μεταβλητή/ Variable του κάθε participant φαίνεται στο παρακάτω σχήμα:

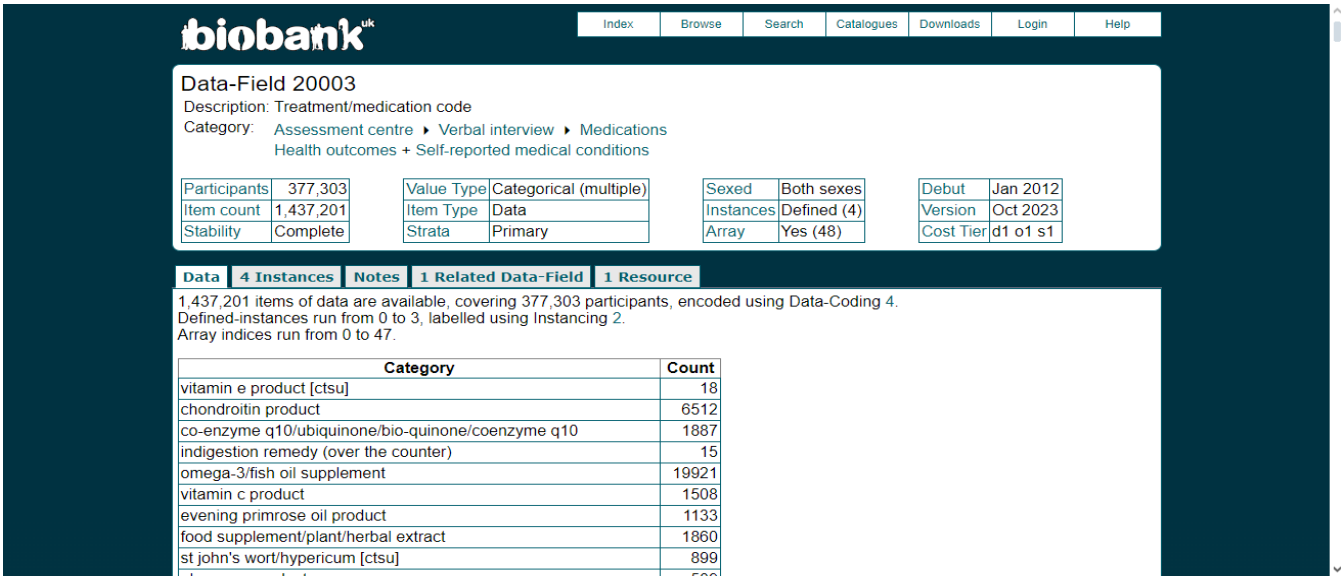


Εικόνα 2

Είναι σημαντικό να τονίσουμε ότι καθε Data Field απαντάει στην ουσία σε μια συγκεκριμένη ερώτηση, μέτρηση ή αποτέλεσμα. Ποια ήταν πχ η συστολική πίεση του ασθενή με eid 123456 στο δευτερο ραντεβού του ή ποια η φαρμακευτική αγωγή του ασθενούς..

Αν για παράδειγμα καταγράφηκε τέσσερις φορές η φαρμακευτική αγωγή του ασθενούς (Field Number 2003), όπως φαίνεται παρακάτω, θα υπάρχουν 4 instances και σε κάθε ένα από αυτά, το κάθε ένα ως ξεχωριστό array. Έστω ότι αναφερόμαστε στο πρώτο instance. Ο μέγιστος αριθμός των φαρμάκων που χρησιμοποιεί ο κάθε συμμετέχων (έστω ότι αυτά είναι 47) θα

καθορίσει των αριθμό των μεταβλητών για αυτό το field, όπως φαίνεται στον παρακάτω εικόνα 3 [6].



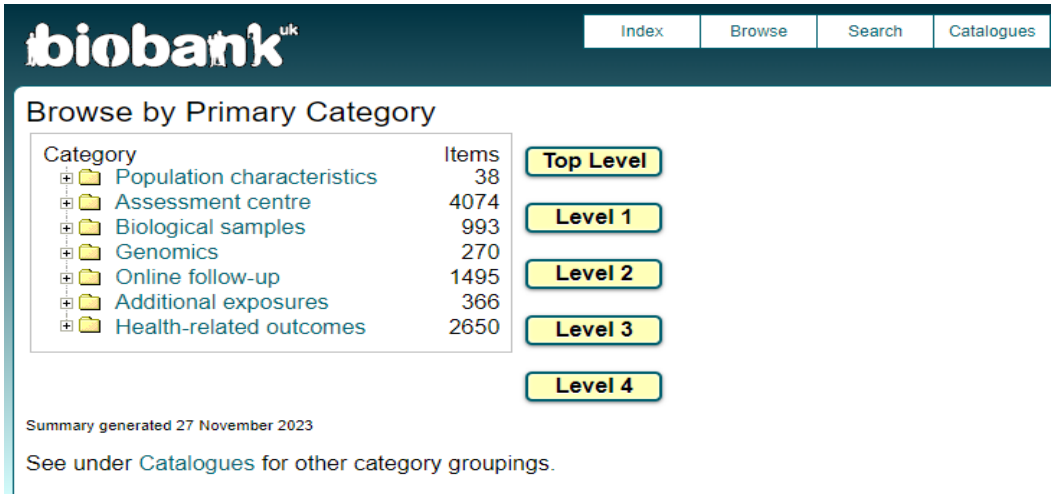
Εικόνα 3

Θα υπάρχουν 4 instances και σε κάθε ένα από αυτά, θα καταγράφηκε το κάθε ένα ως ξεχωριστό array. Έστω ότι αναφερόμαστε στο πρώτο instance. Ο μέγιστος αριθμός των φαρμάκων που χρησιμοποιεί ο κάθε συμμετέχων (έστω ότι αυτά είναι 47) θα καθορίσει των αριθμό των μεταβλητών για αυτό το field, όπως φαίνεται στον παρακάτω πίνακα [6].

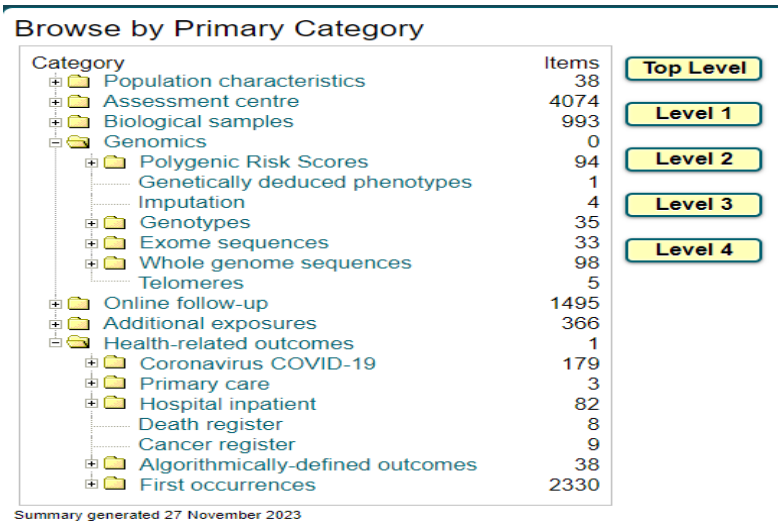
f.eid	f.20003.0.1	f.20003.0.2	f. 20003.0.3	…	f. 20003.0.47
5967229	NA	NA	NA	…	NA
4674807	178	1754	NA	…	NA
1456203	45	NA	NA	…	NA
3723112	1341	161	131	…	14

1.3 Σχετικά με τις Κατηγορίες/ Categories του UK Biobank

Τα δεδομένα στο UK Biobank είναι κατηγοριοποιημένα σε main Categories, όπου ανάλογα με τον τρόπο συλλογής των δεδομένων, αποτελούν το Top level και βρίσκονται μέσα από το browse (εικόνα 4 & 5). Αυτές στην συνέχεια χωρίζονται σε υποκατηγορίες Sub Categories, level 1, level 2, level 3 & level 4 [3][4].



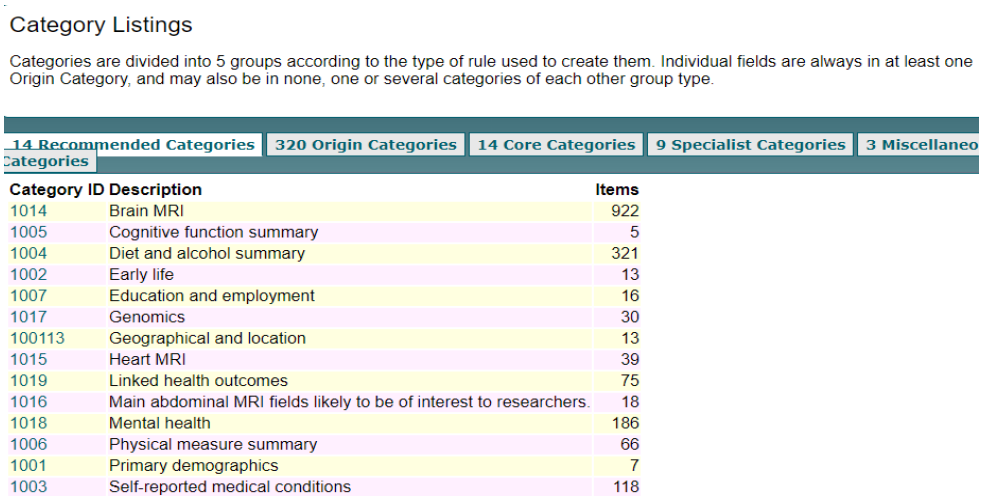
Εικόνα 4



Εικόνα 5

1.3.1 Category & Sub Categories in the UKB

Όπως αναφέρθηκε στην αρχή κάθε Data Field ανήκει σε μια Category – Sub Category. Υπάρχουν 320 sub categories (εικόνα 6), που είναι στην ουσία το level 4. Υπάρχουν 14 προτεινόμενες κατηγορίες οι οποίες έχουν διαμορφωθεί κατά τέτοιον τρόπο, ώστε να ζητούνται ζητούν απευθείας από τους ερευνητές. Επιπροσθέτως υπάρχει κι ένας διαχωρισμός (κυρίως χρηστικός) με 14 core categories όπου είναι στα level 1,2 & 3, όταν χρησιμοποιούμε το browse στην πλατφόρμα [7].



Category Listings

Categories are divided into 5 groups according to the type of rule used to create them. Individual fields are always in at least one Origin Category, and may also be in none, one or several categories of each other group type.

14 Recommended Categories | 320 Origin Categories | 14 Core Categories | 9 Specialist Categories | 3 Miscellaneo Categories

Category ID	Description	Items
1014	Brain MRI	922
1005	Cognitive function summary	5
1004	Diet and alcohol summary	321
1002	Early life	13
1007	Education and employment	16
1017	Genomics	30
100113	Geographical and location	13
1015	Heart MRI	39
1019	Linked health outcomes	75
1016	Main abdominal MRI fields likely to be of interest to researchers.	18
1018	Mental health	186
1006	Physical measure summary	66
1001	Primary demographics	7
1003	Self-reported medical conditions	118

Εικόνα 6

Να σημειώσουμε και πάλι ότι τα Categories περιέχουν ομαδοποιήσεις από Data Fields, τα οποία στην ουσία είναι Variables που μοιράζονται κάποιο κοινό χαρακτηριστικό. Μπορούν επί παραδείγματι να αναφέρονται στην ίδια πάθηση ή να είναι εργαστηριακά δεδομένα.

Παράδειγμα: Έστω θέλουμε να μελετήσουμε το Data Field 4080, με description Systolic blood pressure, automated reading (εικόνα 7).

Data-Field 4080

Description: Systolic blood pressure, automated reading

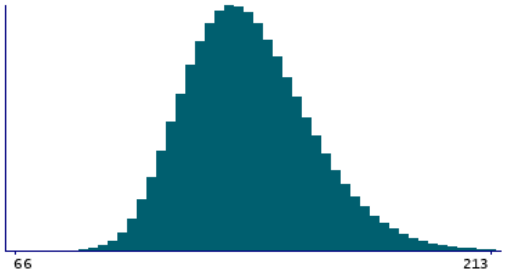
Category: Assessment centre ▸ Physical measures ▸ Blood pressure
Physical measure summary + Physical measures

Participants	475,939	Value Type	Integer, mmHg	Sexed	Both sexes	Debut	Jan 2012
Item count	1,098,848	Item Type	Data	Instances	Defined (4)	Version	Oct 2023
Stability	Complete	Strata	Primary	Array	Yes (2)	Cost Tier	d1 o1 s1

Data | 4 Instances | Notes | 3 Related Data-Fields | 5 Resources

1,098,848 items of data are available, covering 475,939 participants.
Defined-instances run from 0 to 3, labelled using Instancing 2.
Array indices run from 0 to 1.
Units of measurement are mmHg.

Maximum	268
Decile 9	164
Decile 8	154
Decile 7	147
Decile 6	141
Median	137
Decile 4	132
Decile 3	127
Decile 2	122
Decile 1	115
Minimum	51



- There are 197 distinct values.
- Mean = 138.231
- Std.dev = 19.4125
- 6 items below graph minimum of 66
- 1179 items above graph maximum of 213

Εικόνα 7

Από τα δεδομένα του πίνακα μπορούμε να συνάγουμε ότι είχαμε 475939 συμμετέχοντες, με 1098848 καταγραφές. Stability = Complete που σημαίνει ότι τα δεδομένα δεν πρόκειται να μεταβληθούν, η μέτρηση γίνεται σε ακέραιους αριθμούς στήλης υγραργύρου και το item type μας δείνει ότι είναι απλής δομής. Το strata type υποδυκνύει ότι τα δεδομένα συλλέγονται άμεσα, Sexed = both Sexes ότι αναφέρονται και στα δύο φύλλα, instances = 4 ότι έχουν γίνει 4 μετρήσεις, με Array = 2 δηλαδή δύο διαφορετικές καταγραφές[8].

Η τελευταία καταγραφή Cost Tier = d1o1s1 σημαίνει ότι για την πρόσβαση στα δεδομένα μέσω του MTA (Material Transfer Agreement) η ερευνητική ομάδα θα χρειαστεί να πληρώσει το Tier 1 για να έχει πρόσβαση στα δεδομένα για d= Download data from RAP, o = Online accessing data on RAP, s = Showcase downloading Data. Περισσότερες λεπτομέρειες σχετικά με την πληρωμή, στη συνέχεια στο κομμάτι της αίτησης για πρόσβαση στα δεδομένα.

Στο Data Field 4080 μπορούμε να δούμε την περιγραφή, το sub category (Physical measure summary + Physical measures) καθώς και το path από την αρχική κατηγορία (Assessment Center -> Physical measures -> Blood Pressure -> Physical measure summary + Physical measures). Επιπροσθέτως αν χτυπήσουμε πάνω σε κάθε στοιχείο του path βλέπουμε την κατηγορία στην οποία ανήκει, καθώς και τα υπόλοιπα Data Fields αυτής της κατηγορίας. Έτσι το Assessment Center είναι Category 100000, το Physical measures είναι Category 100006, το Blood Pressure είναι Category 100011, το Physical measure summary είναι Category 1006 και το Physical measures είναι Category 706.

Category 1006
Physical measure summary

Description
Main physical measurement fields likely to be of interest to researchers

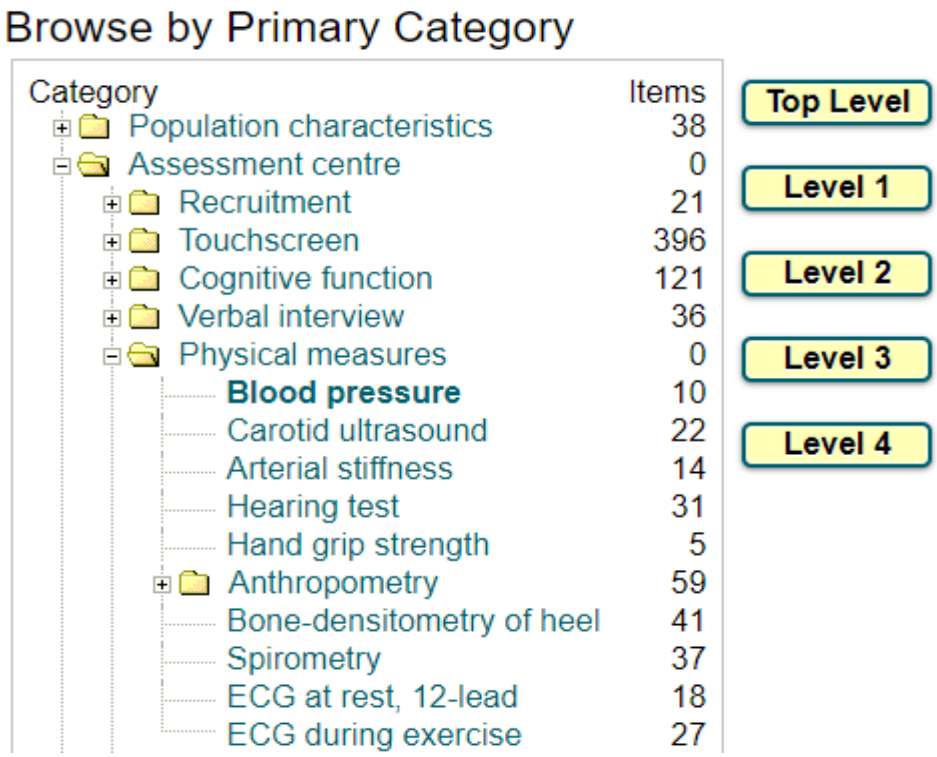
66 Data-Fields | 7 Applications

Field ID	Description	Category
90012	Overall acceleration average	Acceleration averages
4194	Pulse rate	Arterial stiffness
21021	Pulse wave Arterial Stiffness index	Arterial stiffness
4079	Diastolic blood pressure, automated reading	Blood pressure
94	Diastolic blood pressure, manual reading	Blood pressure
4081	Method of measuring blood pressure	Blood pressure
95	Pulse rate (during blood-pressure measurement)	Blood pressure
102	Pulse rate, automated reading	Blood pressure
4080	Systolic blood pressure, automated reading	Blood pressure
93	Systolic blood pressure, manual reading	Blood pressure
23124	Arm fat mass (left)	Body composition by impedance
23120	Arm fat mass (right)	Body composition by impedance
23123	Arm fat percentage (left)	Body composition by impedance
23119	Arm fat percentage (right)	Body composition by impedance
23125	Arm fat-free mass (left)	Body composition by impedance
23121	Arm fat-free mass (right)	Body composition by impedance
23126	Arm predicted mass (left)	Body composition by impedance
23122	Arm predicted mass (right)	Body composition by impedance

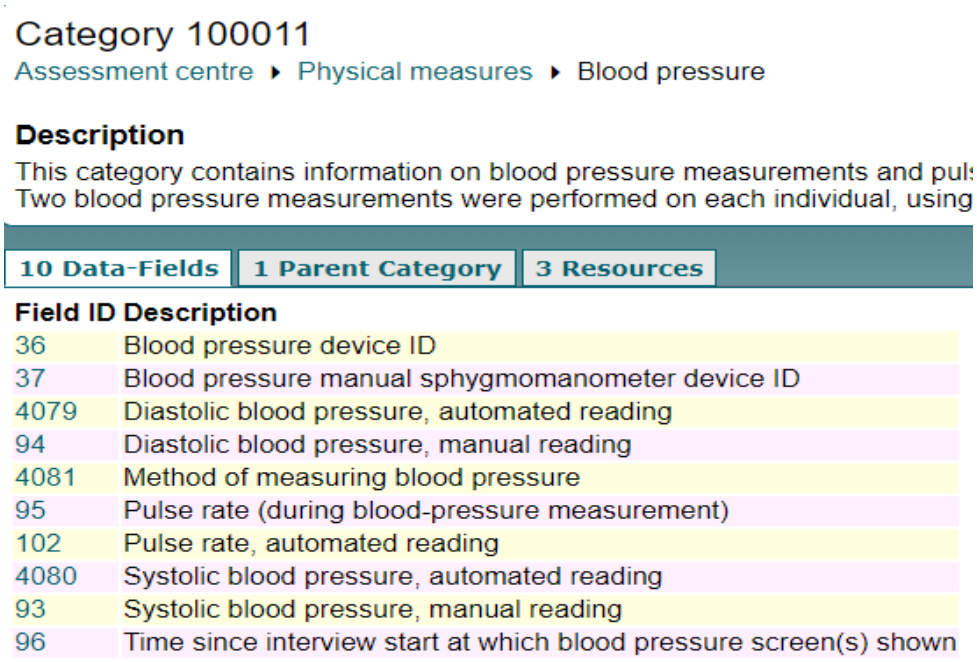
Εικόνα 8

Αν θέλουμε να δουμε που βρίσκεται στο tree των categories browse πατάμε στο browse και βλέπουμε αναλυτικά τα level. Παρατηρούμε ότι βρίσκεται στο level 3 του Assesment center -> Physical measures-> Blood Pressure. Ο αριθμός 10 δίπλα από σην subcategory blood pressure

υποδυκνύει πως υπάρχουν συνολικά 10 Data Fields σε αυτήν την υποκατηγορία,τα οποία φαίνονται στην παρακάτω εικόνα.



Εικόνα 9



Category 100011

Assessment centre ▸ Physical measures ▸ Blood pressure

Description

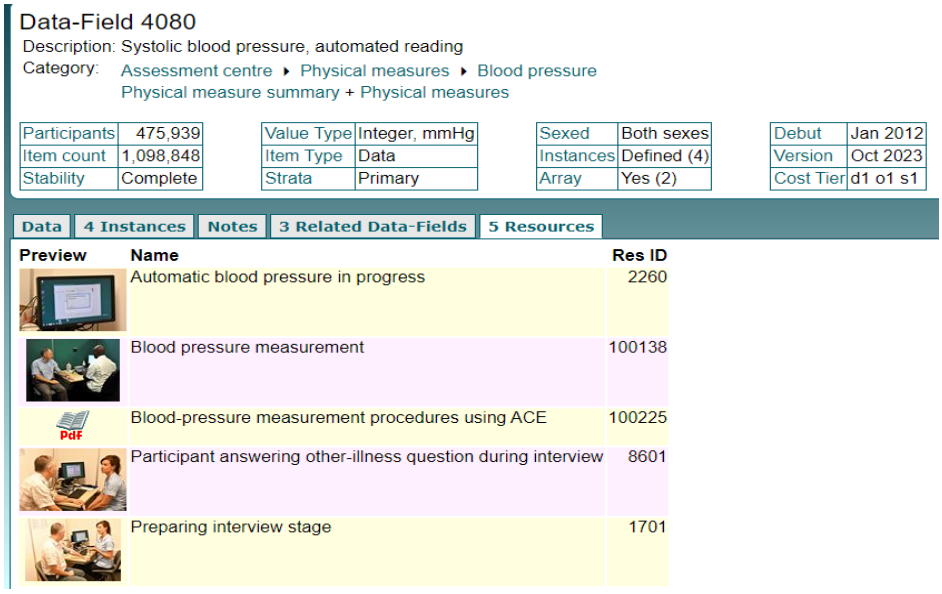
This category contains information on blood pressure measurements and puls
Two blood pressure measurements were performed on each individual, using

10 Data-Fields | 1 Parent Category | 3 Resources

Field ID	Description
36	Blood pressure device ID
37	Blood pressure manual sphygmomanometer device ID
4079	Diastolic blood pressure, automated reading
94	Diastolic blood pressure, manual reading
4081	Method of measuring blood pressure
95	Pulse rate (during blood-pressure measurement)
102	Pulse rate, automated reading
4080	Systolic blood pressure, automated reading
93	Systolic blood pressure, manual reading
96	Time since interview start at which blood pressure screen(s) shown

Εικόνα 10

Χτυπώντας στο resources μπορούμε να δούμε τον τρόπο συλλογής των δεδομένων.



Data-Field 4080

Description: Systolic blood pressure, automated reading

Category: Assessment centre ▸ Physical measures ▸ Blood pressure
Physical measure summary + Physical measures

Participants	475,939	Value Type	Integer, mmHg	Sexed	Both sexes	Debut	Jan 2012
Item count	1,098,848	Item Type	Data	Instances	Defined (4)	Version	Oct 2023
Stability	Complete	Strata	Primary	Array	Yes (2)	Cost Tier	d1 o1 s1

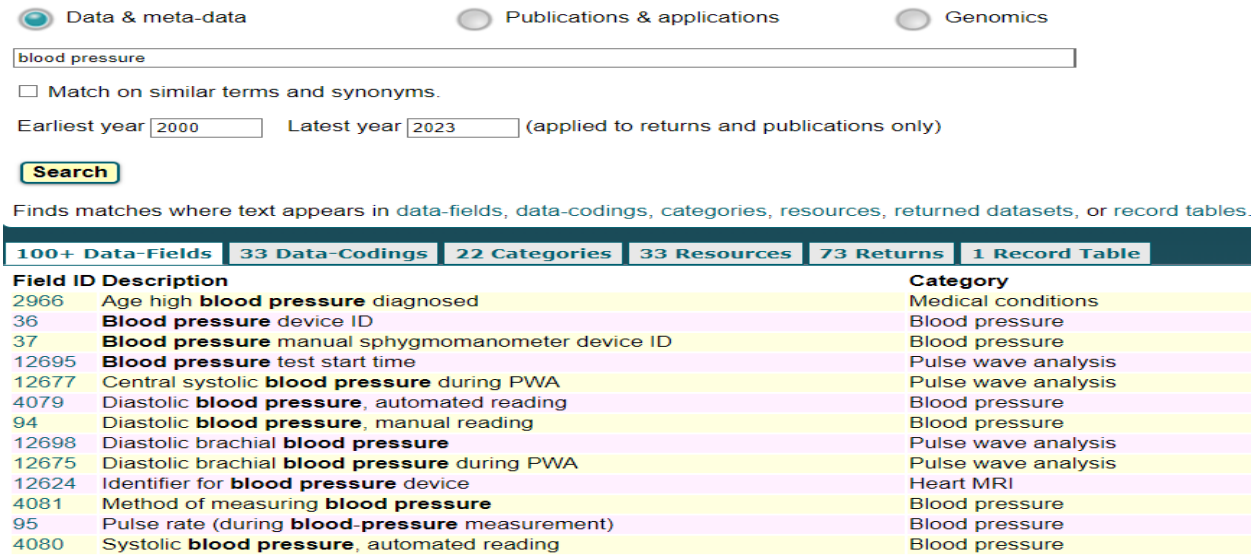
Data | 4 Instances | Notes | 3 Related Data-Fields | 5 Resources

Preview	Name	Res ID
	Automatic blood pressure in progress	2260
	Blood pressure measurement	100138
PDF	Blood-pressure measurement procedures using ACE	100225
	Participant answering other-illness question during interview	8601
	Preparing interview stage	1701

Εικόνα 11

1.4 Αναζήτηση μέσω της επιλογής Search

Έστω ότι στην επιλογή search αναζητούμε blood pressure (εικόνα 12). Τα αποτελέσματα της αναζήτησης δίνουν 100+ ξεχωριστά Data Fields = Variables, τα οποία βρίσκονται σε 22 διαφορετικές subcategories, όπως φαίνεται στις εικόνες 10 & 11.



Data & meta-data Publications & applications Genomics

blood pressure

Match on similar terms and synonyms.

Earliest year 2000 Latest year 2023 (applied to returns and publications only)

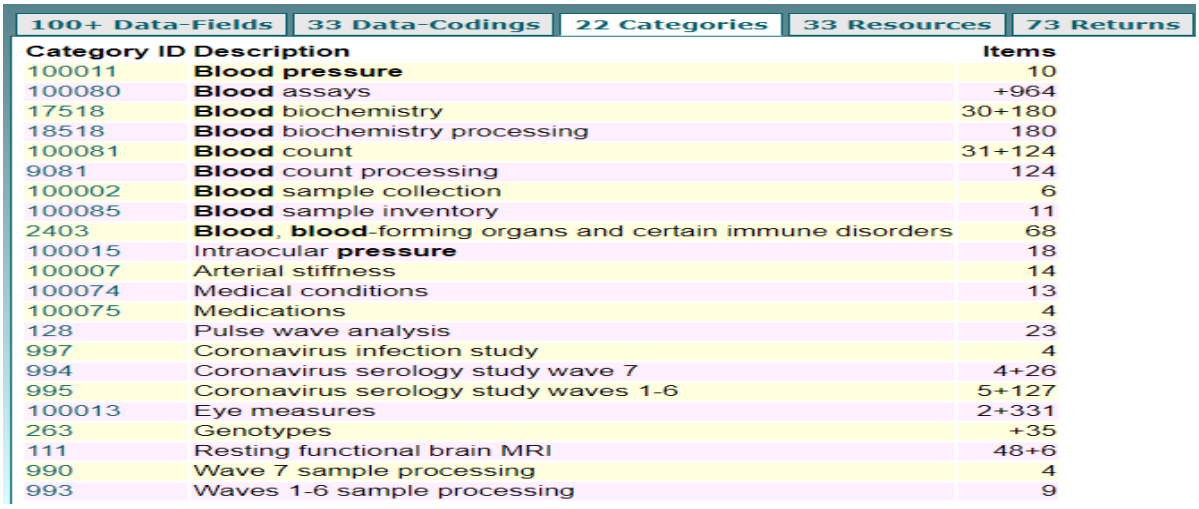
Search

Finds matches where text appears in data-fields, data-codings, categories, resources, returned datasets, or record tables.

100+ Data-Fields | 33 Data-Codings | 22 Categories | 33 Resources | 73 Returns | 1 Record Table

Field ID	Description	Category
2966	Age high **blood pressure** diagnosed	Medical conditions
36	**Blood pressure** device ID	Blood pressure
37	**Blood pressure** manual sphygmomanometer device ID	Blood pressure
12695	**Blood pressure** test start time	Pulse wave analysis
12677	Central systolic **blood pressure** during PWA	Pulse wave analysis
4079	Diastolic **blood pressure**, automated reading	Blood pressure
94	Diastolic **blood pressure**, manual reading	Blood pressure
12698	Diastolic brachial **blood pressure**	Pulse wave analysis
12675	Diastolic brachial **blood pressure** during PWA	Pulse wave analysis
12624	Identifier for **blood pressure** device	Heart MRI
4081	Method of measuring **blood pressure**	Blood pressure
95	Pulse rate (during **blood-pressure** measurement)	Blood pressure
4080	Systolic **blood pressure**, automated reading	Blood pressure

Εικόνα 12



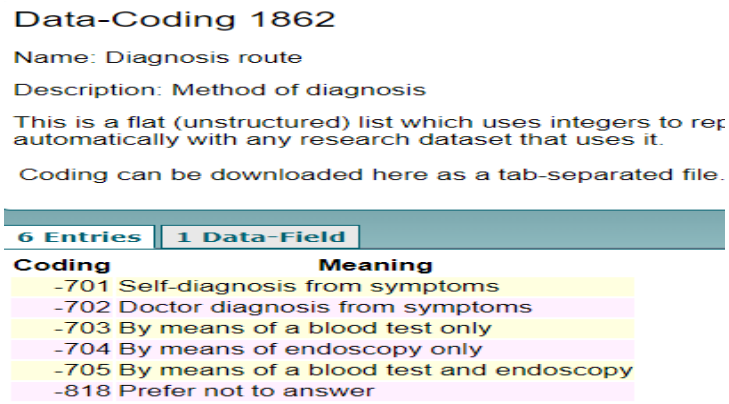
100+ Data-Fields | 33 Data-Codings | 22 Categories | 33 Resources | 73 Returns

Category ID	Description	Items
100011	**Blood pressure**	10
100080	**Blood** assays	+964
17518	**Blood** biochemistry	30+180
18518	**Blood** biochemistry processing	180
100081	**Blood** count	31+124
9081	**Blood** count processing	124
100002	**Blood** sample collection	6
100085	**Blood** sample inventory	11
2403	**Blood**, **blood**-forming organs and certain immune disorders	68
100015	Intraocular **pressure**	18
100007	Arterial stiffness	14
100074	Medical conditions	13
100075	Medications	4
128	Pulse wave analysis	23
997	Coronavirus infection study	4
994	Coronavirus serology study wave 7	4+26
995	Coronavirus serology study waves 1-6	5+127
100013	Eye measures	2+331
263	Genotypes	+35
111	Resting functional brain MRI	48+6
990	Wave 7 sample processing	4
993	Waves 1-6 sample processing	9

Εικόνα 13

Αξίζει να αναφερθούμε στα 33 διαφορετικά Data codings που χρησιμοποιούνται για τα κατηγορικά δεδομένα και στην ουσία είναι η αντιστοίχιση ανάμεσα στην τιμή και την κατηγορία που αποτυπώνεται στην βάση δεδομένων (εικόνα 13).

Παραδειγμα: Το Data Field 1862 (Method of Diagnosis) περιγράφει τον τρόπο της διάγνωσης. Στην εικόνα 14 βλέπουμε τον κωδικό που χρησιμοποιείται ανάλογως της οδού διάγνωσης.


Data-Coding 1862

Name: Diagnosis route

Description: Method of diagnosis

This is a flat (unstructured) list which uses integers to rep
automatically with any research dataset that uses it.

Coding can be downloaded here as a tab-separated file.

6 Entries | 1 Data-Field

Coding	Meaning
-701	Self-diagnosis from symptoms
-702	Doctor diagnosis from symptoms
-703	By means of a blood test only
-704	By means of endoscopy only
-705	By means of a blood test and endoscopy
-818	Prefer not to answer

Εικόνα 14

1.5 Πρόσβαση στα δεδομένα του UK Biobank. Το Access Management System - AMS

Για να μπορέσει να υπάρχει πρόσβαση στα δεδομένα του UK Biobank θα πρέπει κατ' αρχάς να γίνει εγγραφή του ενδιαφερόμενου researcher στην σελίδα https://ams.ukbiobank.ac.uk/ams/ Στο mail που έχει δηλωθεί έρχεται verification link, πατάμε στον σύνδεσμο και ενεργοποιούμε την εγγραφή. Κάνουμε εισαγωγή με τα στοιχεία και πατάμε registration, με την συμπλήρωση των απαραίτητων στοιχείων ολοκληρώμνουμε την εγγραφή.

1.6 Είδη δεδομένων στο UK Biobank Data on the Research Analysis Platform και Interrelations

Η UK Biobank περιέχει δεδομένα από περίπου 500000 εθελοντές. Κάθε συμμετέχων έχει έναν μοναδικό επταψήφιο αναγνωριστικό κωδικό EID, ο οποίος είναι ένας αριθμός ανάμεσα στο 1000000 και το 6000000.

Τα διαθέσιμα δεδομένα είναι αποθηκευμένα στην UK Biobank σε διάφορα format τα οποία έχουν τέσσερις διαφορετικούς τρόπους πρόσβασης [5][18].

1.6.1 The Main Dataset: Είναι αρχείο table, όπου κάθε row είναι κάθε ξεχωριστός participant και όλα τα fields του καθε participant columns. Περιέχει όλες τις μετρήσεις και τις πληροφορίες που έχουν συλλεχθεί στα UKB assessment centers ή από online ερωτηματολόγια. Κάποιες πληροφορίες έχουν συλλεχθεί μέσω του NHS, προέρχονται από δεδομένα από bulk files (MRI Scans) και είναι διαθέσιμα μέσω του Data Portal [5](σελ 11).

1.6.2 The Data Portal. Εδώ είναι προσβάσιμα αρχεία νοσοκομείων, πρωτοβάθμια φροντίδα υγείας, Covid-19, PCR test, Olink & OMOP Data. Στα αρχεία αυτού του τύπου η πρόσβαση είναι εφικτή από την σελίδα Downloads του Showcase[19]. Πρόκειται για tabular data με περιπλοκη δομή, ώστε δεν χωρούν στο Main UKB Dataset [5](σελ 40). Ένας ασθενής μπορεί να έχει έναν πολύ μεγάλο αριθμό εισαγωγών στο νοσοκομείο, όπου η κάθε μια εισαγωγή αντιστοιχεί και στις συνοδευόμενες πληροφορίες, οι οποίες δεν μπορούν να διαταχθούν σε one-row-per-participant structure του Main Dataset. Στην περίπτωση αυτή το record repository χωρίζεται σε έναν αριθμό συνδεόμενων (interconnected) database tables, εκ των οποίων η κάθε μια καλύπτει ένα αριθμό διαφορετικών data areas. Πληροφορίες σχετικά με αυτά τα record tables βρίσκονται στο τμήμα resources του showcase, όπως φαίνεται στην εικόνα 13.

- **Category 2000 for hospital inpatient data.** Note that summary information about hospital diagnoses and procedures is available in a main dataset, but the detail of each admission in the record-level tables is not.

- **Category 100093 for death data.** Note that death data is (uniquely) also available in a main dataset.
- **Category 3000** for primary care (GP) data (covering approximately 45% of the cohort).
- **Field 40100** for COVID-19 PCR test results data.
- **Field 32040** for COVID-19 vaccination data. This field is restricted and is only available for research related to COVID-19.
- **Category 1839** for the Olink proteomics data.
- **Field 20142** for the OMOP Common Data Format (CDF) dataset tables.

Λίστα με το σύνολο των Data Portal Tables βρίσκονται στο Record Tables Catalogue https://biobank.ctsu.ox.ac.uk/crystal/docs.cgi?id=3 του Showcase όπως φαίνονται και στην εικόνα 13.

Record Tables Catalogue

Record tables are resources in which the data is presented as row-based information.

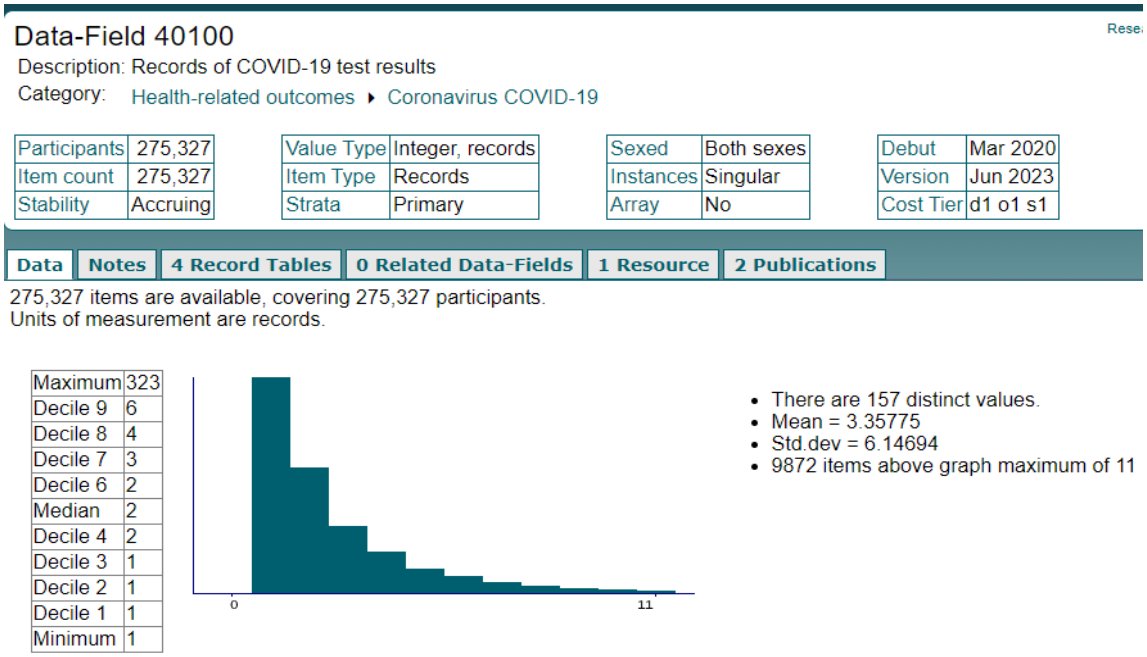
39 Record Tables

Table ID	Table	Title
1063	hesin	HES inpatient core dataset
1065	hesin_critical	HES inpatient critical care
1066	hesin_delivery	HES inpatient delivery
1067	hesin_diag	HES inpatient diagnoses
1068	hesin_maternity	HES inpatient maternity
1069	hesin_oper	HES inpatient operations
1070	hesin_psych	HES inpatient psychiatric
1060	gp_clinical	GP clinical events
1061	gp_registrations	GP registrations
1062	gp_scripts	GP prescriptions
1058	death	Death records
1059	death_cause	Death causes
1053	covid19_result_england	COVID-19 test results (England)
1054	covid19_result_scotland	COVID-19 test results (Scotland)
1055	covid19_result_wales	COVID-19 test results (Wales)
1051	covid19_misc	COVID-19 miscellaneous
1056	covid19_tpp_gp_clinical	GP clinical events (TPP, COVID19)
1057	covid19_tpp_gp_scripts	GP prescriptions (TPP, COVID19)
1049	covid19_emis_gp_clinical	GP clinical events (EMIS, COVID19)

Εικόνα 15

Πηγαίνοντας στο Field 40100 βλέπουμε (εικόνα 16) ότι περιέχει τέσσερα διαφορετικά Record Tables που σχετίζονται με καταγραφές Covid-19 και τα οποία μπορούμε να τα αντιστοιχίσουμε στις αντίστοιχες καταγραφές στην εικόνα 15. Επιπλέον όπως μπορούμε να δούμε στο resources στην εικόνα 18, επιπλέον πληροφορίες είναι διαθέσιμες σχετικά με το data field [20]. Οι πληροφορίες αυτές στην συγκεκριμένη περίπτωση προέρχονται από των άρθρο των

Jacob A. Et al. Dynamic linkage of COVID-19 test results between Public Health England's Second Generation Surveillance System and UK Biobank [21].



Data-Field 40100

Description: Records of COVID-19 test results

Category: Health-related outcomes ▸ Coronavirus COVID-19

Participants	275,327	Value Type	Integer, records	Sexed	Both sexes	Debut	Mar 2020
Item count	275,327	Item Type	Records	Instances	Singular	Version	Jun 2023
Stability	Accruing	Strata	Primary	Array	No	Cost Tier	d1 o1 s1

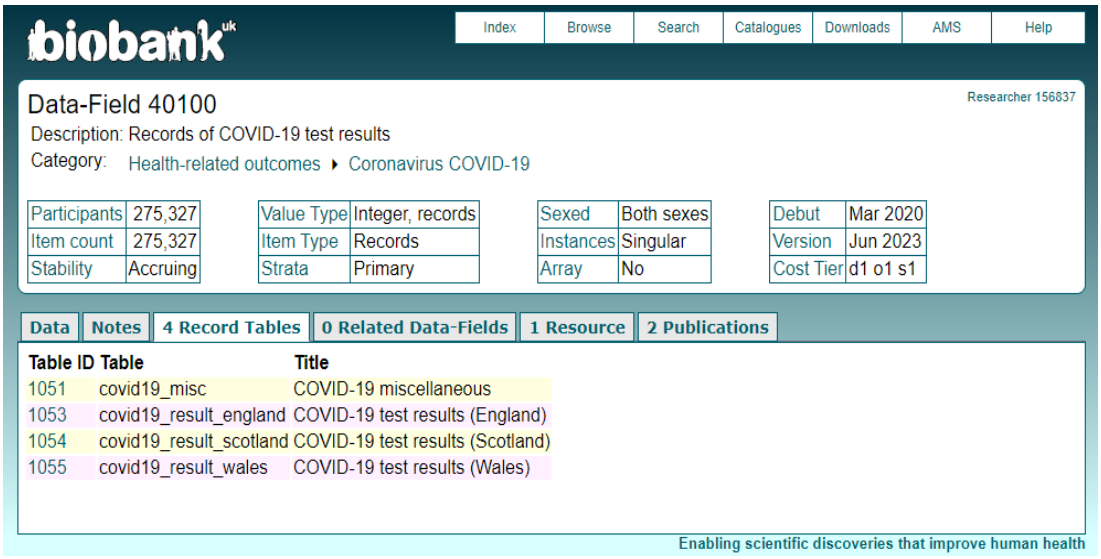
Data | Notes | 4 Record Tables | 0 Related Data-Fields | 1 Resource | 2 Publications

275,327 items are available, covering 275,327 participants.
Units of measurement are records.

Maximum	323
Decile 9	6
Decile 8	4
Decile 7	3
Decile 6	2
Median	2
Decile 4	2
Decile 3	1
Decile 2	1
Decile 1	1
Minimum	1

- There are 157 distinct values.
- Mean = 3.35775
- Std.dev = 6.14694
- 9872 items above graph maximum of 11

Εικόνα 16



biobank uk

Index | Browse | Search | Catalogues | Downloads | AMS | Help

Researcher 156837

Data-Field 40100

Description: Records of COVID-19 test results

Category: Health-related outcomes ▸ Coronavirus COVID-19

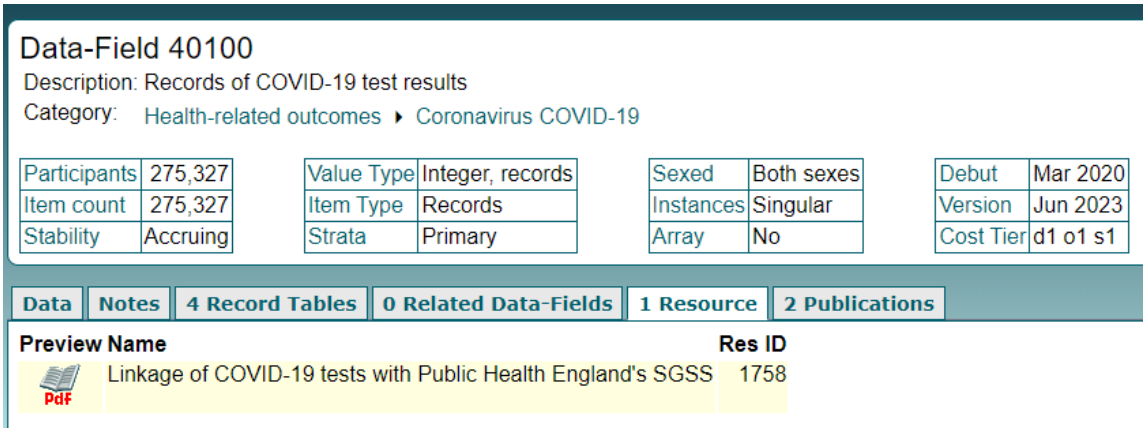
Participants	275,327	Value Type	Integer, records	Sexed	Both sexes	Debut	Mar 2020
Item count	275,327	Item Type	Records	Instances	Singular	Version	Jun 2023
Stability	Accruing	Strata	Primary	Array	No	Cost Tier	d1 o1 s1

Data | Notes | 4 Record Tables | 0 Related Data-Fields | 1 Resource | 2 Publications

Table ID	Table	Title
1051	covid19_misc	COVID-19 miscellaneous
1053	covid19_result_england	COVID-19 test results (England)
1054	covid19_result_scotland	COVID-19 test results (Scotland)
1055	covid19_result_wales	COVID-19 test results (Wales)

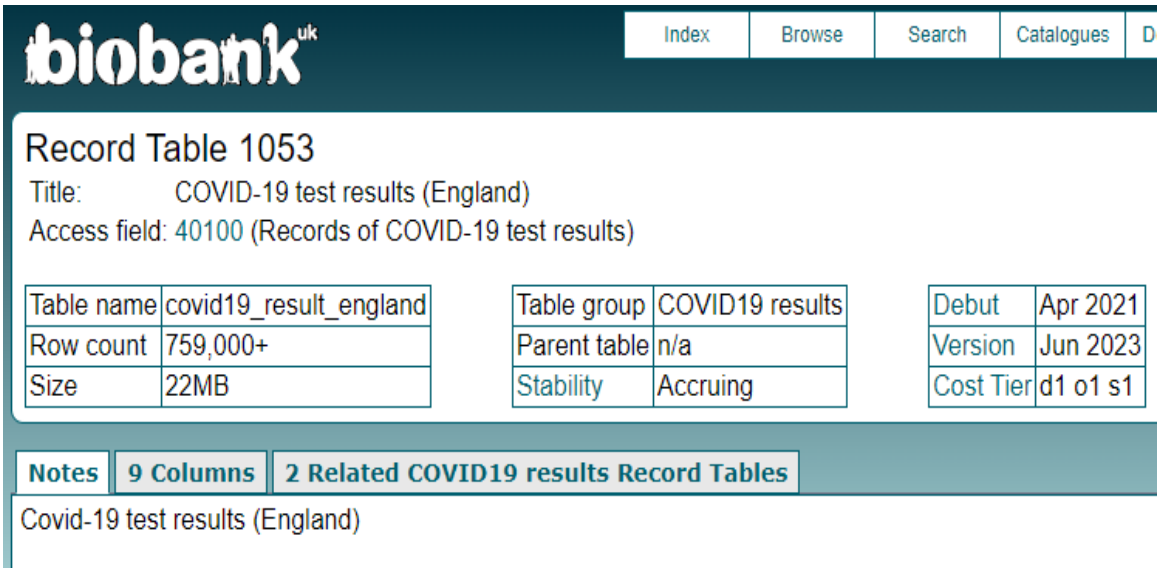
Enabling scientific discoveries that improve human health

Εικόνα 17

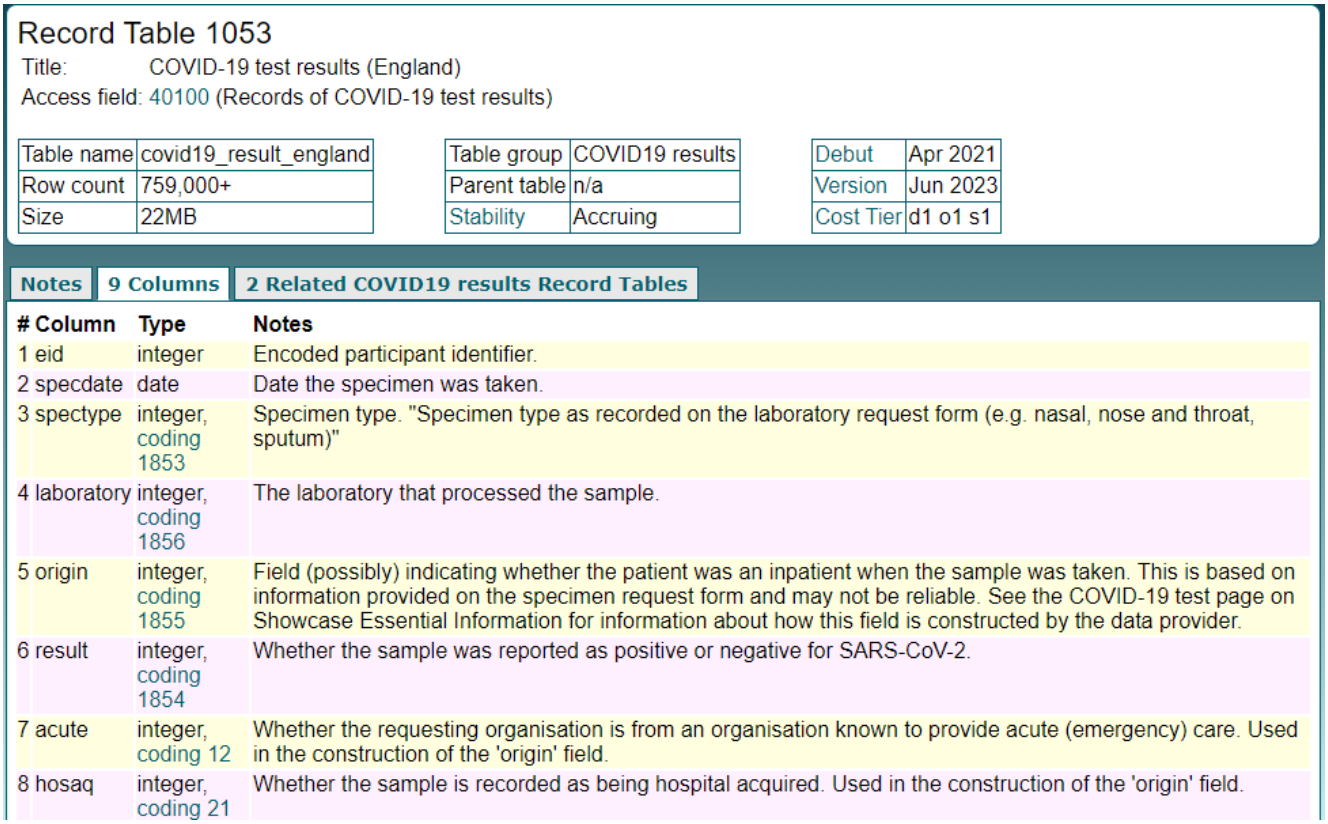


Εικόνα 18

Κάνοντας κλικ στο Table Id παίρνουμε την λίστα των columns του συγκεκριμένου table. Στην περίπτωση που παρουσιάζεται, στην εικόνα 17 εμφανίζεται το Record table 1053, μαζί με τις συνοδευόμενες πληροφορίες για το data type και τις σημειώσεις του field, οι οποίες εμφανίζονται αν πατήσουμε στο ID (εικόνα 17). Επίσης στην εικόνα 18 βλέπουμε τα 9 διαφορετικά variables που καταγράφονται από τα COVID-19 test results in England [22].



Εικόνα 17



Record Table 1053

Title: COVID-19 test results (England)

Access field: 40100 (Records of COVID-19 test results)

Table name	covid19_result_england	Table group	COVID19 results	Debut	Apr 2021
Row count	759,000+	Parent table	n/a	Version	Jun 2023
Size	22MB	Stability	Accruing	Cost Tier	d1 o1 s1

Notes | 9 Columns | 2 Related COVID19 results Record Tables

#	Column	Type	Notes
1	eid	integer	Encoded participant identifier.
2	specdate	date	Date the specimen was taken.
3	spectype	integer, coding 1853	Specimen type. "Specimen type as recorded on the laboratory request form (e.g. nasal, nose and throat, sputum)"
4	laboratory	integer, coding 1856	The laboratory that processed the sample.
5	origin	integer, coding 1855	Field (possibly) indicating whether the patient was an inpatient when the sample was taken. This is based on information provided on the specimen request form and may not be reliable. See the COVID-19 test page on Showcase Essential Information for information about how this field is constructed by the data provider.
6	result	integer, coding 1854	Whether the sample was reported as positive or negative for SARS-CoV-2.
7	acute	integer, coding 12	Whether the requesting organisation is from an organisation known to provide acute (emergency) care. Used in the construction of the 'origin' field.
8	hosaq	integer, coding 21	Whether the sample is recorded as being hospital acquired. Used in the construction of the 'origin' field.

Εικόνα 18

Επιπλέον πληροφορίες σχετικά με τα Tables & Fields είναι ανακτήσιμες στο Schema 17 [23] & Schema 18 [24] του Showcase.

Τα δεδομένα στο Data Table μπορούν με την κατάλληλη διαδικασία να αποθηκευτούν ως σύνολο τοπικά (που είναι αρκετά χρονοβόρα διαδιασία) ή με SQL statements να γίνει ερευνα των αναγκαίων πεδίων και να αποθηκευτούν / χρησιμοποιηθούν μόνο αυτά [5].

1.6.3 The Download Utilities. Η βιβλιοθήκη UKB περιέχει διάφορους τύπους δεδομένων χωρίς πίνακα (non – tabular), όπως εικόνες, δεδομένα γονιδιώματος και επιστρεφόμενα σύνολα δεδομένων/ returned datasets από τους ερευνητές μετά από ερευνητικές εργασίες, τα οποία δεν είναι κατάλληλα να συμπεριληφθούν στον κύριο κορμό των δεδομένων [5]. Τα δεδομένα αυτά μπορούν να αποθηκευθούν τοπικά με την βοήθεια των utilities ukbfetch, gfetch & ukblink, που χρησιμοποιούνται για το κατέβασμα μεγάλων ή περίπλοκων αρχείων (non tabular data), όπως

genotyping array, απεικονιστικές φωτογραφίες (MRI, ακτινογραφίες), genomics data, returned datasets και fitness data. Αυτά τα utilities υπάρχουν διαθέσιμα στην ιστοσελίδα του UKB [25].

Απαραίτητο για να συνδεθεί κανείς στα UKB remote Repositories θα χρειαστεί να αποθηκεύσει το keyfile από το notification email του project στο ίδιο directory με το download utility.

1.6.4 The Research Analysis Platform (RAP) Η οργάνωση των δεδομένων [26]. Η UK Biobank περιέχει δεδομένα από περίπου 500,000 εθελοντές. Καθε συμμετέχων έχει έναν μοναδικό επταψήφιο αναγνωριστικό κωδικό EID, ο οποίος είναι ένας αριθμός ανάμεσα στο 1000000 και το 6000000. Τα δεδομένα στην Research Analysis Platform (RAP) είναι αποθηκευμένα ως Bulk Files ή Tabular Data.

1.6.4.1 Bulk Files. Τα Bulk Files περιέχουν subolders και έχουν την ακόλουθη δομή:

- Για καθε Bulk Field Category υπάρχει ένα subfolder. Για παράδειγμα whole genome CRAM files είναι αποθηκευμένα στο subfolder **Whole genome sequences**. Αυτά τα categories ορίζονται από την UK Biobank, ειδικά για την RAP.
- Εντός της κάθε Category Subfolder υπάρχει ένας νέος υποφάκελος (SubfolderII) για κάθε bulk Field. Ο subfolder **Whole genome sequences** θα περιέχει εσωτερικά και άλλους υποφακέλους π.χ. Whole genome CRAM files σχετικά με αυτό το field.
- Στην Συνέχεια εντός του κάθε field subfolder υπάρχουν αποθηκευμένα τα δεδομένα -σε ξεχωριστά files - που αφορούν στον κάθε ξεχωριστό participant με κωδικούς απο το prefix του participant συνήθως ανάμεσα στο “10” & το “60”.

Η RAP χρησιμοποιεί την ακόλουθη ονοματοδοσία για τα bulk data files:

Files που περιέχουν δεδομένα για κάποιον συμμετέχοντα ονοματίζονται κατά τον ακόλουθο τρόπο:

<EID>_<FIELD-ID>_<INSTANCE-ID>_<ARRAY-ID>.<SUFFIX>

Για παράδειγμα ένα whole genome CRAM files (field ID #23193) ονοματίζονται: **<EID>_23193_0_0.cram**

Κάποιες εξαιρέσεις στον κανόνα αυτό υπάρχουν όταν ένα field είναι συμπληρωματικό στο κύριο field, όπως ένα CRAI index που συνοδεύει έναν CRAM file, ή ένα TBI index που συνοδεύει ένα VCF file. Στην περίπτωση αυτή το σύστημα χρησιμοποιεί ως prefix αυτό του κυρίου field. Για παράδειγμα ένα whole genome CRAM indices (field ID #23194) ονοματίζεται ως εξής:

<EID>_23193_0_0.cram.crai

Files που περιέχουν data από ένα σύνολο/ cohort συμμετεχόντων (όπως PLINK, BGEN ή pVCF files) ονοματίζονται ως ακολούθως:

ukb<FIELD-ID>_c<CHROM>_b<BLOCK>_v<VERSION>.<SUFFIX>

Όπου **<CHROM>** αντιπροσωπεύει το αντίστοιχο χρωμόσωμα (όπως "1", "2" ή "X"), **<BLOCK>** αντιστοιχεί σε ένα index (αρχίζοντας από το "0") για datasets που έχουν χωριστεί σε πολλαπλά κομμάτια και **<VERSION>** αντιστοιχή στην dataset version του UK Biobank.

1.6.4.1.A Ιδιότητες των Bulk Data Files. Όταν Bulk Data Files διατίθονται απο την RAP για ερευνητικούς σκοπούς αντιστοιχίζονται σε αυτούς κάποιες αρχικές ιδιότητες των φακέλων. Στην εικόνα 19 μπορούμε να παρατηρήσουμε τις αντιστοιχίσεις των Key Value Pairs (string).

Key	Value	Which files have this property?
eid	The corresponding participant EID	Files that correspond to a single participant.
field_id	The corresponding data-field id.	All files.
instance_id	The corresponding instance id (typically a visit to an assessment centre).	Files that correspond to data-fields with multiple instances.
array_id	The corresponding array index.	Files that correspond to array data-fields.
resource_id	The corresponding UK Biobank resource id.	Auxiliary files to a resource on the UK Biobank Showcase.

Εικόνα 19

Τεχνικές οδηγίες για την χρήση των Bulk Files είναι διαθέσιμες στο documentation toy UKB [27].

1.6.4.2 Tabular Data: Η RAP αποθηκεύει Tabular Data-Fields και Linked Health Data σε SQL Database [26]. Αυτή η Database βρίσκεται στο root folder του project και συνήθως ακολουθεί το ακόλουθο μοτίβο:

app<APPLICATION-ID>_<CREATION-TIME> (e.g. App12345_20210101123456)

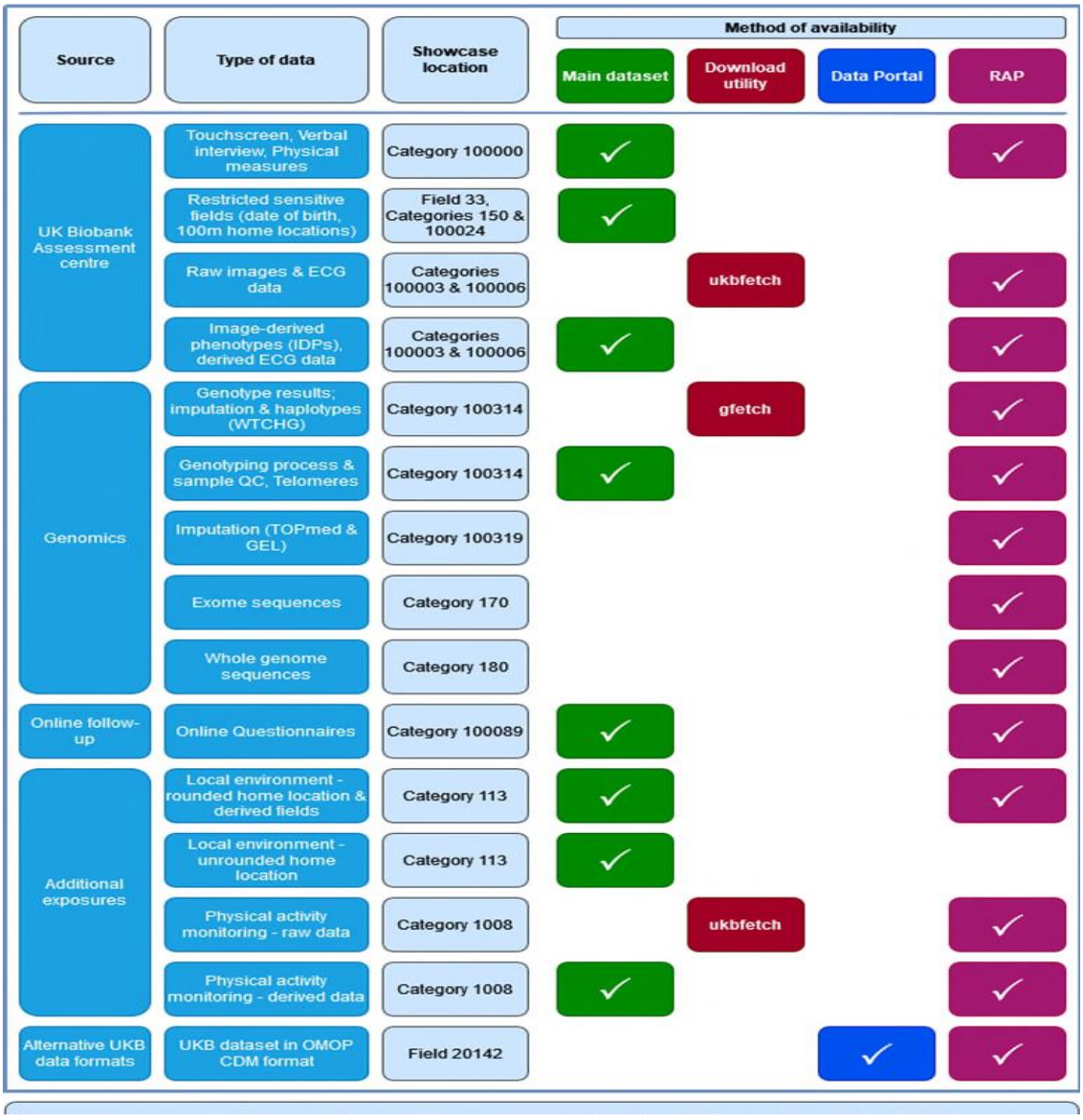
Στον ίδιο φάκελο βρίσκεται και ένα συσχετιζόμενο Dataset (Apollo Dataset) το οποίο λαμβάνει το όνομα του από το την Database με το .dataset στο τέλος και συνδιάζει low-level SQL columns με field-level metadata from the UK Biobank Showcase. Ένα Apollo Dataset είναι ένα DNA Nexus Record τύπου Dataset που περιέχει data και metadata, και συσχετίζει μεταξύ logical data structure (phenotypes, genotypes, etc.) και physical layout των υποκείμενων database(s) και metadata lookups. Τα Apollo Datasets, τα οποία όπως αναφέρθηκε περιέχουν Tabular data-fields

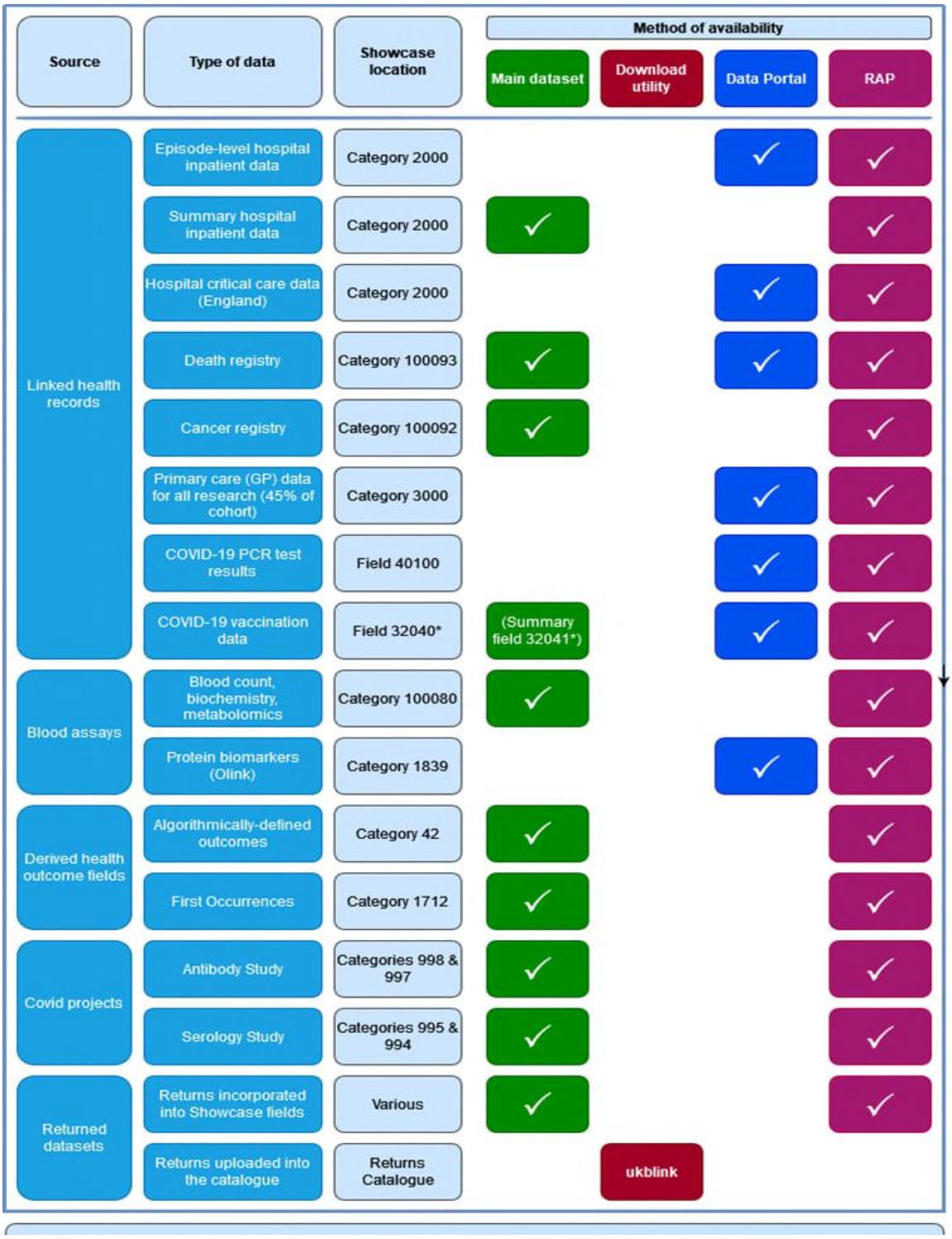
και Linked Health Data είναι δυνατόν να προσπελαστούν είτε μέσω της εφαρμογής Cohort Browser του DNAnexus, είτε μέσω του Sparkl στο JupyterLab [28][29].

Περισσότερες οδηγίες σχετικά με την αναζήτηση Phenotyping Data Files [30].

Περισσότερες οδηγίες σχετικά με την αναζήτηση Tabular Data [31].

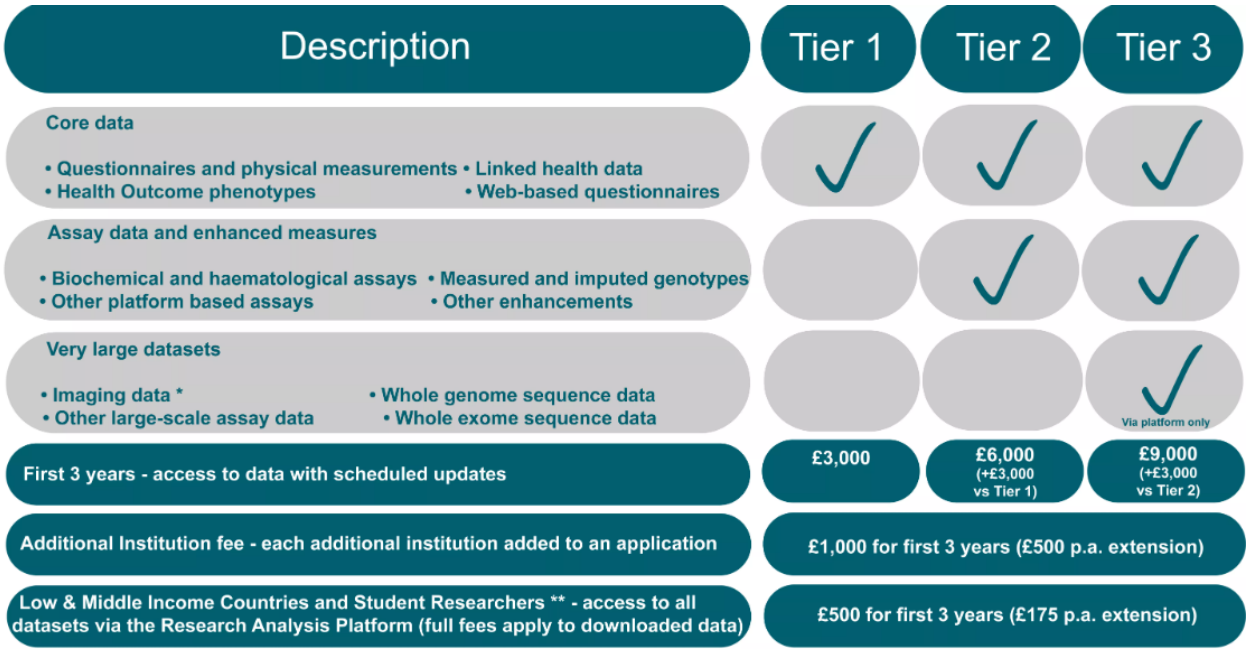
1.5 Συνοπτική απεικόνιση τρόπων ανάκτησης δεδομένων από UK Biobank [5]




Source
Type of data
Showcase location
Method of availability
Main dataset
Download utility
Data Portal
RAP
Linked health records
Episode-level hospital inpatient data
Category 2000
Summary hospital inpatient data
Category 2000
Hospital critical care data (England)
Category 2000
Death registry
Category 100093
Cancer registry
Category 100092
Primary care (GP) data for all research (45% of cohort)
Category 3000
COVID-19 PCR test results
Field 40100
COVID-19 vaccination data
Field 32040*
(Summary field 32041*)
Blood assays
Blood count, biochemistry, metabolomics
Category 100080
Protein biomarkers (Olink)
Category 1839
Derived health outcome fields
Algorithmically-defined outcomes
Category 42
First Occurrences
Category 1712
Covid projects
Antibody Study
Categories 998 & 997
Serology Study
Categories 995 & 994
Returned datasets
Returns incorporated into Showcase fields
Various
Returns uploaded into the catalogue
Returns Catalogue
ukblink

Εικόνα 20

1.7 Κόστος πρόσβασης δεδομένων [32]



Εικόνα 21

Costs exclude VAT.

* The imaging data will be available for download with the Tier 3 payment: but at no additional cost over and above the Tier 2 payment via the Research Analysis Platform in 2022. Imaging derived phenotypes will be available as part of Tier 1.
** Applications from Student Researchers must be for the sole purpose of performing a postgraduate student project (e.g. MSc or PhD or equivalent), submitted by the student or their supervisor.
N.B. First 3 years - access to data with scheduled updates (Duration Extensions are pro-rated, for example: Tier 3 for 2 years extension = £6,000).

Υπάρχει η δυνατότητα χρηματοδότησης για early career researchers μέσω του UK Biobank Platform Credits Programme σύμφωνα με τα ακόλουθα κριτήρια:

UK Biobank defines early career researchers as "*an individual within an academic institution within four years of the award of their PhD or equivalent professional training, or within four years of starting their first employment position (full-time or part-time), excluding career breaks)".* Early career researchers also include those bona fide students eligible for reduced Access fees.

Κεφάλαιο 2

National Inpatient Sample (NIS) Documentation

2.1 Τα δεδομένα του National Inpatient Sample (NIS)

Το National Inpatient Sample (NIS) είναι η μεγαλύτερη ενδονοσοκομειακή βάση υγείας στις ΗΠΑ. Έχει δημιουργηθεί ως τμήμα του Healthcare Cost and Utilization Project (HCUP)και περιέχει διαπολιτειακά δεδομένα υγείας (multistate health data system). Το NIS δημιουργήθηκε αρχικά από τις βάσεις δεδομένων των νοσοκομείων community hospitals, που είχαν ως σκοπό την καταγραφή των διαχειριστικών διαδικασιών και του κόστους, εν τούτοις παρέχει σημαντικές πληροφορίες για τα αποτελέσματα της περίθαλψης των ασθενών σε εθνικό επίπεδο[44]. Ο ορισμός του community hospital ακολουθεί αυτόν της Αμερικανικής Ένωσης Νοσοκομείων American Hospital Assosiation- AHA): μη ομοσπονδιακό (π.χ. μη στρατιωτικό, όχι νοσοκομείο βετεράνων, ή Indian Health Service), βραχίας νοσηλείας γενικό ή εξειδικευμένο νοσοκομείο, όπως του μαιευτικού-γυναικολογικού, ωτορινολαρυγγολογικό, βραχυπρόθεσμης αποκατάστασης, ορθοπεδικό, και παιδιατρικό. Εξαιρούνται τα νοσοκομεία μακροχρόνιας περίθαλψης, τα ψυχιατρικά νοσοκομεία, μονάδες θεραπείας αλκοολισμού/φαρμακευτικής εξάρτησης, και νοσοκομειακές μονάδες εντός ιδρυμάτων (πχ φυλακές) [44].

Η διαδικασία δημιουργίας του αρχείου είναι η ακόλουθη. Για κάθε ασθενή που φτάνει σε κάποιο νοσοκομείο (Federal or Non Federal) καταγράφεται το προσωπικό του ιατρικό αρχείο. Μετά την θεραπεία και το εξιτήριο, δημιουργείται το billing record/discharge record, καταγραφονται οι ασθένειες βάσει του International classification of Diseases ICD-10-CD καθώς και οι διαδικασίες που ακολουθήθηκαν με κωδικοποίση Procedure Categories - PCS. Το τμήμα

κοστολόγησης συντάσει τον λογαριασμό χρησιμοποιώντας τους κωδικούς που καταγράφηκαν, ενώ αποτυπώνει και δημογραφικά δεδομένα όπως ηλικία και φύλο. Τα δεδομένα μεταφέροντα σε State Level Data Organizations. Όσα από τα Data Organizations συμμετέχουν στο HCUP στέλνουν τα δεδομένα τους στο Agency for Healthcare Research and Quality (AHRQ) [34].

Έχοντας ως βασή το 2012, το NIS δημιουργεί την βάση δεδομένων λαμβάνοντας δείγματα με διαστρωματοποιημένο τρόπο (stratified), από όλες τα εξιτήρια (discharges), τα οποία έχουν συμπεριληφθεί στο SID (State Inpatient Databases). Τα νοσοκομεία ταξινομούνται ανά περιοχή (New England, Middle Atlantic, Mountain, Pacific κοκ), τοποθεσία (αστική ή αγροτική), αριθμό κλινών (μικρά, μεσαία, μεγάλα), ιδιοκτησία (Government- non federal/Public, Private non – Profit/ Voluntary, Private investor owned/propriatery) καθώς και αν έχουν ταυτόχρονα και εκπαιδευτική δράση. Κάθε τέτοια κατηγοριοποίηση αποτελεί ένα strata.

Το 20% των εξιτηρίων από το κάθε στρώμα που έχει καταγραφεί στο SID, χρησιμοποιείται ως δείγμα. Τα δείγματα αυτά έχουν αντίστοιχη βαρύτητα (discharge sample weights), ανάλογα με την ταξινόμηση τους. Όσον αφορά στις ασθένεις που καταγράφονται το HCUP, γίνονται κάποιες προσαρμογές σχετικά με την σοβαρότητα τους με διάφορα HCUP Tools and Software. Επιπροσθέτως χρησιμοποιούνται και finite population correction factors (fpc) για τον τελικό υπολογισμό της διακύμανσης του πληθυσμού, καθώς θεωρείται πως τα δειγματοληπτικά εξιτήρια (sample discharges) εντός του ίδιου νοσοκομείου είανι περισσότερο ομογενοποιημένα σε σχέση με τα δειγματοληπτικά εξιτήρια (sample discharges) διαφορετικών νοσοκομίων [35].

Είναι σημαντικό να τονιστεί η διαφορά στην καταγραφή μεταξύ του UKB & του NIS. Στο πρώτο είναι καταγεγραμένα όλα τα δεδομένα του κάθε ασθενούς από το 2006, ανεξαρτήτως αν ο ασθενής νοσηλεύτηκε ενδονοσοκομειακά, ως εξωτερικός ασθενής ή και καθόλου. Στα δεδομένα αυτά ένα πολύ σημαντικό κομμάτι είναι το κομμάτι του γονιδιώματος του κάθε συμμετέχοντος.

Αντίθετα στο στο NIS καταγράφονται σε πολιτειακό επίπεδο μόνο οι περιπτώσεις για τις οποίες εκδίδεται εξιτήριο μετά από ενδονοσοκομειακή περίθαλψη. Τα στοιχεία που καταγράφονται έχουν να κάνουν και με το κόστος της θεραπείας , φαρμακευτική και θεραπευτική αγωγή,οι διαδικασίες, καθώς και δημογραφικά δεδομένα ενδονοσοκομειακών ασθενών που έλαβαν. Η Πλήρης λίστα των δεδομένων που καταγράφονται βρίσκεται στον

ακόλουθο σύνδεσμο [33]. Στην συνέχεια με σταθμισμένο δείγμα γίνεται αναγωγή στο σύνολο του πληθυσμού.

2.1.1 Συνοπτικές διαφορές UKB & NIS

Συνοπτικά το UKB είναι ένα σύστημα που πρωταρχικά δημιουργήθηκε με σκοπό να καταγράφει όσο το δυνατόν πληρέστερα όλα τα δεδομένα υγείας, αλλά και δημογραφικά και lifestyle δεδομένα των συμμετεχόντων, ενώ το NIS έχει ως αφετηρία τον υπολογισμό του κόστους του ασθενούς κατά την παραμονή του στο νοσοκομείο (community hospital), καταγράφοντας δεδομένα υγείας και διαδικασιών που λάβαν χώρα ενδονοσοκομειάκα. Με το εξιτήριο του ασθενούς, που δίνεται μετά από ενδονοσοκομειακή περίθαλψη, τα δεδομένα αποστέλονται στο NIS και με βάση δείγμα από όλα τα εξιτήρια, μέσω διαστρωματοποιημένης συλλογής και ανάλυσης, υπολογίζονται τα δεδομένα σε εθνικό επόπεδο.

Είναι εμφανές πως ως βάση δεδομένων υγείας το UKB είναι πληρέστερο, καθώς εξ αρχής δημιουργήθηκε για να καταγράφει ιατρικά και βιοιατρικά δεδομένα, ενώ η στόχευση του NIS είναι πρωτίστως ο υπολογισμός του κόστους, μέσω της αποτύπωσης των διαχειριστικών διαδικασιών. Εν τούτοις και από το NIS μπορούν να αντληθούν σημαντικές πληροφορίες για την περίθαλψη των ασθενών σε εθνικό επίπεδο[44].

2.2 Δομή αρχείων NIS

Εκτός από το έτος 2015, το οποίο χωρίστηκε σε δύο μέρη λόγω της αλλαγής από την κωδικοποίηση ICD-9-CM σε ICD-10-CM/PCS, το NIS παρέχει ένα ετήσιο αρχείο για κάθε ημερολογιακό έτος. Υπάρχουν τρεις φάκελοι σε επίπεδο εξιτηρίου (discharge-level files) και ένας φάκελος σε επίπεδο νοσοκομείου (hospital-level file).

2.2.1 Discharge-level files.

2.2.1.A Core Files. Το Core File είναι ένα μεμονωμένο αρχείο που περιέχει στοιχεία δεδομένων που χρησιμοποιούνται συνήθως (π.χ. ηλικία, αναμενόμενος κύριος πληρωτής, κατάσταση υγείας κατά το εξιτήριο, κωδικοί ICD-10-CM/PCS, συνολικές χρεώσεις). Αυτό το αρχείο είναι διαθέσιμο για όλα τα έτη του NIS. Η σύνδεση μεταξύ των discharge-level files

προ του έτους 2012, γίνεται με το μοναδικό αναγνωριστικό εγγραφής - unique record identifier HCUP (KEY) παρείχε τη σύνδεση μεταξύ των αρχείων σε επίπεδο εκφόρτισης.

2.2.1.B Severity Files. Το Severity File είναι ένα μεμονωμένο αρχείο που περιέχει πρόσθετα στοιχεία δεδομένων που βοηθούν στον προσδιορισμό της σοβαρότητας της κατάστασης για μια συγκεκριμένη εκκένωση. Αρχεία αυτού του τύπου είναι διαθέσιμα από το NIS του 2002.

2.2.1.C Diagnosis and Procedure Groups Files. Το **Diagnosis and Procedure Groups File** είναι ένα ενιαίο αρχείο που περιέχει πρόσθετες πληροφορίες για τις διαγνώσεις ICD-10-CM και τις διαδικασίες ICD-10-PCS και δημιουργείται από το Agency for Healthcare Research and Quality (AHRQ) software tools. Αυτά τα αρχεία είναι διαθέσιμα από το NIS του 2005. Για τα έτη δεδομένων 2016-2017, αυτό το αρχείο δεν ήταν διαθέσιμο στα NIS. Δεδομένα που εξήγχθησαν από τα εργαλεία λογισμικού ICD-10-CM/PCS AHRQ δεν συμπεριλήφθηκαν στα NIS επειδή βρίσκονταν ακόμη σε δοκιμαστική φάση ανάπτυξης.

Ξεκινώντας από το έτος δεδομένων 2018, αυτό το αρχείο περιλαμβάνει δεδομένα που προέρχονται από το Clinical Classifications Software Refined (CCSR) για διαγνώσεις ICD-10-CM. Ξεκινώντας με το έτος δεδομένων 2019, δεδομένα που προέρχονται από το Elixhauser Comorbidity Software Refined for ICD-10-CM, το CCSR για τις διαδικασίες ICD-10-PCS και τις Κατηγορίες Διαδικασιών Refined για το ICD-10-CM είναι επίσης διαθέσιμα σε αυτό το αρχείο.

2.2.2Hospital-level files. Το Hospital-level file είναι ένα ενιαίο αρχείο που περιέχει πληροφορίες για τα χαρακτηριστικά του νοσοκομείου. Αυτό το αρχείο είναι διαθέσιμο σε όλα τα έτη του NIS.

Σύνδεση μεταξύ του Βασικού Αρχείου Εσωτερικού Νοσοκομείου και του Αρχείου Νοσοκομείου. Πριν από το 2012, το αναγνωριστικό νοσοκομείου HCUP - hospital identifier (HOSPID) παρείχε τη σύνδεση μεταξύ του NIS Inpatient Core File και του Hospital File. Από το ετός 2012, ο αριθμός νοσοκομείου NIS - hospital number (HOSP_NIS) παρέχει τη σύνδεση μεταξύ του NIS Inpatient Core File και του Hospital File. Οι τιμές HOSP_NIS δημιουργούνται εκ νέου κάθε χρόνο, επομένως δεν μπορούν να χρησιμοποιηθούν για τη σύνδεση νοσοκομείων μεταξύ των ετών [36].

Κεφάλαιο 3

Ανάλυση δεδομένων NIS με κλασσικές μεθόδους στατιστικής

3.1 Σκοπός της έρευνας

Στο πρώτο σκέλος της έρευνας μας, σκοπός είναι μελέτη των δεδομένων που προέρχονται από το NIS για το έτος 2013. Εξετάζονται πιθανές συσχετίσεις μεταβλητών με τον συνολικό κόστος νοσηλείας καθώς και το ποσοστό θνησιμότητας. Η ανάλυση πραγματοποιείται με κλασσική μέθοδο στατιστικής αναλυσης και το στατιστικό πακέτο SPSS.

Στο δύτερο σκέλος δημιουργούμε, χρησιμοποιώντας τον XGBoost classifier, ένα επεξηγηματικό Machine Learning μοντέλο, με προβλεπτική ισχύ για το τελικό κόστος του ασθενή καθώς και με την πιθανότητα επιβίωσης. Τα δεδομένα μας αφορούν ενδονοσοκομειακούς ασθενείς, όπως αυτοί έχουν καταγραφεί στο NIS Inpatient Core File.

3.2 Δεδομένα

Για την έρευνα χρησιμοποιήθηκαν δεδομένα το έτους 2013. Αποτελούνται από περίπου 7*106 εγγραφές με 141 ξεχωριστά variables. Οι μεταβλητές αυτές περιλαμβάνουν δημογραφικά δεδομένα, οικονομικά στοιχεία, δεδομένα σχετικά με την διαστωμάτωση (strata) του νοσοκομείου, την κατάσταση του ασθενή, τις ασθένειες όπως καταχωρούνται στο γενικευμένο CCS (Clinical Classifications Software) του NIS, τις διαδικασίες που ακολουθήθηκαν ενδονοσοκομειακά και πάλι βάσει του Single-Level CCS – Procedures του NIS.

3.3 Στατιστική Ανάλυση SPSS.

Μέθοδος: Με σκοπό να μελετήσω πιο ομαδοποιημένα τα δεδομένα, δημιουργήσα κατ' αρχάς έντεκα ξεχωριστές κατηγορίες, χρησιμοποιώντας ως βάση την κατηγοριοποίηση του CCS (Clinical Classifications Software) του NIS. Επιλέχθησαν 11 ξεχωριστές κατηγορίες ασθενειών με

σκοπό να μελετηθεί κατά πόσο υπάρχει επεξηγηματική δυνατότητα σχετικά με το κόστος, αλλά και της πιθανότητας θανάτου

Το Clinical Classifications Software (CCS) είναι ένα εργαλείο για τη ομαδοποίηση διαγνώσεων καταστάσεων υγείας και διαδικασιών, το οποίο αναπτύχθηκε από το Agency for Healthcare Research and Quality (AHRQ). Το CCS προσφέρει την δυνατότητα ομαδοποίησης ασθενειών και διαδικασιών από ένα πλήθος πιο πολύπλοκων κωδικών. Αυτή η ομαδοποίηση διευκολύνει την γρήγορη κατανόηση των μοτίβων των διαγνώσεων και των κλινικών πρακτικών, ώστε να διευκολύνει την ανάλυση του κόστους, των διαδικασιών σε σχέση με το τελικό αποτέλεσμα της θεραπείας. Το CCS συνενώνει πιο πολύπλοκους και ιεραρχικά ταξινομημένους κωδικούς διάγνωσης και κλινικών διαδικασιών κατά τη Διεθνή Ταξινόμηση Ασθενιών (International Classification of Diseases- IDC-9-CM), η οποία περιέχει περισσότερους από 14.000 κωδικούς διάγνωσης και 3.900 κωδικούς κλινικών διαδικασίων [37].

Οι κατηγορίες που δημιουργήθηκαν είναι οι ακόλουθες:

1. Cancer
2. Cardio & Circulatory
3. Pneumo
4. Gastro
5. Renal
6. Genital
7. Pregnancy
8. Bones_Joints_Chronic
9. Congenital
10. Myoskeletal_Acute
11. Emergency

Στην κατηγορία εμφανίζεται κάθε ασθενής με κωδικό 1, αν έχει διαγνωσθεί έστω και μία φορά με πάθηση που να εμπίπτει στην συγκεκριμένη κατηγορία, και με κωδικό 0 εάν δεν έχει

ποτέ διαγνωστεί με αντίστοιχη πάθηση. Το μάξιμουμ των πιθανών διαγνώσεων είναι εικοσιπέντε. Ο κώδικας που χρησιμοποιηθηκε για την ομαδοποίηση είναι ο ακόλουθος:

```
1   * Encoding: UTF-8.
2   * Create a new variable named Cancer.
3   COMPUTE Cancer = 0.
4
5   * Loop through each of the 25 columns and check the conditions.
6   DO REPEAT var = DXCCS1 TO DXCCS25.
7     IF (Cancer = 0 AND var > 10 AND var < 45) Cancer = 1.
8   END REPEAT.
9
10  * Display the values of the new variable.
11  EXECUTE.
12
13  * (Optional) Label the new variable.
14  VARIABLE LABELS Cancer 'Cancer(1 if any of DXCCS1 to DXCCS25 is between 10 and 45, else 0)'.
15
16  * (Optional) Display variable labels.
17  EXECUTE.
18
```

Εικόνα 22

3.3.1 Descriptive Statistics on Categories

Παρακάτω εμφανίζονται οι πίνακες με τα Frequency Tables για τις 11 αυτές κατηγορίες:

Cancer(1 if any of DXCCS1 to DXCCS25 is between 10 and 45, else 0)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	6136167	86,2	86,2	86,2
	1	983396	13,8	13,8	100,0
	Total	7119563	100,0	100,0	

Εικόνα 23

Cardio & Circulatory (1 if any of DXCCS1 to DXCCS25 is between 95 and 122, else 0)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	3087441	43,4	43,4	43,4
	1	4032122	56,6	56,6	100,0
	Total	7119563	100,0	100,0	

Εικόνα 24

Pneumo(1 if any of DXCCS1 to DXCCS25 is between 121 and 135, else 0)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	3444852	48,4	48,4	48,4
	1	3674711	51,6	51,6	100,0
	Total	7119563	100,0	100,0	

Εικόνα 25

Gastro(1 if any of DXCCS1 to DXCCS25 is between 137 and 156, else 0)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	4629102	65,0	65,0	65,0
	1	2490461	35,0	35,0	100,0
	Total	7119563	100,0	100,0	

Εικόνα 26

Renal(1 if any of DXCCS1 to DXCCS25 is between 155 and 164, else 0)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	5156516	72,4	72,4	72,4
	1	1963047	27,6	27,6	100,0
	Total	7119563	100,0	100,0	

Εικόνα 27

Genital(1 if any of DXCCS1 to DXCCS25 is between 164 and 177, else 0)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	6830071	95,9	95,9	95,9
	1	289492	4,1	4,1	100,0
	Total	7119563	100,0	100,0	

Εικόνα 28

Pregnancy(1 if any of DXCCS1 to DXCCS25 is between 175 and 197, else 0)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	6269286	88,1	88,1	88,1
	1	850277	11,9	11,9	100,0
	Total	7119563	100,0	100,0	

Εικόνα 29

Bones_Joints_Chronic(1 if any of DXCCS1 to DXCCS25 is between 200 and 213, else 0)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	5207539	73,1	73,1	73,1
	1	1912024	26,9	26,9	100,0
	Total	7119563	100,0	100,0	

Εικόνα 30

Congenital(1 if any of DXCCS1 to DXCCS25 is between 212 and 225, else 0)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	6179269	86,8	86,8	86,8
	1	940294	13,2	13,2	100,0
	Total	7119563	100,0	100,0	

Εικόνα 31

Myoskeletal_Acute(1 if any of DXCCS1 to DXCCS25 is between 224 and 233, else 0)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	6801257	95,5	95,5	95,5
	1	318306	4,5	4,5	100,0
	Total	7119563	100,0	100,0	

Εικόνα 32

Emergency(1 if any of DXCCS1 to DXCCS25 is between 232 and 254, else 0)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	5079166	71,3	71,3	71,3
	1	2040397	28,7	28,7	100,0
	Total	7119563	100,0	100,0	

Εικόνα 33

Ενδεικτικά αναφέρουμε πως το 51,6% των ασθενών για το έτος 2013 έχει εμφανίσει πνευμονολογικά προβλήματα, ενώ το 56,6% έχει εμφανίσει καρδιολογικά και κυκλοφορικά προβλήματα.

Επιπλέον εξετάσαμε την πιθανότητα συνοσηρότητας, εξετάζοντας τα περιγραφικά στατιστικά της μεταβλητής Number of diagnoses on this record. Τα περιγραφικά της στατιστικά παρατίθενται παρακάτω.

Statistics

Number of diagnoses on this record

N	Valid	7119563
	Missing	0
Mean		9,32
Median		8,00
Mode		2
Std. Deviation		6,069
Variance		36,829
Skewness		,715
Std. Error of Skewness		,001
Kurtosis		-,218
Std. Error of Kurtosis		,002
Range		25
Minimum		0
Maximum		25
Sum		66328462
Percentiles	25	4,00
	50	8,00
	75	13,00

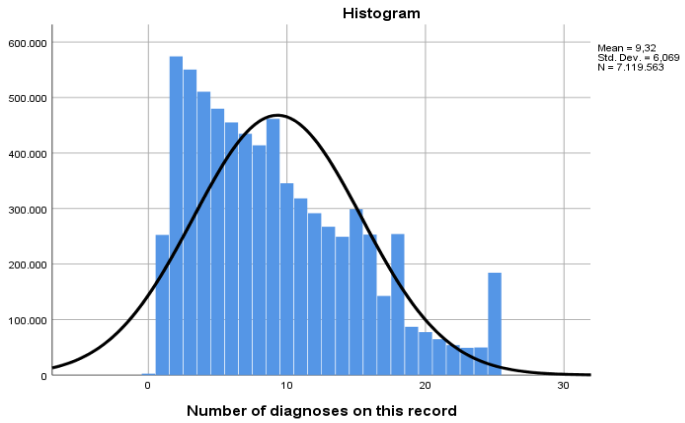
Εικόνα 34

Number of Diagnosis on this Record

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	2424	,0	,0	,0
	1	252381	3,5	3,5	3,6
	2	573997	8,1	8,1	11,6
	3	550270	7,7	7,7	19,4
	4	510381	7,2	7,2	26,5
	5	479769	6,7	6,7	33,3
	6	455067	6,4	6,4	39,7
	7	434678	6,1	6,1	45,8
	8	413769	5,8	5,8	51,6
	9	461547	6,5	6,5	58,1
	10	345461	4,9	4,9	62,9
	11	318178	4,5	4,5	67,4
	12	291565	4,1	4,1	71,5
	13	267197	3,8	3,8	75,2
	14	249093	3,5	3,5	78,7
	15	298875	4,2	4,2	82,9
	16	253007	3,6	3,6	86,5
	17	142433	2,0	2,0	88,5
	18	254004	3,6	3,6	92,1
	19	86995	1,2	1,2	93,3
	20	77331	1,1	1,1	94,4
	21	64556	,9	,9	95,3
	22	53865	,8	,8	96,0
	23	48947	,7	,7	96,7
	24	49533	,7	,7	97,4
	25	184240	2,6	2,6	100,0

Εικόνα 35

Παρατηρούμε ότι οι ασθενείς σε κατά μέσο όρο εμφανίζουν 9,32 διαγνώσεις, διάμεση τιμη 8, συχνότερη τιμή 2 και τυπική απόκλιση 6,069.



Εικόνα 36

Επιπλέον παρατηρούμε ότι το bar chart έχει ισχυρή θετική λοξότητα (skewness) με τιμή 0,715.

Τέλος εξετάζουμε τα περιγραφικά στατιστικά της διάρκειας παραμονής, του αριθμού των διαδικασιών που λάβαν χώρα ενδονοσοκομειακά, ποσοστού θανάτων και του συνολικού κόστους θεραπείας του ασθενούς.

Died during hospitalization

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	6982429	98,1	98,1	98,1
	1	134502	1,9	1,9	100,0
	Total	7116931	100,0	100,0	
Missing	System	2632	,0		
Total		7119563	100,0		

Εικόνα 37

Statistics

		Number of diagnoses on this record	Age in years at admission	Number of procedures on this record	Total charges (cleaned)	Died during hospitalization
N	Valid	7119563	7117978	7119563	6978139	7116931
	Missing	0	1585	0	141424	2632
Mean		9,32	48,65	1,63	39513,32	,02
Median		8,00	54,00	1,00	21166,00	,00
Mode		2	0	0	5207	0
Std. Deviation		6,069	27,615	2,097	74879,987	,136
Variance		36,829	762,579	4,398	5607012411	,019
Skewness		,715	-,378	2,301	14,449	7,066
Std. Error of Skewness		,001	,001	,001	,001	,001
Kurtosis		-,218	-,982	7,583	455,164	47,932
Std. Error of Kurtosis		,002	,002	,002	,002	,002
Range		25	90	15	4991588	1
Minimum		0	0	0	100	0
Maximum		25	90	15	4991688	1
Sum		66328462	346315768	11577850	3,E+11	134502
Percentiles	25	4,00	28,00	,00	10180,00	,00
	50	8,00	54,00	1,00	21166,00	,00
	75	13,00	71,00	2,00	43607,00	,00

Εικόνα 38

Τα Summary statistics των παραπάνω μεταβλητών αναπτύχθηκαν,κατά το πρότυπο δημοσιευμένων εργασιών [44] θεωρώντας ότι υπαρχει πιθανότητα να μας δώσουν κάποιες ενδείξεις, σχετικά με το αν οι μεταβλητές αυτές μπορούν να συσχετιστούν με το κόστος και το ποσοστό θνησιμότητας, που θα εξετάσουμε στη συνέχεια.

Παρατηρούμε ότι το ποσοστό θνησιμότητας είναι 1,9%, ενώ οι υπόλοιπες τρεις μεταβλητές δεν έχουν κανονική κατανομή, γεγονός που συνάγεται από τις μεγάλες αποκλίσεις μεταξύ μέσης τιμής (mean), διάμεσου (median) και συχνότερης τιμής (node).

Στην συνέχεια διενεργούμε έλεγχο σχετικά με το αν υπάρχουν συσχετίσεις correlations, μεταξύ των έντεκα μεταβλητών που δημιουργήσαμε και των ακόλουθων μεταβλητών:

1. Total charges
2. Died during Hospitalization
3. Race
4. Length of Stay
5. Number of diagnoses on this Record
6. Number of Chronic Conditions
7. Number of Procedures on this Record
8. Age in years at admission

Correlations

		Length of stay (cleaned)	Number of chronic conditions	Number of diagnoses on this record	Number of procedures on this record	Emergency	Cardio & Circ	Pneumo	Renal	Myoskeletal_Acute	Total charges	Died during hospitalization
Length of stay (cleaned)	Pearson Correlation	1	,201**	,314**	,330**	,126**	,126**	,153**	,156**	,039**	,641**	,061**
	Sig. (2-tailed)		,000	,000	,000	,000	,000	,000	,000	,000	,000	,000
	N	7119064	7119064	7119064	7119064	7119064	7119064	7119064	7119064	7119064	6977830	7116433
Number of chronic conditions	Pearson Correlation	,201**	1	,860**	,114**	,238**	,682**	,493**	,455**	,025**	,201**	,105**
	Sig. (2-tailed)	,000		,000	,000	,000	,000	,000	,000	,000	,000	,000
	N	7119064	7119563	7119563	7119563	7119563	7119563	7119563	7119563	7119563	6978139	7116931
Number of diagnoses on this record	Pearson Correlation	,314**	,860**	1	,214**	,363**	,599**	,496**	,516**	,061**	,284**	,160**
	Sig. (2-tailed)	,000	,000		,000	,000	,000	,000	,000	,000	,000	,000
	N	7119064	7119563	7119563	7119563	7119563	7119563	7119563	7119563	7119563	6978139	7116931
Number of procedures on this record	Pearson Correlation	,330**	,114**	,214**	1	,122**	,081**	,051**	,059**	,032**	,496**	,126**
	Sig. (2-tailed)	,000	,000	,000		,000	,000	,000	,000	,000	,000	,000
	N	7119064	7119563	7119563	7119563	7119563	7119563	7119563	7119563	7119563	6978139	7116931
Emergency	Pearson Correlation	,126**	,238**	,363**	,122**	1	,205**	,204**	,161**	,093**	,156**	,078**
	Sig. (2-tailed)	,000	,000	,000	,000		,000	,000	,000	,000	,000	,000
	N	7119064	7119563	7119563	7119563	7119563	7119563	7119563	7119563	7119563	6978139	7116931
Cardio & Circ	Pearson Correlation	,126**	,682**	,599**	,081**	,205**	1	,379**	,361**	,046**	,165**	,088**
	Sig. (2-tailed)	,000	,000	,000	,000	,000		,000	,000	,000	,000	,000
	N	7119064	7119563	7119563	7119563	7119563	7119563	7119563	7119563	7119563	6978139	7116931
Pneumo	Pearson Correlation	,153**	,493**	,496**	,051**	,204**	,379**	1	,243**	-,001**	,151**	,099**
	Sig. (2-tailed)	,000	,000	,000	,000	,000	,000		,000	,010	,000	,000
	N	7119064	7119563	7119563	7119563	7119563	7119563	7119563	7119563	7119563	6978139	7116931
Renal	Pearson Correlation	,156**	,455**	,516**	,059**	,161**	,361**	,243**	1	,025**	,133**	,109**
	Sig. (2-tailed)	,000	,000	,000	,000	,000	,000	,000		,000	,000	,000
	N	7119064	7119563	7119563	7119563	7119563	7119563	7119563	7119563	7119563	6978139	7116931
Myoskeletal_Acute	Pearson Correlation	,039**	,025**	,061**	,032**	,093**	,046**	-,001**	,025**	1	,057**	,002**
	Sig. (2-tailed)	,000	,000	,000	,000	,000	,000	,010	,000		,000	,000
	N	7119064	7119563	7119563	7119563	7119563	7119563	7119563	7119563	7119563	6978139	7116931
Total charges (cleaned)	Pearson Correlation	,641**	,201**	,284**	,496**	,156**	,165**	,151**	,133**	,057**	1	,106**
	Sig. (2-tailed)	,000	,000	,000	,000	,000	,000	,000	,000	,000		,000
	N	6977830	6978139	6978139	6978139	6978139	6978139	6978139	6978139	6978139	6978139	6975532
Died during hospitalization	Pearson Correlation	,061**	,105**	,160**	,126**	,078**	,088**	,099**	,109**	,002**	,106**	1
	Sig. (2-tailed)	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	
	N	7116433	7116931	7116931	7116931	7116931	7116931	7116931	7116931	7116931	6975532	7116931

Εικόνα 39

Στο παραπάνω πίνακα παρατηρούμε ότι όλες οι συσχετίσεις παρουσιάζονται στατιστικά σημαντικές, και οι ισχυρότερες R^2 στις μεταβλητές έχουν επισημανθεί με κόκκινο. Εμφανίζονται ισχυρές συσχετίσεις μεταξύ Total charges & Length of Stay, μεταξύ Total charges & Number of Procedures, μεταξύ Number of chronic conditions & Cardio and Circulatory, μεταξύ Number of chronic conditions & Number of diagnoses, μεταξύ Number of chronic conditions & Pneumo, μεταξύ Number of diagnoses με Number of chronic conditions , Cardio and Circulatory, Pneumo, Renal. Παράλληλα υπάρχουν στατιστικά σημαντικές, ασθενέστερες μεν συσχετίσεις μεταξύ των υπόλοιπων μεταβλητών.

3.3.2Principal Component Analysis - PCA

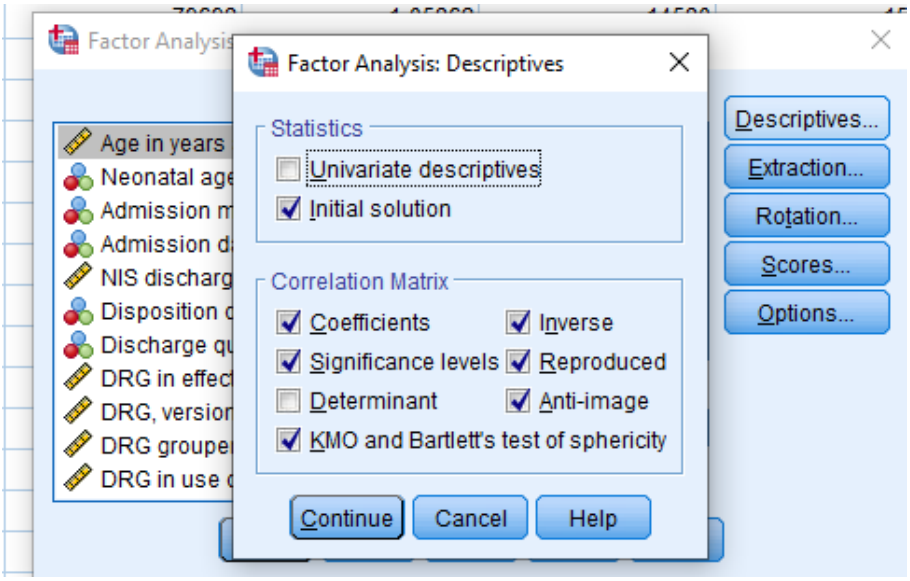
Επειδή ό αριθμός των μεταβλητών είναι ήδη μεγάλος, ενώ ταυτόχρονα, πολλές είναι μεν στατιστικά σημαντικές με μικρή όμως Pearson Correlation, πραγματοποιήσαμε Principal Component Analysis – PCA με σκοπό την δημιουργία λιγότερων Components τα οποία θα μπορούν να επεξηγήσουν καλύτερα την διακύμανση της εξαρτημένης μεταβλητής. Οι μεταβλητές που χρησιμοποιήθηκαν είναι οι ακόλουθες:

1. Cancer
2. Cardio & Circulatory
3. Pneumo
4. Gastro
5. Renal
6. Genital
7. Pregnancy
8. Bones_Joints_Chronic
9. Congenital
10. Myoskeletal_Acute
11. Emergency
12. Length of Stay
13. Number of diagnoses on this Record

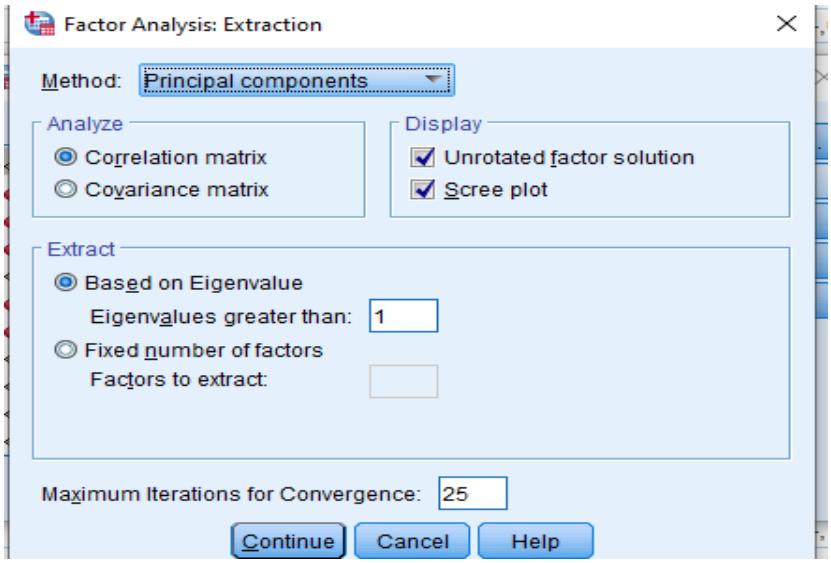
14. Number of Chronic Conditions

15. Number of Procedures on this Record

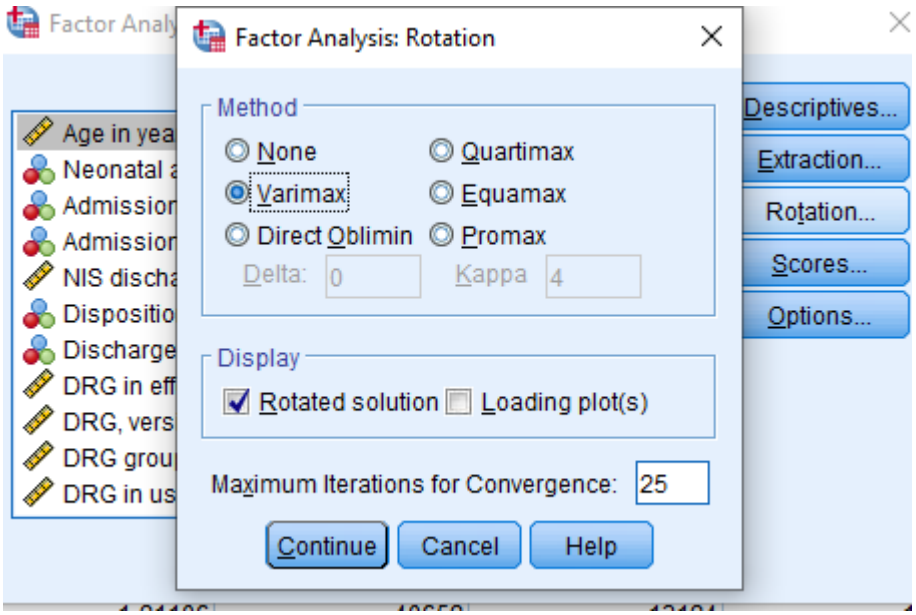
Στην Factor Analysis ενεργοποιήθηκαν οι εξής παράμετροι:



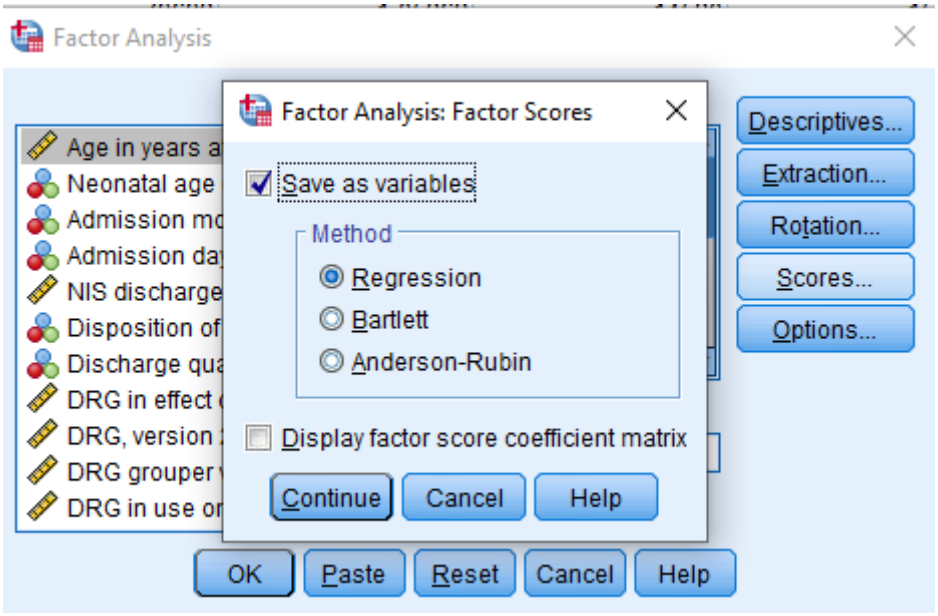
Εικόνα 40



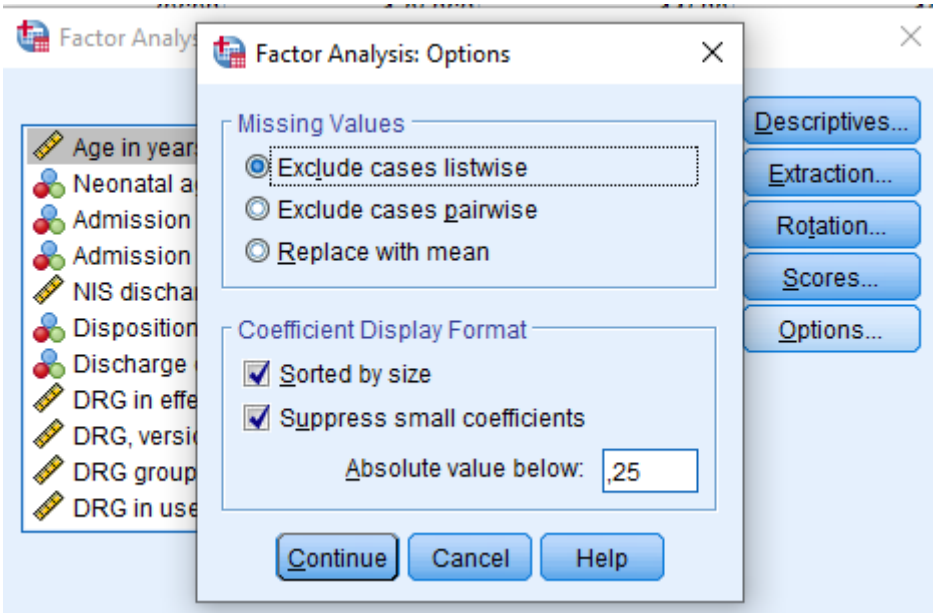
Εικόνα 41



Εικόνα 42



Εικόνα 43



Εικόνα 44

3.3.2.Α Επεξήγηση των αποτελεσμάτων της PCA. Εξετάζοντας το KMO Test παρατηρούμε ότι η τιμή που λαμβάνουμε είναι 0,739 συνεπώς επαρκώς καλή ώστε να επιβεβαιώσει την καταλληλότητα των δεδομένων για Factor Analysis. Επιπροσθέτως στο Bartlett Test λαμβάνουμε τιμή p=0,000 επιβεβαιώνοντας πως στην περίπτωση μας έχουμε να κάνουμε με Correlation Matrix και όχι με Identity Matrix.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,739
Bartlett's Test of Sphericity	Approx. Chi-Square	33815846,990
	df	105
	Sig.	,000

Εικόνα 45

Στην συνέχεια εξετάζουμε τα communalities του πίνακα που προκύπτουν και παρατηρούμε ότι υπάρχουν τιμές με υψηλή communality > 0.6 καθώς και τιμές με επαρκή

communality > 0,4 οι οποίες υποδεικνύουν σημαντικές συσχετίσεις μεταξή των items (Cancer, Cardio & Circulatory, Pneumo, Gastro, Renal, Genital, Pregnancy, Bones_Joints_Chronic, Congenital Myoskeletal_Acute, Emergency Length of Stay, Number of diagnoses on this Record, Number of Chronic Conditions, Number of Procedures on this Record) και των components.

Communalities

	Initial	Extraction
Length of stay (cleaned)	1,000	,657
Number of chronic conditions	1,000	,819
Number of diagnoses on this record	1,000	,814
Number of procedures on this record	1,000	,653
Cancer(1 if any of DXCCS1 to DXCCS25 is between 10 and 45, else 0)	1,000	,133
Cardio & Circ (1 if any of DXCCS1 to DXCCS25 is between 95 and 122, else 0)	1,000	,647
Pneumo(1 if any of DXCCS1 to DXCCS25 is between 121 and 135, else 0)	1,000	,844
Gastro(1 if any of DXCCS1 to DXCCS25 is between 137 and 156, else 0)	1,000	,835
Renal(1 if any of DXCCS1 to DXCCS25 is between 155 and 164, else 0)	1,000	,448
Myoskeletal_Acute(1 if any of DXCCS1 to DXCCS25 is between 224 and 233, else 0)	1,000	,754
Emergency(1 if any of DXCCS1 to DXCCS25 is between 232 and 254, else 0)	1,000	,312
Genital(1 if any of DXCCS1 to DXCCS25 is between 164 and 177, else 0)	1,000	,318
Pregnancy(1 if any of DXCCS1 to DXCCS25 is between 175 and 197, else 0)	1,000	,682
Bones_Joints_Chronic(1 if any of DXCCS1 to DXCCS25 is between 200 and 213, else 0)	1,000	,362
Congenital(1 if any of DXCCS1 to DXCCS25 is between 212 and 225, else 0)	1,000	,684

Extraction Method: Principal Component Analysis.

Εικόνα 46

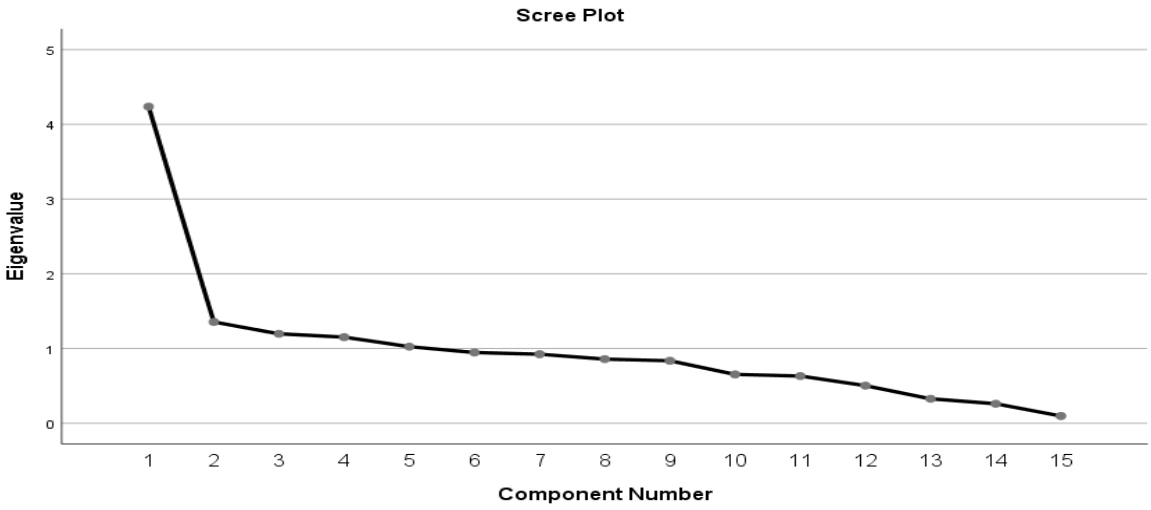
Στον πίνακα Total Variance Explained, παρατηρούμε ότι τα πέντε πρώτα components, τα οποία είναι και αυτά που έχουν eigenvalue μεγαλύτερη της μονάδας, εξηγούν το 59,766% της διακύμανσης.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4,237	28,244	28,244	4,237	28,244	28,244	3,219	21,459	21,459
2	1,356	9,037	37,281	1,356	9,037	37,281	1,899	12,662	34,121
3	1,196	7,975	45,257	1,196	7,975	45,257	1,414	9,426	43,547
4	1,152	7,678	52,935	1,152	7,678	52,935	1,238	8,252	51,800
5	1,025	6,831	59,766	1,025	6,831	59,766	1,195	7,966	59,766
6	,947	6,311	66,077						
7	,923	6,156	72,233						
8	,858	5,718	77,951						
9	,836	5,572	83,523						
10	,653	4,355	87,878						
11	,631	4,209	92,088						
12	,503	3,353	95,440						
13	,327	2,180	97,621						
14	,261	1,737	99,358						
15	,096	,642	100,000						

Extraction Method: Principal Component Analysis.

Εικόνα 47



Εικόνα 48

Στο Rotated Component Matrix μπορούμε να αποτιμήσουμε την βαρύτητα (loadings) με την οποία συμμετέχουν οι μεταβλητές μας σε κάθε component. Χάριν ευκολίας μεταβλητές που επηρεάζουν το component κάτω από 0,25 παραλήφθηκαν από την αποτύπωση.

Rotated Component Matrix[a]

	Component				
	1	2	3	4	5
Number of chronic conditions	,854	,270			
Number of diagnoses on this record	,801	,282	,276		
Cardio & Circ (1 if any of DXCCS1 to DXCCS25 is between 95 and 122, else 0)	,773				
Renal(1 if any of DXCCS1 to DXCCS25 is between 155 and 164, else 0)	,654				
Bones_Joints_Chronic(1 if any of DXCCS1 to DXCCS25 is between 200 and 213, else 0)	,483				,330
Cancer(1 if any of DXCCS1 to DXCCS25 is between 10 and 45, else 0)	,307				
Gastro(1 if any of DXCCS1 to DXCCS25 is between 137 and 156, else 0)		,897			
Pneumo(1 if any of DXCCS1 to DXCCS25 is between 121 and 135, else 0)	,301	,865			
Length of stay (cleaned)			,783		
Number of procedures on this record			,774		
Pregnancy(1 if any of DXCCS1 to DXCCS25 is between 175 and 197, else 0)	-,356			,705	
Congenital(1 if any of DXCCS1 to DXCCS25 is between 212 and 225, else 0)	-,397			-,600	-,301

Genital(1 if any of DXCCS1 to DXCCS25 is between 164 and 177, else 0)				,547	
Myoskeletal_Acute(1 if any of DXCCS1 to DXCCS25 is between 224 and 233, else 0)					,857
Emergency(1 if any of DXCCS1 to DXCCS25 is between 232 and 254, else 0)					,416

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.[a]

a. Rotation converged in 6 iterations.

Εικόνα 49

3.3.3Correlation Analysis on Total Charges & Died During Hospitilization

Στην συνέχεια τρέξαμε Pearson Coefficient R^2 με τα 5 components της PCA κι εξαρτημένη μεταβλητή το Total Charges και το Died During Hospitalization. Παρατίθενται τα αποτελέσματα στους παρακάτω πίνακες:

Correlations

		Total charges (cleaned)	PCA Factor1 without Death	PCA Factor2 without Death	PCA Factor3 without Death	PCA Factor4 without Death	PCA Factor5 without Death
Total charges (cleaned)	Pearson Correlation	1	,132**	,094**	,644**	-,033**	,053**
	Sig. (2-tailed)		,000	,000	,000	,000	,000
	N	6978139	6975224	6975224	6975224	6975224	6975224
PCA Factor1 without Death	Pearson Correlation	,132**	1	,000	,000	,000	,000
	Sig. (2-tailed)	,000		1,000	1,000	1,000	1,000
	N	6975224	7116433	7116433	7116433	7116433	7116433
PCA Factor2 without Death	Pearson Correlation	,094**	,000	1	,000	,000	,000
	Sig. (2-tailed)	,000	1,000		1,000	1,000	1,000
	N	6975224	7116433	7116433	7116433	7116433	7116433
PCA Factor3 without Death	Pearson Correlation	,644**	,000	,000	1	,000	,000
	Sig. (2-tailed)	,000	1,000	1,000		1,000	1,000
	N	6975224	7116433	7116433	7116433	7116433	7116433
PCA Factor4 without Death	Pearson Correlation	-,033**	,000	,000	,000	1	,000
	Sig. (2-tailed)	,000	1,000	1,000	1,000		1,000
	N	6975224	7116433	7116433	7116433	7116433	7116433
PCA Factor5 without Death	Pearson Correlation	,053**	,000	,000	,000	,000	1
	Sig. (2-tailed)	,000	1,000	1,000	1,000	1,000	
	N	6975224	7116433	7116433	7116433	7116433	7116433

**. Correlation is significant at the 0.01 level (2-tailed).

Εικόνα 50

Correlations

		Died during hospitalization	PCA Factor1 without Death	PCA Factor2 without Death	PCA Factor3 without Death	PCA Factor4 without Death	PCA Factor5 without Death
Died during hospitalization	Pearson Correlation	1	,213**	-,098**	,321**	-,070**	-,368**
	Sig. (2-tailed)		,000	,000	,000	,000	,000
	N	7116931	7116433	7116433	7116433	7116433	7116433
PCA Factor1 without Death	Pearson Correlation	,213**	1	,000	,000	,000	,000
	Sig. (2-tailed)	,000		1,000	1,000	1,000	1,000
	N	7116433	7116433	7116433	7116433	7116433	7116433
PCA Factor2 without Death	Pearson Correlation	-,098**	,000	1	,000	,000	,000
	Sig. (2-tailed)	,000	1,000		1,000	1,000	1,000
	N	7116433	7116433	7116433	7116433	7116433	7116433
PCA Factor3 without Death	Pearson Correlation	,321**	,000	,000	1	,000	,000
	Sig. (2-tailed)	,000	1,000	1,000		1,000	1,000
	N	7116433	7116433	7116433	7116433	7116433	7116433
PCA Factor4 without Death	Pearson Correlation	-,070**	,000	,000	,000	1	,000
	Sig. (2-tailed)	,000	1,000	1,000	1,000		1,000
	N	7116433	7116433	7116433	7116433	7116433	7116433
PCA Factor5 without Death	Pearson Correlation	-,368**	,000	,000	,000	,000	1
	Sig. (2-tailed)	,000	1,000	1,000	1,000	1,000	
	N	7116433	7116433	7116433	7116433	7116433	7116433

**. Correlation is significant at the 0.01 level (2-tailed).

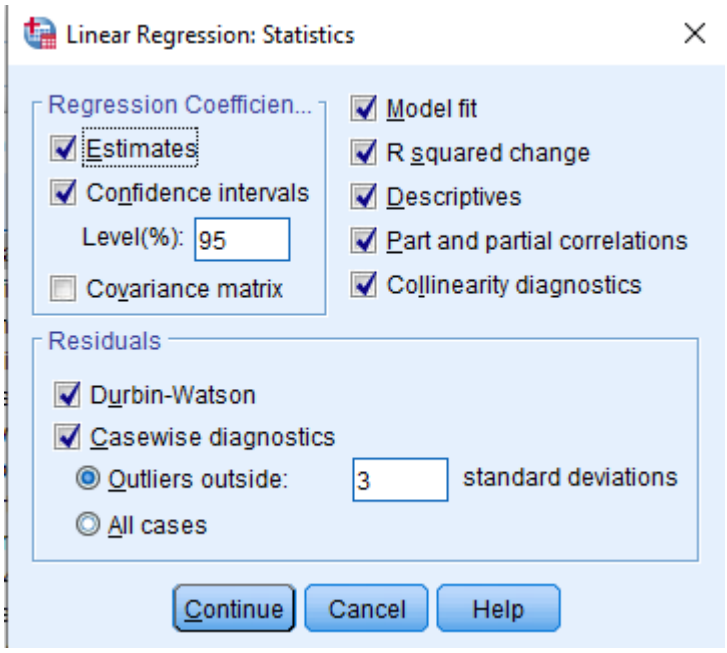
Εικόνα 51

Από τα αποτελέσματα παρατηρούμε πως η εξαρτημένη μεταβλητή Total Charges σχετίζεται με το Component 3. Αν ανατρέξουμε στις μεταβλητές που έχουν τα μεγαλύτερα loading σε αυτά τα components θα παρατηρήσουμε ότι το PCA Factor 3 περιέχει ως πιο επιδραστικές μεταβλητές τις Length of stay, Number of procedures on this record. Αν ανατρέξουμε στον αρχικό πίνακα με τα NIS_2013_Correlations, θα παρατηρήσουμε πως όντως, το Total Charges συσχετιζόταν ισχυρά με Length of stay, Number of procedures on this record με συντελεστή Pearson 0.641 & 0.496 αντίστοιχα και p value = 0.000

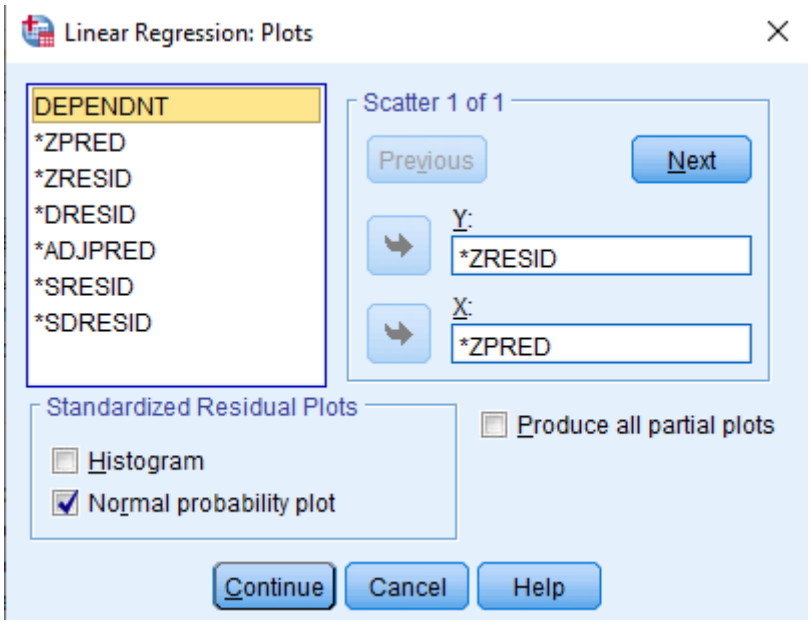
Κατ' αντίστοιχο τρόπο η εξαρτημένη μεταβλητή Died during hospitalization συσχετίζεται – αν και ασθενώς- επίσης με τα PCA Factor 1 & PCA Factor 3.

3.3.4Regression Analysis on Total Charges

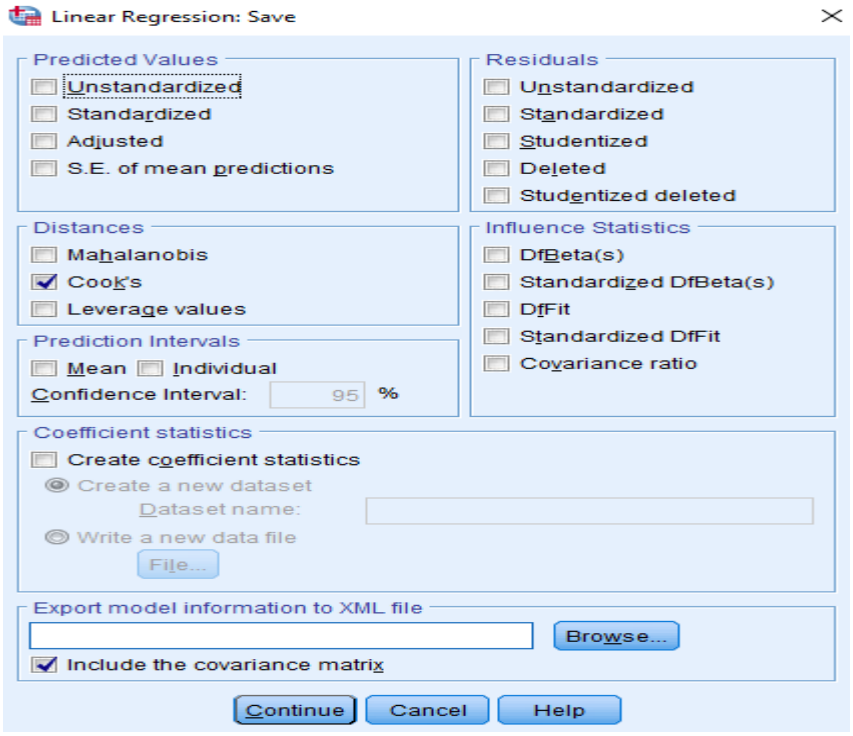
Κλείνοντας την ανάλυση μας πραγματοποιήσαμε δύο ξεχωριστές Linear Regression Analysis με εξαρτημένη μεταβλητή την Total Charges την πρώτη φορά και την Died during hospitalization την δεύτερη. Στην ανάλυση μας χρησιμοποιήσαμε τους ακόλουθους ελέγχους στο SPSS.



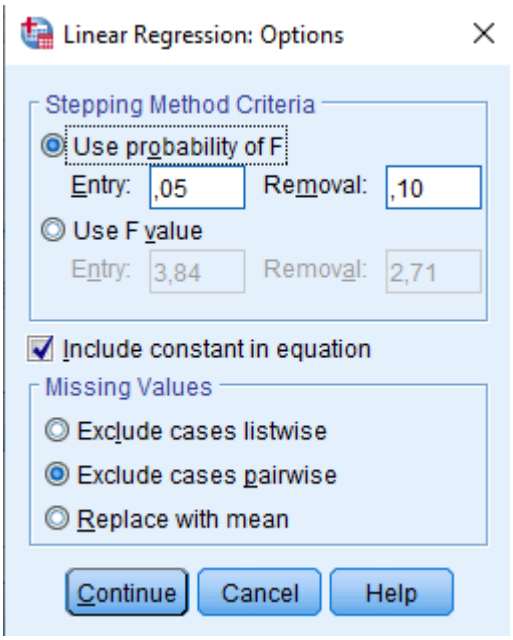
Εικόνα 52



Εικόνα 53



Εικόνα 54



Εικόνα 55

3.3.4.A Regression Analysis on Total Charges ανάλυση αποτελεσμάτων

Τα αποτελέσματα που λάβαμε για την πρώτη ανάλυση παρουσιάζονται στους παρακάτω πίνακες:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95,0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	39513,320	20,744		1904,774	,000	39472,662	39553,978					
	Factor score1	11714,723	20,744	,156	564,718	,000	11674,065	11755,381	,156	,209	,156	1,000	1,000
	Factor score2	5652,318	20,744	,075	272,475	,000	5611,659	5692,976	,075	,103	,075	1,000	1,000
	Factor score3	48880,017	20,744	,653	2356,303	,000	48839,359	48920,675	,653	,666	,653	1,000	1,000
	Factor score4	-2794,043	20,744	-,037	-134,689	,000	-2834,702	-2753,385	-,037	-,051	-,037	1,000	1,000
	Factor score5	6165,101	20,744	,082	297,194	,000	6124,442	6205,759	,082	,112	,082	1,000	1,000

a. Dependent Variable: Total charges (cleaned)

Εικόνα 55

Αρχικά παρατηρούμε ότι δεν υπάρχει collinearity μεταξύ των προβλεπτικών μεταβλητών. Η tolerance είναι 1 συνεπώς μεγαλύτερη από το όριο > 0,1 ενώ και η VIF έχει επίσης τιμή 1, μικρότερη από το όριο < 10. Οι συντελεστές standardized coefficients b έχουν καλές τιμές με προεξέχουσα για το Factor Score 3 = 0.653, ενώ και οι partial correlations δίνουν αρκετά καλές τιμές με 0,666 στο Factor score 3 και σχετικά ικανοποιητικές στα Factor score 1 = 0.209, Factor score 2 = 0.103 & Factor score 5 = 0.112.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	,682[a]	,464	,464	54797,444	,464	1210356,093	5	6977824	,000	1,666

a. Predictors: (Constant), Factor score5, Factor score4, Factor score3, Factor score2 , Factor score1

b. Dependent Variable: Total charges (cleaned)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1,817E+16	5	3,634E+15	1210356,093	,000[b]
	Residual	2,095E+16	6977824	3002759901		
	Total	3,912E+16	6977829			

a. Dependent Variable: Total charges (cleaned)

b. Predictors: (Constant), Factor score5, Factor score4, Factor score3, Factor score2 , Factor score1

Εικόνα 56

Παρατηρούμε ότι το τεστ ANOVA έχει p value = 0.000 γεγονός που δίνει στο μοντέλο προβλεπτική ισχυ. Το μοντέλο με συντελεστή adjusted R^2 επεξηγεί το 46,44% της διακύμανσης.

ΚάνονταςΤο Darbin Watson Test Βρίσκεται στο 1.666, το οποίο μας δείχνει αρνητική αυτοσχεσχέτιση των residuals.

3.3.4.B Regression Analysis on Died during hospitalization

Στην συνέχεια, θέλοντας να εξετάσουμε αν μπορούμε να προβλέψουμε την πιθανότητα επιβίωσης ενός εσθενή, εφαρμόζουμε regression analysis με τις ίδιες προβλεπτικές μεταβλητές, αλλά αυτήν την φορά με εξαρτημένη μεταβλητή το Died during hospitalization και παίρνουμε τους ακόλουθους πίνακες.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	,169[a]	,028	,028	,134	,028	41741,433	5	7116427	,000	1,995

a. Predictors: (Constant), Factor score5, Factor score4, Factor score2 , Factor score21, Factor score3

b. Dependent Variable: Died during hospitalization

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3757,543	5	751,509	41741,433	,000[b]
	Residual	128123,441	7116427	,018		
	Total	131880,984	7116432			

a. Dependent Variable: Died during hospitalization

b. Predictors: (Constant), Factor score5, Factor score4, Factor score2 , Factor score21, Factor score3

Εικόνα 57

Με βάση τα δεδομένα που προκύπτουν απο την Regression Analysis παρατηρούμε ότι το τεστ ANOVA έχει και αυτήν τη φορά p value = 0.000. Αυτό μας δείχνει κατ' αρχάς ότι το μοντέλο μας έχει προβλεπτική ισχύ. Εν τούτοις επεξηγεί μόνο το 2,80% =R^2 της διακύμανσης. Επιπλέον τόσο οι συντελεστές b, όσο και τα standardized coefficients είναι πολύ κοντά στο 0, ενώ και η partial correlation των factors είναι επίσης χαμηλή, με υψηλότερη τιμή του factor 3 που ανέρχεται στο 0.118 επεξηγεί δηλαδή μόνος του το 11,80% της εξαρτημένης μεταβλητής. Το Darbin Watson Test Βρίσκεται στο 1.995, το οποίο μας δείχνει ουδέτερη αυτοσυσχέτιση των residuals.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	,019	,000		375,638	,000	,019	,019					
	Factor score21	,015	,000	,113	305,485	,000	,015	,015	,113	,114	,113	1,000	1,000
	Factor score2	,006	,000	,045	122,817	,000	,006	,006	,045	,046	,045	1,000	1,000
	Factor score3	,016	,000	,117	316,376	,000	,016	,016	,117	,118	,117	1,000	1,000
	Factor score4	-,001	,000	-,004	-10,963	,000	-,001	,000	-,004	-,004	-,004	1,000	1,000
	Factor score5	,000	,000	,004	9,906	,000	,000	,001	,004	,004	,004	1,000	1,000

a. Dependent Variable: Died during hospitalization

Εικόνα 58

Συνεπώς, συμπεραίνουμε ό,τι σε αντίθεση με την πρώτη περίπτωση που εξετάσαμε τα total charges σε συνάρτηση με τους factors 1-5, αντίστοιχο μοντέλο πρόβλεψης επιβίωσης δεν αποδεικνύεται αποτελεσματικό, χρησιμοποιώντας τις ίδιες εξαρτημένες μεταλητές.

Κεφάλαιο 4

Ανάλυση δεδομένων NIS και δημιουργία προβλεπτικού μοντέλου με χρήση Machine Learning (ML) Model

4.1 The predictive Machine Learning Model (ML)

Στην κλασσική στατιστική ανάλυση οι μεταβλητές που επηρεάζουν το τελικό αποτέλεσμα (output) θα πρέπει να επιλεγούν εξ αρχής, μειώνοντας την δυνατότητα του μοντέλου να διαχειριστεί και να επεξηγήσει μαζικά δεδομένα, αξιοποιώντας εν τέλει σε χαμηλό βαθμό τις διαθέσιμες πληροφορίες [13]. Αντίθετα στην μηχανική μάθηση, η επιλογή μεταβλητών (feature selection methods) γίνεται αυτόματα από το μοντέλο, που επιλέγει τις μεταβλητές οι οποίες συνεισφερουν σε μεγαλύτερο βαθμό στο τελικό αποτέλεσμα (output) [40]. Έτσι δεν περιορίζει τον αριθμό των μεταβλητών που θα χρησιμοποιηθούν στην επεξήγηση του τελικού αποτελέσματος, εμποδίζοντας την απώλεια πληροφορίας [38].

Στην περίπτωση των ιατρικών δεδομένων, παρ' όλα αυτά, είναι ουσιώδες πέρα από την επιλογή των μεταβλητων, να υπάρχει και ιατρική επεξηγηματικότητα του μοντέλου, και κατανόηση των παραγόντων που επηρεάζουν το τελικό αποτέλεσμα [46] καθώς η έλλειψη επεξηγηματικότητας είναι πιθανόν να κρύβει biases καθώς και συγχυτικές μεταβλητές (confounding variables) [6]. Ειδικά τα deep learning models, θεωρούνται "black box", καθώς με η πρόβλεψη στηρίζεται στις σχέσεις μεταξύ των μεταβλητών και όχι στην ιατρική εξειδίκευση. Έτσι το τελικό αποτέλεσμα είναι πιθανό να είναι biased, καθώς επηρεάζεται κατά κύριο λόγο από την παρουσία (representativeness) των δεδομένων και όχι την και την κλινική παρατήρηση [38][40].

4.2 Μεθοδολογία

Για την έρευνα χρησιμοποιήθηκαν, όπως και στην στατιστική ανάλυση με το SPSS δεδομένα το έτους 2013. Οι μεταβλητές αυτές περιλαμβάνουν δημογραφικά δεδομένα, οικονομικά στοιχεία, δεδομένα σχετικά με την διαστωμάτωση (strata) του νοσοκομείου, την κατάσταση του ασθενή, τις ασθένειες όπως καταχωρούνται στο γενικευμένο CCS (Clinical Classifications Software) του NIS, τις διαδικασίες που ακολουθήθηκαν ενδονοσοκομειακά και

πάλι βάσει του Single-Level CCS – Procedures του NIS. Οι μεταβλητές αυτές περιλαμβάνουν δημογραφικά δεδομένα, οικονομικά στοιχεία, δεδομένα σχετικά με την διαστωμάτωση (strata) του νοσοκομείου, την κατάσταση του ασθενή, τις ασθένειες όπως καταχωρούνται στο γενικευμένο CCS (Clinical Classifications Software) του NIS, τις διαδικασίες που ακολουθήθηκαν ενδονοσοκομειακά και πάλι βάσει του Single-Level CCS – Procedures του NIS. Απο το αρχικό dataset, το οποίο αποτελούνταν από περίπου $7*10^6$ εγγραφές με 141 ξεχωριστά variables, κρατήθηκε για λόγους υπολογιστικής δυνατότητας το 10% των συμμετεχόντων, περίπου 700000. Επιπλέον από τις 141 μεταβλητές κρατήθηκαν οι 39, ενώ για τις μεταβλητές που απορρίφθηκαν η αιτία ήταν ο πολύ υψηλός αριθμός από missing values. Ειδικά στην πρόβλεψη θνησιμότητας αφαιρέθηκε και η DISPUNIFORM καθώς αυτή η μεταβλητή κωδικοποιούσε το προς τα που (άλλη μονάδα, επιστροφή στο σπίτι κλπ) κατευθύνθηκε ο ασθενής κατά το εξιτήριο, έχοντας και ως πιθανή επιλογή το Died, οπότε στην ουσία ήταν 100% predictive μεταβλητή.

Στην περίπτωση της πρόβλεψης θνησιμότητας, όπου το αποτέλεσμα αποτελεί κατηγορική μεταβλητή, κάθε σύνολο δεδομένων χωρίστηκε σε training (75%) και testing set (25%). Προτιμήθηκε extreme Gradient Boosting (XGBoost) για τη δημιουργία του προγνωστικού μοντέλου, καθώς το XGBoost παρέχει μια υπερπαράμετρο σχεδιασμένη να να ρυθμίζει (tuning) τη συμπεριφορά του αλγορίθμου για imbalanced datasets σε classification problems.

Ένας δεύτερος λόγος που προτιμήθηκε ο συγκεκριμένος αλγόριθμος οφείλεται στην επεκτασιμότητα (scalability) του, τόσο ως προς το είδος των διαφορετικών σεναρίων που μπορεί να διαχειριστεί, όσο και ως προς το μέγεθος του δείγματος. Στην επιλογή του συνηγορεί και η ταχύτητα του σε συνδιασμό με την περιορισμένη ανάγκη για υπολογιστική ισχύ [43].

4.3. Πρόβλεψη με XGBoost της πιθανότητας θνησιμότητας [48]

Στην περίπτωση της δημιουργίας του ML μοντέλου για την πρόβλεψη της θνησιμότητας χρησιμοποιήθηκε ο XGBoost Classifier. Στην στήλη y = DIED (1,0) αφαιρέθηκαν όσα rows δεν είχαν τιμή. Επιπλέον αξίζει να σημειωθεί ότι στην περίπτωση μας είχαμε να διαχειριστούμε ένα ήταν ένα ιδιαιτέρως imbalanced dataset με 12108 τιμές 1=DIED & 620800 τιμές 0 = NOT DIED. Σε μια πρώτη προσπάθειαδημιουργίας προβλεπτικού μοντέλου, αυτό μας οδηγούσε στα παρακάτω αποτελέσματα: F1 Score: 0.0000 ROC-AUC: 0.5000 Accuracy: 0.9796 Precision: 0.0000 Recall:

0.0000. Για να αντιμετωπίσουμε την αδυναμία του XGBoost Classifier στο imbalanced dataset και δεδομένου ότι το πρόβλημα μας ήταν ένα binary classification problem, προβήκαμε σε προεπεξεργασία των δεδομένων με την μέθοδο undersampling παίρνοντας 12108 τιμές με κωδικοποίηση 1=DIED & 12108 τιμές με κωδικοποίηση 0 = NOT DIED.

Για την βελτιστοποίηση της απόδοσης του αλγορίθμου, ρυθμίστηκαν μια σειρά παραμέτρων. Ακριβώς επειδή το dataset ήταν imbalanced δημιουργήσαμε ένα 5 Fold stratified cross validation (CV), όπου η κατανομή των θετικών και των αρνητικών εκβάσεων (1,0) ήταν ίδια σε κάθε ένα απο τα πέντε αυτά folds [39].

Η διαδικασία του Cross Validation επαναλήφθηκε 400 φορές n_estimators για να μειωθεί τόσο η variance όσο και η bias, συνεπώς σε κάθε επανάληψη το μοντέλο δημιουργούργησε και αξιολόγησε 2000 μοντέλα. Το testing set δεν ήταν μέρος του training ή του validation set και συνεπώς αξιολόγησε την απόδοση του μοντέλου, σε άγνωστο dataset.

Σε αυτή τη μελέτη, χρησιμοποιήσαμε την SHapley Additive Explanation framework (SHAP), το οποίο είναι ένα επεξήγηματικό μοντέλο που βασίζεται στις τιμές Shapley. Η τιμή Shapley είναι η μέση οριακή συνεισφορά (marginal contribution) της τιμής μιας μεταβλητής, εν μέσω όλων των πιθανών συνδιασμών μεταβλητών [8][9]. Οι τιμές Shapley, αν και δεν είναι ακριβώς το ίδιο ούτε και προέρχονται από την ίδια μεθοδολογία, ομοιάζουν στις partial coefficients του SPSS, καθώς όπως και αυτές βελτιώνουν την διορατικότητα μας (insight), σχετικά με την συνεισφορά των μεταβλητών στο τελικό αποτέλεσμα. Στην συνέχεια, έχοντας πλέον ένας balanced dataset τρέξαμε τον XGBoost Classifier με το ακόλουθο Grid of hyperparameters:

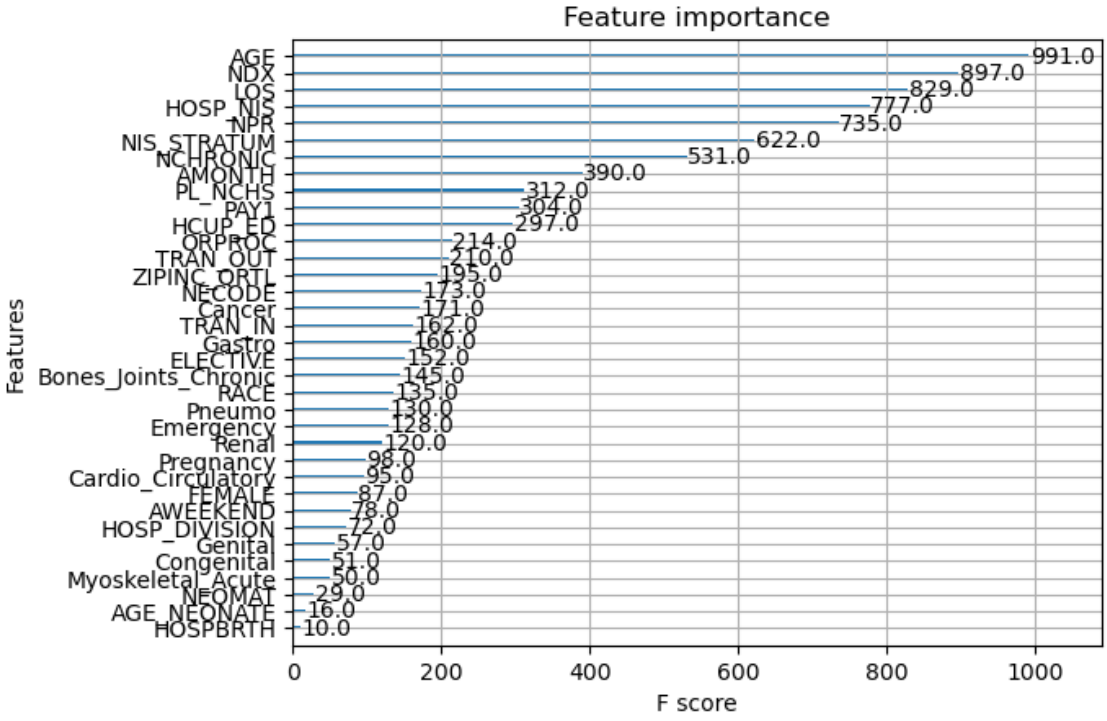
```
paramGrid = {
    "learning_rate": [0.05],
    "n_estimators": [200, 300, 400],
    "max_depth": [3, 4, 5],
    'subsample': [0.8],
    "colsample_bylevel": [0.8],
    "colsample_bytree": [0.8],
    'gamma': [0.2]
}
```

Εικόνα 59

Ως βέλτιστη επιλογή προκρίθηκε με Best ROC AUC score στο training set: 0.95 ο ακόλουθος συνδιασμός υπερπαραμέτρων: Best parameters: {'colsample_bylevel': 0.8, 'colsample_bytree': 0.8, 'gamma': 0.2, 'learning_rate': 0.05, 'max_depth': 5, 'n_estimators': 400, 'subsample': 0.8}. Τα αποτελέσματα που πήραμε στο test set ήταν τα ακόλουθα: F1 Score: 0.8865, ROC-AUC: 0.8852, Accuracy: 0.8849, Precision: 0.8624, Recall: 0.9119 .

Με βάση αυτά τα αποτελέσματα, μπορούμε να θεωρήσουμε τον classifier ως πολύ καλό στην πρόβλεψη της θνησιμότητας.

Όσον αφορά το κομμάτι επιμέρους μεταβλητών στο παρακάτω σχήμα μπορούμε να δούμε την συνεισφορά στην πρόβλεψη της κάθε μεταβλητής στην παρακάτω εικόνα:

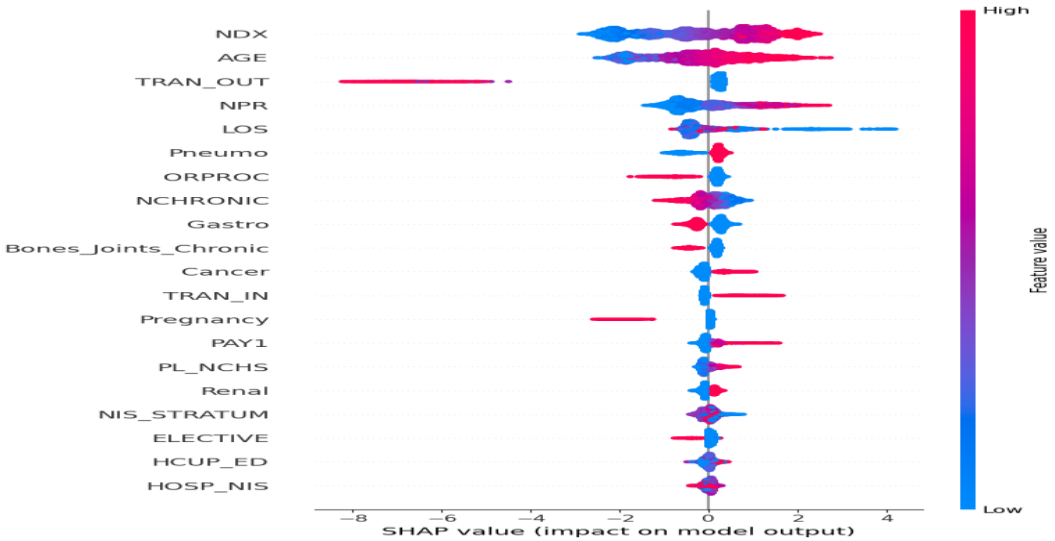


Εικόνα 60

Εξετάζοντας τις μεταβλητές με την μεγαλύτερη προβλεπτική συνεισφορά παρατηρούμε πως πρώτη έρχεται η ηλικία του ασθενούς, ακολουθούμενη από τον δείκτη NDX [14] που

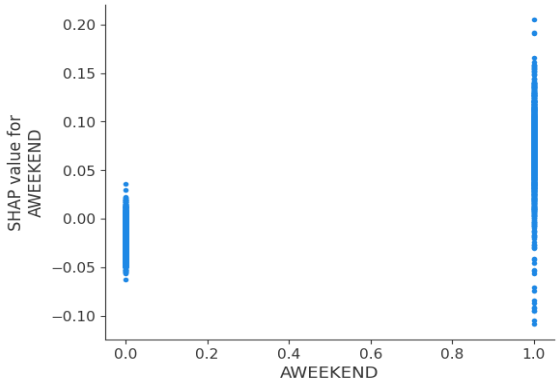
αναφέρεται στον αριθμό των κωδικοποιημένων διαγνώσεων του κατά την διάρκεια της νοσηλείας του. Όσο μεγαλύτερο το νούμερο ο ασθενής έχει διαγνωστεί με περισσότερες ασθένειες. Τρίτος παράγοντας είναι το LOS (Length of Stay), ενώ τέταρτος ακόλουθεί το HOSP_NIS. Το HOSP_NIS είναι ένας αριθμός ο οποίος σχετίζεται με το νοσοκομείο και αναφέρεται στην βαρύτητα του δείγματος των δεδομένων του κάθε τοπικού νοσοκομείου, που αποστέλονται στο NIS κατά το εξιτήριο του ασθενούς. Υψηλότερος αριθμός HOSP_NIS υποδεικνύει νοσοκομείο με μεγαλύτερες δυνατότητες, μεγαλύτερο αριθμό ασθενών, υψηλότερο αριθμό κλινών, περισσότερες διαδικασίες και συνεπώς τα εξιτήρια του σαν όγκος αλλά και σαν δεδομένα, είναι πολύ περισσότερα σε σχέση με ένα περιφερειακό νοσοκομείο. Κατα πόσο αυτη μεταβλητή είναι κλινικά ουσιαστικά προβλεπτική παραμένει ένα ερώτημα. Ακολουθεί , πέμπτη, η μεταβλητή NPR (Number of Procedures), οποία υποδεικνύει τον αριθμό των διαδιασιών που πραγματοποιήθηκαν για τον συγκεκριμένο ασθενή. NIS_STRATUM, ως έκτη, είναι η επόμενη μεταβλητή με σημαντική επίδραση στο μοντέλο. Πρόκειται και σε αυτήν την περίπτωση για έναν αριθμό που κατατάσσει το νοσοκομείο στην αντιστοιχη κατηγορίας διαστρωμάτωσης που του έχει αποδοθεί από το NIS. Και σε αυτήν την περίπτωση έχουμε να κάνουμε με την βαρύτητα των δεδομένων που αποστέλονται στο NIS.

Στην συνέχεια εξετάσαμε τις Shap Values παίρνοντας το παρακάτω διάγραμμα.

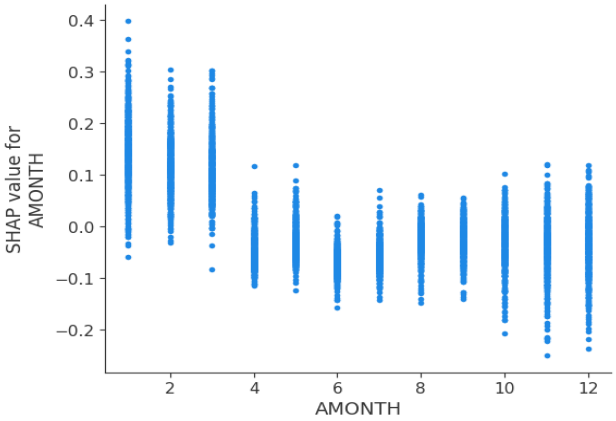


Εικόνα 61

Κι εδώ παρατηρούμε ότι υψηλές τιμέ σε AGE, NDX & NPR (Number of Procedures) συσχετίζονται με αύξηση της πιθανότητας θνησιμότητας ενώ χαμηλές τιμές LOS (Length of Stay) επίσης συσχετίζονται με αύξηση της πιθανότητας θανάτου [39]. Μελετώντας την συνεισφορά της κάθε μεταβλητής στο τελικό αποτέλεσμα, όσο μεταβάλεται η τιμή της, ανακαλύπτουμε ένα κάπως «παράξενο» εύρημα. Όσο υψηλότερος ο καταγεγραμμένος αριθμός των χρόνιων ασθενειών, η συνεισφορά τους στο ποσοστό ενδονοσοκομειακών θανάτων είναι αρνητική. Επίσης δύο απροσδόκητα ευρύματα, συνδέουν αυξημένη θετική συνεισφορά στην θνησιμότητα αν η εισαγωγή στο νοσοκομείο έγινε Σαββατοκύριακο ή από Ιανουάριο μέχρι Μάρτιο. Παρακάτω βλέπουμε και τις Shap Values για αυτά τα ευρύματα.

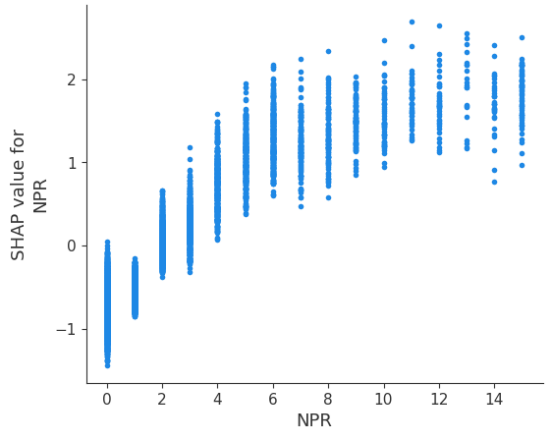


Εικόνα 62



Εικόνα 63

Ενδεικτικά και προς επίρρωση του F- Score βλέπουμε ακόμα ότι η αύξηση της τιμής NPR αυξάνει την θετική συνεισφορά της μεταβλητής στο τελικό output[39].



Εικόνα 64

4.4.Α Πρόβλεψη με XGBoost του κόστους θεραπείας, median (Binary Solution)[49]

Στην συνέχεια εξετάσαμε κατά πόσο θα μπορούσαμε να δημιούργήσουμε ένα ML model το οποίο να προβλέπει το κόστος για τον ασθενή, υπολογίζοντας τις προηγούμενες παραμέτρους. Στην περίπτωση αυτή το πρώτο πρόβλημα ήταν, όπως είδαμε και στην ανάλυση του SPSS υπάρχει πολύ μεγάλη τυπική απόκλιση, ενώ και το εύρος τιμών είναι επίσης εξαιρετικά μεγάλο. Επίσης υπάρχει πολύ μεγάλη λοξότητα (skewness), όπως υποδεικνύεται από την μεγάλη διαφορά του μέσου όρου με τον διάμεσο. Επιπλέον το 98,6% των τιμών ήταν από 100 ως 10^5 $. Αυτή η κατανομή δημιουργούσε προβλημα αξιοπιστίας στον Classifier, καθώς έδινε ένα τεράστιο Mean Square Error, καθιστώντας το μοντέλο αναξιόπιστο.

Με σκοπό να υπερβούμε τα παραπάνω προβλήματα, δημιουργήσαμε το μοντέλο μας σε τρία ξεχωριστά στάδια. Στην πρώτη περίπτωση κάναμη ένα binary classification, όπου όσα y είχαν τιμή άνω του διαμέσου παίρναν ως ψευδομεταβλητή την τιμή 1 ενώ όσα y είχαν τιμή κάτω του διαμέσου παίρναν ως ψευδομεταβλητή την τιμή 0.

Statistics

Total charges (cleaned)

N	Valid	698470
	Missing	14110
Mean		39560,41
Median		21194,00
Std. Deviation		74518,054
Range		4991588
Minimum		100
Maximum		4991688

Εικόνα 65

Ακολούθως, μετά από πολλούς συνδιασμούς τρέξαμε τον XGBoost Classifier με το παρακάτω Grid of hyperparameters, το οποίο έδινε το καλύτερο αποτέλεσμα:

```
paramGrid = {
    "learning_rate": [0.05],
    "n_estimators": [400],
    "max_depth": [6],
    'subsample': [0.8],
    "colsample_bylevel": [0.8],
    "colsample_bytree": [0.8],
    'gamma': [0.2]
}
```

Εικόνα 66

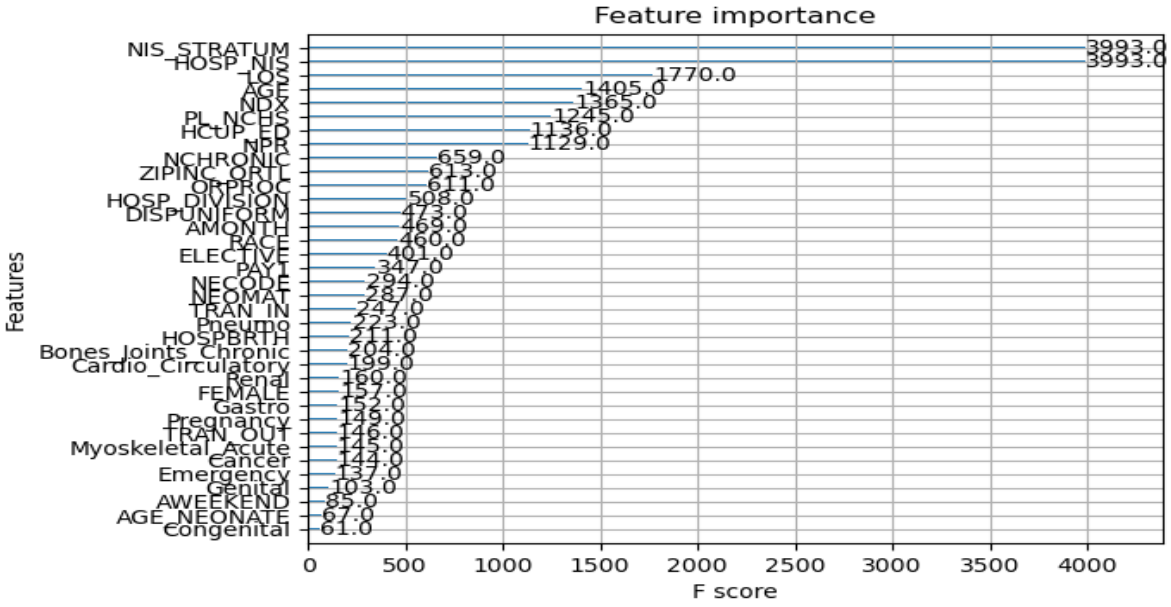
Ως βέλτιστη επιλογή προκρίθηκε με Best ROC AUC score στο training set: 0.95 ο ακόλουθος συνδιασμός υπερπαραμέτρων: Best parameters: {'colsample_bylevel': 0.8, 'colsample_bytree': 0.8, 'gamma': 0.2, 'learning_rate': 0.05, 'max_depth': 6, 'n_estimators': 400, 'subsample': 0.8}. Τα αποτελέσματα που πήραμε στο test set ήταν τα ακόλουθα: F1 Score: 0.8662, ROC-AUC: 0.8665, Accuracy: 0.8665, Precision: 0.8691, Recall: 0.8633 .

```
F1 Score: 0.8662
ROC-AUC: 0.8665
Accuracy: 0.8665
Precision: 0.8691
Recall: 0.8633
```

Εικόνα 67

Έχοντας υπόψη τα νούμερα που λάβαμε στο test set μπορούμε κατ' αρχάς να θεωρήσουμε ότι ο classifier που χρησιμοποιήσαμε, έχει ισχυρή προβλεπτική ισχύ, όσον αφορά την πρόβλεψη του κόστους και την κατάταξη του πάνω ή κάτω από την διάμεσο.

Όσον αφορά το κομμάτι επιμέρους μεταβλητών στο παρακάτω σχήμα μπορούμε να δούμε την συνεισφορά στην πρόβλεψη της κάθε μεταβλητής στην παρακάτω εικόνα:

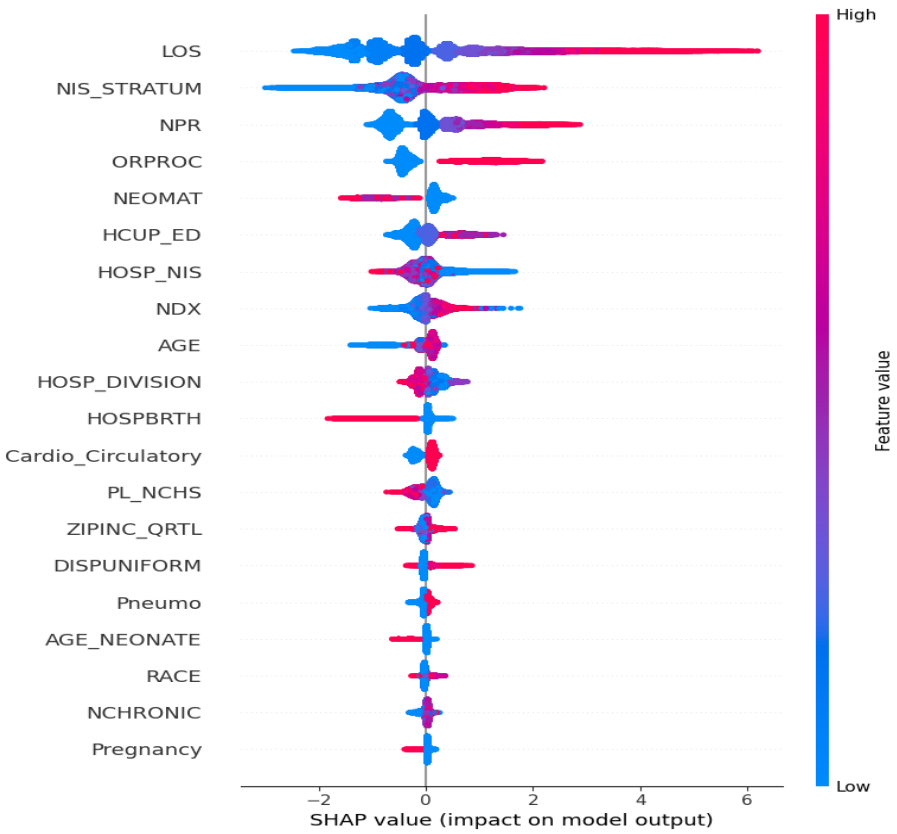


Εικόνα 68

Παρατηρούμε πως πρώτη στην συνεισφορά έρχεται η διαστρωμάτωση του νοσοκομείου NIS_STRATUM, ακολουθούμενη από το HOSP_NIS, το οποίο όπως αναφέρθηκε είναι ένας αριθμός ο οποίος σχετίζεται με την βαρύτητα του δείγματος των δεδομένων του κάθε τοπικού νοσοκομείου και τα οποία αποστέλονται στο NIS κατά το εξιτήριο του ασθενούς. Τρίτος παράγοντας είναι το LOS (Length of Stay), ενώ τέταρτος ακόλουθεί ο δείκτης NDX [14] που

αναφέρεται στον αριθμό των κωδικοποιημένων διαγνώσεων του κατά την διάρκεια της νοσηλείας του.

Στην συνέχεια εξετάσαμε τις SHap Values παίρνοντας το παρακάτω διάγραμμα.

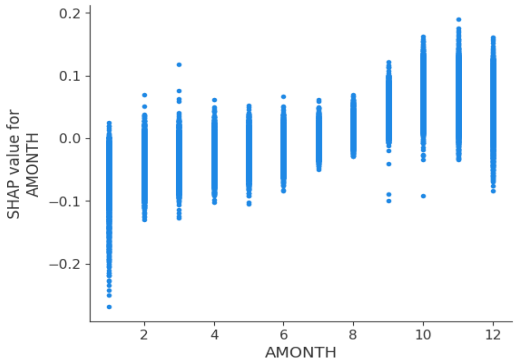


Εικόνα 69

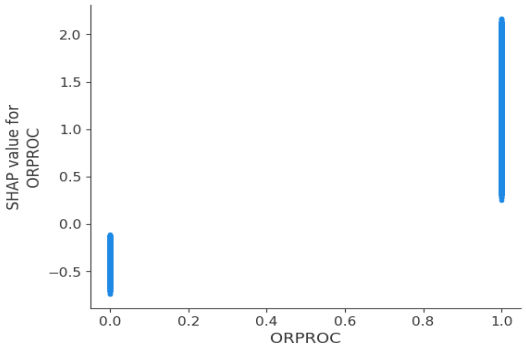
Η παρατήρηση των πρώτων τη τάξει Shap Values, έχει μεγάλη επεξηγηματικότητα. Εν προκειμένου υψηλές τιμές LOS (Length of Stay), NIS_STRATUM, ORPROC (Major operating room ICD-9-CM procedure indicator- 1 αν χρησιμοποιήθηκε 0 αν δεν χρησιμοποιήθηκε), NPR (Number of Procedures), δίνουν υψηλές SHAP Values (εικόνα 68).

Επίσης μελετήσαμε την την συνεισφορά της κάθε μεταβλητής Shapley Values ως προς τον εαυτό της και βλέπουμε πως όσο κινούμαστε προς το τέλος του χρόνου η SHap Value (εικόνα 70)

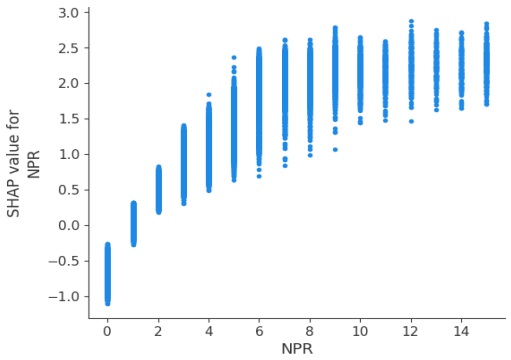
υποδηλώνει πως υπάρχει αυξητικός αντίκτυπος στο τελικό κόστος. Επίπλεον, χρήση operating room (Εικόνα 71) καθώς και Number of Procedures (Εικόνα 72) επιδρούν αυξητικά στο τελικό κόστος επίσης, φαινόμενο που κρινεται λογικό.



Εικόνα 70



Εικόνα 71



Εικόνα 72

4.4.Β Πρόβλεψη με XGBoost του κόστους θεραπείας >75% (Binary Solution)[50]

Στην συνέχεια εξετάσαμε κατά πόσο τοίδιο μοντέλο θα μπορούσε να λιτουργήσει προβλεπτικά, ώστε με βάση τις προηγούμενες παραμέτρους, να προβλέπει κατά πόσο οι συνολικές χρεώσεις για τον ενδονοσοκομειακό ασθενή βρίσκονται βρίσκονται το τελευταίο quartile, βρίσκονται στην περιοχή άνω του 75% των χρεώσεων, που αντιστοιχεί σε total charges ανω τον 44504$.

Και στην περίπτωση αυτό επιλέχθηκε το ακόλουθο grid υπερπαραμέτων:

```
paramGrid = {
    "learning_rate": [0.05],
    "n_estimators": [400],
    "max_depth": [6],
    'subsample': [0.8],
    "colsample_bylevel": [0.8],
    "colsample_bytree": [0.8],
    'gamma': [0.2]
}
```

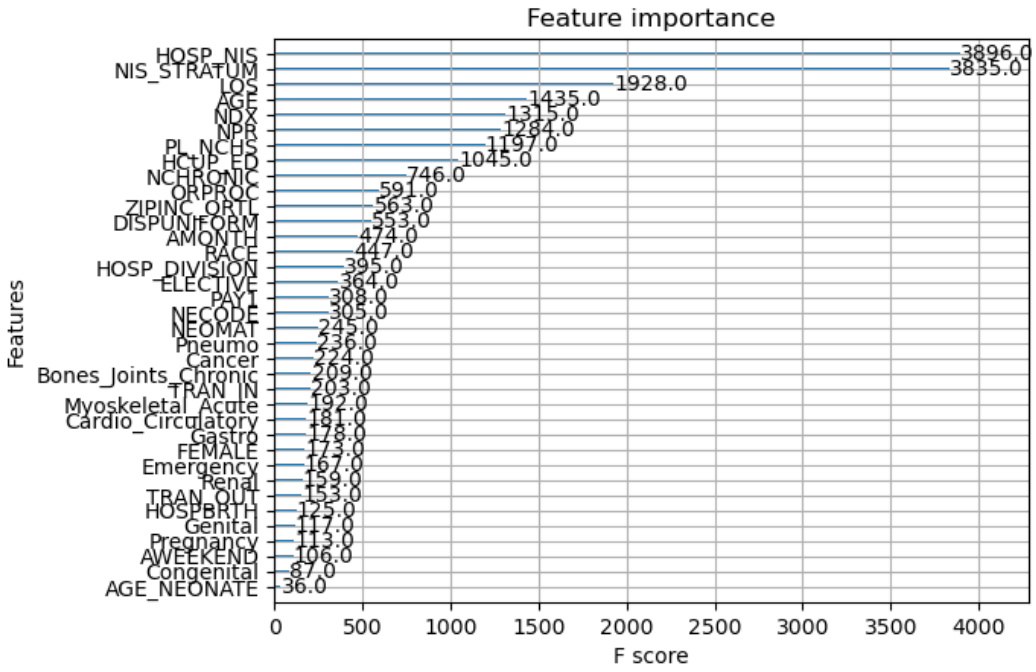
Εικόνα 73

Ως βέλτιστη επιλογή προκρίθηκε με Best ROC AUC score στο training set: 0.948 ο ακόλουθος συνδιασμός υπερπαραμέτρων: Best parameters: {'colsample_bylevel': 0.8, 'colsample_bytree': 0.8, 'gamma': 0.2, 'learning_rate': 0.05, 'max_depth': 6, 'n_estimators': 400, 'subsample': 0.8}. Τα αποτελέσματα που πήραμε στο test set ήταν τα ακόλουθα: F1 Score: 0.7791, ROC-AUC: 0.8445, Accuracy: 0.8943, Precision: 0.8168, Recall: 0.7447 .

```
F1 Score: 0.7791
ROC-AUC: 0.8445
Accuracy: 0.8943
Precision: 0.8168
Recall: 0.7447
```

Εικόνα 74

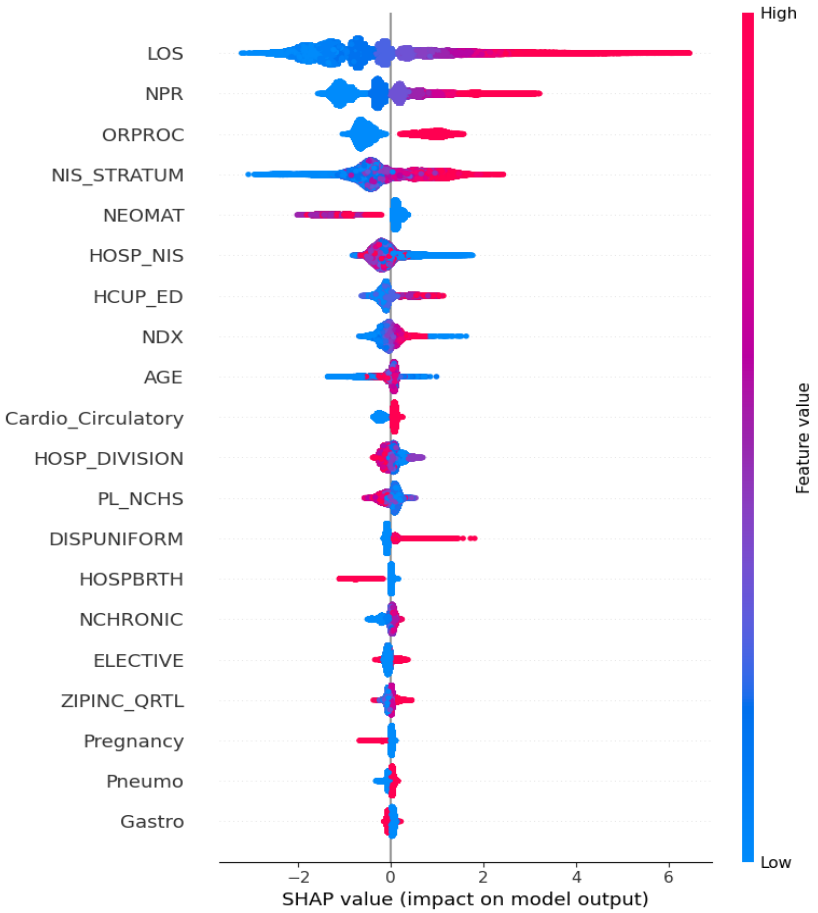
Όσον αφορά το κομμάτι επιμέρους μεταβλητών στο παρακάτω σχήμα μπορούμε να δούμε την συνεισφορά στην πρόβλεψη της κάθε μεταβλητής στην παρακάτω εικόνα:



Εικόνα 75

Παρατηρούμε πως πρώτη στην συνεισφορά έρχεται το HOSP_NIS, το οποίο όπως αναφέρθηκε είναι ένας αριθμός ο οποίος σχετίζεται με την βαρύτητα του δείγματος των δεδομένων του κάθε τοπικού νοσοκομείου που αποστέλονται στο NIS κατά το εξιτήριο του ασθενούς, δεύτερη η διαστρωμάτωση του νοσοκομείου NIS_STRATUM, ακολουθούμενη από τις ημέρες νοσηλείας LOS (Length of Stay), ενώ τέταρτος ακόλουθεί ο δείκτης NDX [14] που αναφέρεται στον αριθμό των κωδικοποιημένων διαγνώσεων του κατά την διάρκεια της νοσηλείας του και πέπμπτος ο αριθμός των procedures που έλαβαν χώρα ενδονοσοκομειακά.

Στο διάγραμμα SHap Values (Εικόνα 76) παρατηρούμε:



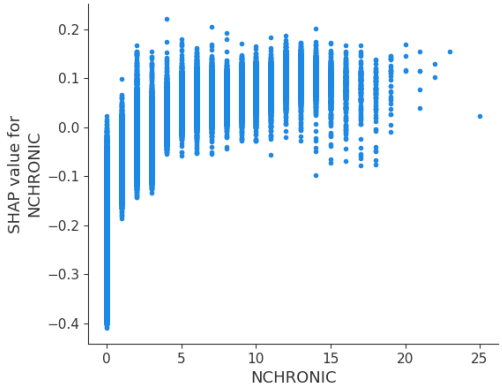
Εικόνα 76

Η παρατήρηση των πρώτων τη τάξει Shap Values, έχει μεγάλη επεξηγηματικότητα. Εν προκειμένου υψηλές τιμές LOS (Length of Stay), NPR (Number of Procedures), ORPROC (Major operating room ICD-9-CM procedure indicator- 1 αν χρησιμοποιήθηκε 0 αν δεν χρησιμοποιήθηκε),NIS_STRATUM, ORPROC (Major operating room ICD-9-CM procedure indicator- 1 αν χρησιμοποιήθηκε 0 αν δεν χρησιμοποιήθηκε), δίνουν υψηλές SHAP Values (εικόνα 79).

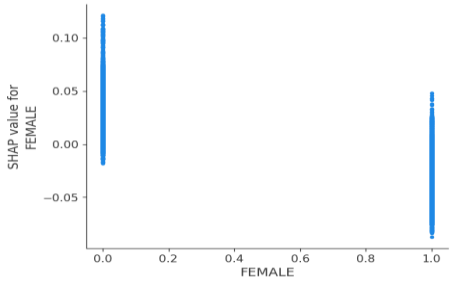
Μελετώντας την την συνεισφορά της κάθε μεταβλητής Shapley Values ως προς τον εαυτό της μπορούμε να ξεχωρήσουμε μερικά ενδιαφέροντα στοιχεία. Η αύξηση τουαριθμύ των χρόνιων ασθενειών αυξάνει την συνεισφορά στην κατηγοριοποίηση στην κλάση άνω του 75% του κόστους (Εικόνα 77). Το φύλο (Άντρας, γυναίκα), επηρεάζει επίσης την συνεισφορά στην

SHap Value. Στις μεν γυναίκες παρουσιάζεται σχετικά με αρνητική συνεισφορά, ενώ στους άντρες με ουδέτερη.

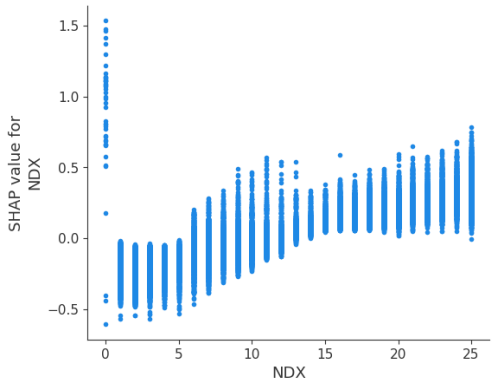
Τέλος (Εικόνα 79) η αύξηση του NDX (αριθμός καταγεγραμμένων ασθενειών κατά το εξιτήριο) επιδρά αυξητικά ώς προς την κατάταξη στην υπψηλά κατηγορία χρεώσεων, γεγονός που κρίνεται λογικό.



Εικόνα 77



Εικόνα 78



Εικόνα79

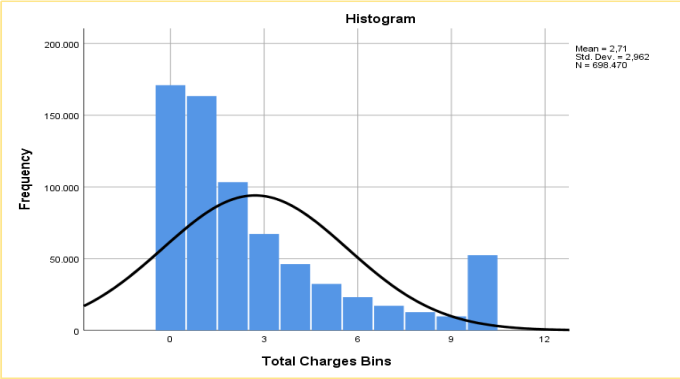
4.4.C Πρόβλεψη με XGBoost του κόστους θεραπείας (Multiclass Solution)[51]

Το τρίτο βήμα του ML μοντέλου μας, με την χρήση του XGBoost ήταν η δημιουργία bin ανά 10000$. Η λογική πίσω από την δημιουργία αυτών των bins εγκειτε στο γεγονός, ότι επι της ουσίας το 92,5% όλων των Total Charges (TOTCHG) βρίσκονται εντός αυτού του εύρους. Η κατανομή τους φαίνεται στις εικόνες 80 και 81.

Total Charges Bins

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	170936	24,0	24,5	24,5
	1	163365	22,9	23,4	47,9
	2	103342	14,5	14,8	62,7
	3	67207	9,4	9,6	72,3
	4	46169	6,5	6,6	78,9
	5	32351	4,5	4,6	83,5
	6	23144	3,2	3,3	86,8
	7	17119	2,4	2,5	89,3
	8	12709	1,8	1,8	91,1
	9	9739	1,4	1,4	92,5
	10	52389	7,4	7,5	100,0
	Total	698470	98,0	100,0	
Missing	System	14110	2,0		
Total		712580	100,0		

Εικόνα 80



Εικόνα 81

Επί της ουσίας μετατρέψαμε ένα πρόβλημα παλινδρόμισης - regression, σε ένα πρόβλημα κατηγοριοποίησης – multiclass classification.

Στην συνέχεια τρέξαμε τον XGBoost Classifier με το παρακάτω Grid of hyperparameters:

```
paramGrid = {
    "learning_rate": [0.05],
    "n_estimators": [400],
    "max_depth": [3],
    'subsample': [0.8],
    "colsample_bylevel": [0.8],
    "colsample_bytree": [0.8],
    'gamma': [0.2]
}
```

Εικόνα 82

Ως βέλτιστη επιλογή προκρίθηκε με Best ROC AUC score στο training set: 0.945 ο ακόλουθος συνδιασμός υπερπαραμέτρων: Best parameters: {'colsample_bylevel': 0.8, 'colsample_bytree': 0.8, 'gamma': 0.2, 'learning_rate': 0.05, 'max_depth': 3, 'n_estimators': 400, 'subsample': 0.8}. Τα αποτελέσματα που πήραμε στο test set ήταν τα ακόλουθα: Accuracy: 0.47 και το ακόλουθο Confusion Matrix:

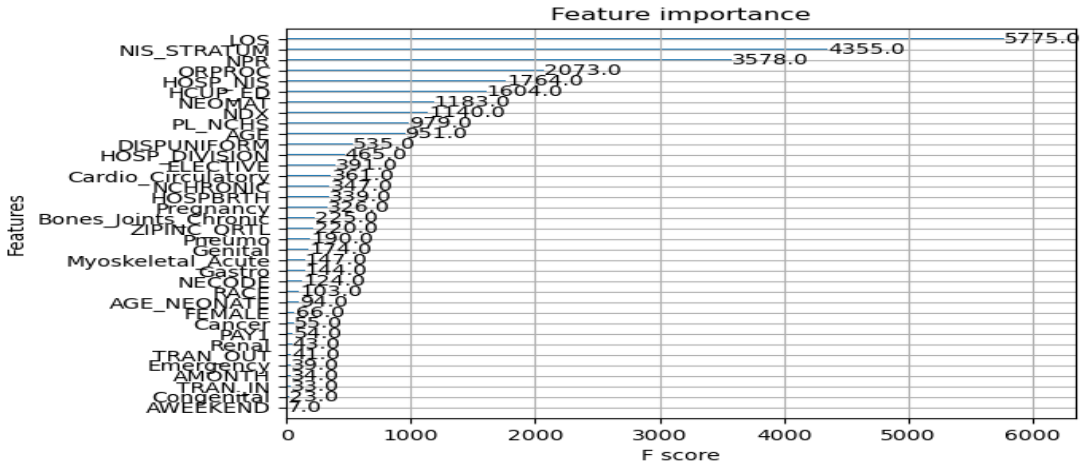
Actual Values (Rows) / Predicted Values (Columns)

```
Confusion Matrix:
[[27458  9513   376    88     4     0     0     0     0    42]
 [ 7087 25172  3578   636    51     1     0     0     0   243]
 [ 1247 13402  6292  1616   196     5     1     0     0   851]
 [  321  5508  5662  2123   305     4     0     0     0  1491]
 [  125  2218  3881  2103   385     9     0     0     0  1919]
 [   47   969  2326  1639   325    19     0     0     0  2155]
 [   21   475  1299  1134   212    10     0     0     0  2190]
 [   15   241   813   709   157     6     0     0     0  2061]
 [    8   135   446   485   122     5     0     0     0  1815]
 [   19   258   937   854   199     4     0     0     0 12204]]
```

Εικόνα 83

Έχοντας υπόψη τα νούμερα που λάβαμε στο test set μπορούμε κατ' αρχάς να θεωρήσουμε ότι ο classifier χρησιμοποιήσαμε, έχει μέτρια προβλεπτική ισχύ, όσον αφορά την πρόβλεψη του κόστους και την κατάταξη του στα 10 διαφορετικά bins. Αυτό συνάγεται από τους πολλούς αριθμούς εκτός της διαγωνίου, που σημαίνει ότι ενώ πολλές περιπτώσεις προβλέπονται σε μια κατηγορία (columns), στην πραγματικότητα ανήκουν σε άλλη κατηγορία.

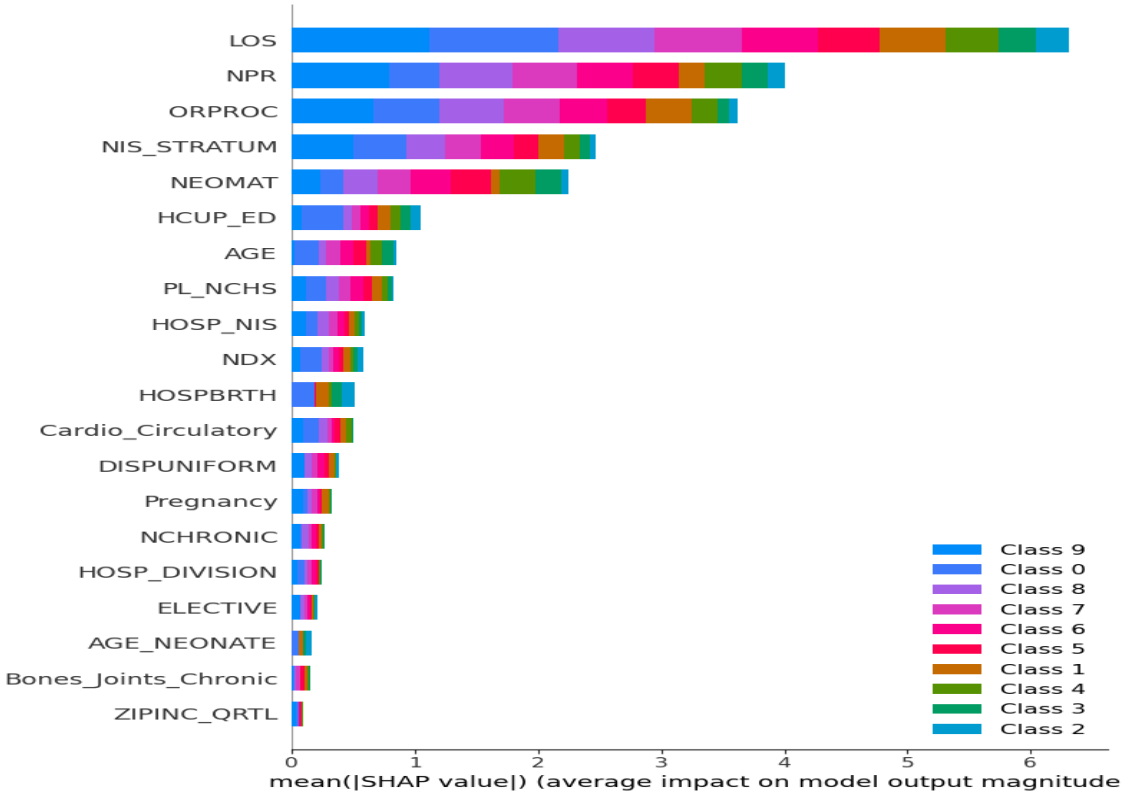
Όσον αφορά το κομμάτι επιμέρους μεταβλητών στο παρακάτω σχήμα μπορούμε να δούμε την συνεισφορά στην πρόβλεψη της κάθε μεταβλητής στην παρακάτω εικόνα:



Εικόνα 84

Βλέπουμε και εδώ ότι την μεγαλύτερη συνεισφορά κατέχει η διάρκεια παραμονής LOS (Length of Stay), με δεύτερη την διαστρωμάτωση του νοσοκομείου NIS_STRATUM. Τρίτη σε συνεισφορά μεταβλητή είναι ο αριθμός των διαδικασιών που πραγματοποιήθηκαν Number of Procedures, με τέταρτη την κατηγορία που αναφέρεται στην πραγματοποίηση ή όχι κάποιας Major Operation Procedure – ORPROC, και πέμπτο τον δείκτη βαρύτητας του δείγματος των δεδομένων του νοσοκομείου (HOSP_NIS), το οποίο όπως αναφέρθηκε αποδίδεται από το NIS.

Το SHap Values Diagramm στην εικόνα 76, μας δείχνει επιπροσθέτως τον βαθμό συνεισφοράς κάθε μεταβλητής, που στην περίπτωση μας είναι μόνο θετική, σε κάθε κλάση. Είναι εμφανές ότι όσο υψηλότερη η κλάση, τόσο το ποσό της συνεισφοράς σε απόλυτα νούμερα κάθε μεταβλητής αυξάνεται. Ταυτόχρονα φαίνεται ότι σε κάθε class το ποσοστό της συνεισφοράς κάθε μεταβλητής είναι διαφορετικό.



Εικόνα 85

Κεφάλαιο 5. Συμπεράσματα

Συμπερασματικά έχοντας μελετήσει τα δεδομένα μας τόσο με παραδοσιακές μεθόδους στατιστικές, αλλά και με την δημιουργία ενός μοντέλου πρόβλεψης που στηρίζεται σε AI μπορούμε να καταλήξουμε στα ακόλουθα.

Η παραδοσιακή προσέγγιση μέσω στατιστικών μεθόδων μας δίνει αρκετά καλή συσχέτιση - correlation μεταξύ των components της PCA και του Total Charges με adjusted R^2 = 46,44%, ενώ δίνει πολύ ασθενή correlation R^2 = 2,80% μεταξύ των components της PCA και της πιθανότητας επιβίωσης.

Με τη σειρά του το ML model που δημιουργήσαμε με classifier τον XGBoost, δίνει εξαιρετική προβλεπτική ισχύ όσον αφορά τη θνησιμότητα. F1 Score: 0.8865, ROC-AUC: 0.8852, Accuracy: 0.8849, Precision: 0.8624, Recall: 0.9119, καθώς εξαιρετική πρόβλεψη αν το κόστος θεραπείας θα υπερβεί την διάμεση τιμή ισχύ. F1 Score: 0.8662, ROC-AUC: 0.8665, Accuracy: 0.8665, Precision: 0.8691, Recall: 0.8633 .

Πολύ καλά επίσης ήταν και τα αποτελέσματα, σχετικά με το αν οι total charges του ενδοσοκομειακού ασθενή θα υπερβούν τα 44504$, αν θα κατηγοριοποιηθούν δηλαδή στο ανώτερο 75% των χρεώσεων. Οι τιμές που λάβαμε ήταν οι ακόλουθες: F1 Score: 0.7791, ROC-AUC: 0.8445, Accuracy: 0.8943, Precision: 0.8168, Recall: 0.7447

Τέλος μέτρια προβλεπτική ισχύ έδειξε το μοντέλο (όπως φαίνεται από το accuracy = 0.47 και το confusion matrix, εικόνα 83) αναφορικά με το σε ποια κατηγορία θα Total Charges θα κατηγοριοποιήσει τον ασθενή.

Βιβλιογραφία

[1] https://www.ukbiobank.ac.uk

[2] https://www.ukbiobank.ac.uk/enable-your-research/about-our-data

[3] https://biobank.ctsu.ox.ac.uk/crystal/browse.cgi?id=100039&cd=category

[4] UK Biobank Showcase User Guide

[5] UK Biobank Data Access Guide. Version 3.4. October 2023.

[6] https://choishingwan.gitlab.io/ukb-administration/pheno/data_manipulation/#format-of-data-fields-in-uk-biobank

[7] https://biobank.ndph.ox.ac.uk/showcase/cats.cgi

[8] https://biobank.ctsu.ox.ac.uk/crystal/field.cgi?id=4080

[9] https://biobank.ndph.ox.ac.uk/showcase/help.cgi?cd=data_field

[10] https://biobank.ndph.ox.ac.uk/showcase/help.cgi?cd=stability

[11] https://biobank.ndph.ox.ac.uk/showcase/help.cgi?cd=value_type

[12] https://biobank.ndph.ox.ac.uk/showcase/help.cgi?cd=item_type

]13] https://biobank.ndph.ox.ac.uk/showcase/help.cgi?cd=strata

[14] https://biobank.ndph.ox.ac.uk/showcase/help.cgi?cd=sexed

[15] https://biobank.ndph.ox.ac.uk/showcase/help.cgi?cd=instances

[16] https://biobank.ndph.ox.ac.uk/showcase/help.cgi?cd=array

[17] DNA Nexus UKB Research Analysis Platform Overview Webinar. August 2022. https://www.youtube.com/watch?v=8bcHeoEggBI

[18] https://dnanexus.gitbook.io/uk-biobank-rap/getting-started/working-with-ukb-data#bulk-data-files

[19] https://biobank.ndph.ox.ac.uk/showcase/download.cgi?tk=WPUCCYbol72eIDLL2ujSgghADPIEcler390620

[20] https://biobank.ndph.ox.ac.uk/showcase/refer.cgi?tk=p48rONVAYtPw1gZoGeINcff4JuVSVT6y390618&id=1758

[21] Jacob A. , Justine K. R. , Naomi A., Derrick W. C. , Daniel J. W., David H. W. and Anne Marie O'Connell5. Dynamic linkage of COVID-19 test results between Public Health England's Second Generation Surveillance System and UK Biobank.

[22] https://biobank.ndph.ox.ac.uk/showcase/rectab.cgi?tk=p48rONVAYtPw1gZoGeINcff4JuVSVT6y390618&id=1053

[23] https://biobank.ndph.ox.ac.uk/ukb/schema.cgi?id=17

[24] https://biobank.ndph.ox.ac.uk/ukb/schema.cgi?id=18

[25] https://biobank.ndph.ox.ac.uk/ukb/download.cgi

[26] https://dnanexus.gitbook.io/uk-biobank-rap/getting-started/working-with-ukb-data

[27] https://dnanexus.gitbook.io/uk-biobank-rap/working-on-the-research-analysis-platform/working-with-bulk-data-files

[28] https://platform.dnanexus.com/panx/projects

[29] https://documentation.dnanexus.com/developer/datasets

[30] https://dnanexus.gitbook.io/uk-biobank-rap/working-on-the-research-analysis-platform/accessing-phenotypic-data-as-a-file

[31] https://dnanexus.gitbook.io/uk-biobank-rap/working-on-the-research-analysis-platform/using-spark-to-analyze-tabular-data

[32] https://www.ukbiobank.ac.uk/enable-your-research/costs

[33] https://hcup-us.ahrq.gov/db/nation/nis/nisdde.jsp

[34] (https://hcup-us.ahrq.gov/HCUP_Overview/HCUP_Overview/index.html)

[35] (HCUP Methods Series. Calculating National Inpatient Sample (NIS) Variances for Data Years 2012 and Later. Report # 2015-09)

[36] (https://hcup-us.ahrq.gov/nisoverview.jsp#about)

[37] (Clinical Classifications Software (CCS) 2015. Issued March 2016. Healthcare Cost and Utilization Project – HCUP. A Federal-State-Industry Partnership in Health Data. Sponsored by the Agency for Healthcare Research and Quality)

[38] Kwon J-M, Jeon KH, Kim HM, et al. Deep-learning-based risk stratification formortality of patients with acute myocardial infarction. PLoS One 2019; 14:1–15.

[39] Constantine Tarabanis, MD, Evangelos Kalampokis, PhD, Mahmoud Khalil, MD, Carlos L. Alviar, MD, Larry A. Chinitz, MD, FHRS, Lior Jankelson, MD, PhD. Explainable SHAP-XGBoost models for in-hospital mortality after myocardial infarction.

[40] Li X, Liu H, Yang J, Xie G, Xu M, Yang Y. Using machine learning models to predict in-hospital mortality for st-elevation myocardial infarction patients. Stud Health Technol Inform 2017;245:476–480.

[41] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 2017;2017. Decem:4766–4775.

[42] Shapley LS. A value for n-person games. Contrib Theory Games 1953; 2:307–317.

[43] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, New York: Association for Computing Machinery; 2016. p. 785–794.

[44] Fizan Abdullah, Alodia Gabre-Kidan, Yiyi Zhang, Leilani Sharpe, David C. Chang. Report of 2,087,915 surgical admissions in U.S. Children: Inpatient mortality rates by procedure and specialty

[45] Augustine Manadan, Shilpa Arora , Millan Whittier, Ehizogie Edigin, Preeti Kansal. Patients admitted on weekends have higher in-hospital mortality than those admitted on weekdays: Analysis of national inpatient sample.

[46] Areti Karamanou, EvangelosKalampokis, KonstantinosTarabanis. Linked Open Government Data to Predict and Explain House Prices: The Case of Scottish Statistics Portal.

[47] https://hcup-us.ahrq.gov/db/vars/ndx/nisnote.jsp

[48] https://github.com/VasileiosVasileiou/Thesis-MBADS/blob/main/NIS_DIED_BINARY%20(1).ipynb

[49] https://github.com/VasileiosVasileiou/Thesis-MBADS/blob/main/NIS_TOTCHG_BINARY_MEDIAN%20(1).ipynb

[50] https://github.com/VasileiosVasileiou/Thesis-MBADS/blob/main/NIS_TOTCHG_BINARY_QUARTILE%20(1).ipynb

[51] https://github.com/VasileiosVasileiou/Thesis-MBADS/blob/main/NIS_TOTCHG_MULTICLASS%20(1).ipynb