Ανοικτό Πανεπιστήμιο Κύπρου

Σχολή Θετικών και Εφαρμοσμένων Επιστημών

Μεταπτυχιακή Διατριβή
Στην Ασφάλεια Υπολογιστών και Δικτύων

Ψευδωνυμοποίηση Δεδομένων Υγείας Για Ασφαλή Διαμοιρασμό Τους: Προκλήσεις Και Τεχνικές

Ιωάννης Δουλγεράκης

Επιβλέπων Καθηγητής
Κωνσταντίνος Λιμνιώτης

Δεκέμβριος 2023

Ανοικτό Πανεπιστήμιο Κύπρου

Σχολή Θετικών και Εφαρμοσμένων Επιστημών

Ψευδωνυμοποίηση Δεδομένων Υγείας Για Ασφαλή Διαμοιρασμό Τους: Προκλήσεις Και Τεχνικές

Ιωάννης Δουλγεράκης

Επιβλέπων Καθηγητής
Κωνσταντίνος Λιμνιώτης

Η παρούσα μεταπτυχιακή διατριβή υποβλήθηκε
προς μερική εκπλήρωση των απαιτήσεων για απόκτηση

μεταπτυχιακού τίτλου σπουδών
στην Ασφάλεια Υπολογιστών και Δικτύων

από τη Σχολή Θετικών και Εφαρμοσμένων Επιστημών
του Ανοικτού Πανεπιστημίου Κύπρου

Δεκέμβριος 2023

Περίληψη

Στις μέρες μας, λαμβάνοντας υπόψη τις τεχνολογικές εξελίξεις, αλλά και της ανάγκης συνδρομής τους κατά το μέγιστο στον τομέα της υγειονομικής περίθαλψης, τα ηλεκτρονικά δεδομένα υγείας αποτελούν ένα πολύ σημαντικό παράγοντα εκμετάλλευσης των τεχνολογιών αυτών. Η αυξημένη διαθεσιμότητα τέτοιων δεδομένων έχει φέρει επανάσταση στον τομέα της υγειονομικής περίθαλψης καθώς και την ιατρική έρευνα, προσφέροντας πλέον μεγάλες ευκαιρίες για επιστημονική εξέλιξη. Ωστόσο παρά τα οφέλη αυτά, η χρήση τέτοιων δεδομένων θα πρέπει να συνοδεύεται από αυστηρούς κανόνες και μέτρα προστασίας της ιδιωτικής ζωής καθώς και των ευαίσθητων πληροφοριών υγείας των ατόμων.

Με την έρευνα μας, στοχεύουμε να εξετάσουμε σε βάθος τους κανόνες αυτούς, αλλά και τρόπους με τους οποίους μπορούμε να αξιοποιήσουμε τέτοια δεδομένα με ασφαλή τρόπο για τις ανάγκες έρευνας, τηρώντας παράλληλα τις νομικές απαιτήσεις που έχει ορίσει ο Γενικός Κανονισμός Προστασίας Δεδομένων (GDPR).

Μέσα από τη διατριβή μας σκοπεύουμε να διερευνήσουμε μεθόδους, όπως η ψευδωνυμοποίηση και η ανωνυμοποίηση, που θα συνδράμουν στον ασφαλή διαμοιρασμό των δεδομένων υγείας για ερευνητικούς σκοπούς, τηρώντας τους κανονισμούς που έχουν οριστεί για τη προστασία των δεδομένων αυτών αλλά και των υποκειμένων τους. Ωστόσο πέρα από τα οφέλη των παραπάνω μεθόδων θα έρθουμε αντιμέτωποι και με αρκετές προκλήσεις και επιπτώσεις που σχετίζονται με τη διασφάλιση της ιδιωτικής ζωής, επιτρέποντας παράλληλα την εξαγωγή χρήσιμων πληροφοριών μέσα από τα δεδομένα αυτά.

Τα αποτελέσματα της διατριβής μας καταδεικνύουν τις δυσκολίες και τις προκλήσεις αυτές, που συνοδεύονται με την αξιοποίηση των δεδομένων υγείας με τρόπο τέτοιο ώστε να διατηρούν τη χρησιμότητά τους για τις ανάγκες έρευνας και παράλληλα να συμμορφώνονται πλήρως με τους Κανονισμούς του GDPR, προστατεύοντας έτσι τα υποκείμενα της έρευνας. Ωστόσο διαφαίνεται ότι, παρόλο που δεν υπάρχει κάποια λύση που να είναι εγγυημένα η βέλτιστη για όλες τις περιπτώσεις, απαιτείται σε κάθε περίπτωση μία συστηματική προσέγγιση προσανατολισμένη στις απαιτήσεις της εκάστοτε περίπτωσης προκειμένου να είναι σε θέση να αποδείξει ότι έχει λάβει τα κατάλληλα εχέγγυα αναφορικά με το διαμοιρασμό τέτοιων ευαίσθητων δεδομένων.

Summary

Taking into consideration the technological advancements in our days, and the need for their maximum contribution in the field of healthcare, electronic health data constitutes a crucial factor for leveraging these technologies. The increased availability of such data has revolutionized the healthcare sector and medical research, offering great opportunities for scientific progress. However, despite these benefits, the use of such data should be accompanied by strict rules and measures to protect individuals' privacy and sensitive health information.

Through our research, our aim is to thoroughly examine these rules and explore ways to utilize such data safely for research purposes while complying with the legal requirements set by the General Data Protection Regulation (GDPR).

In our thesis, we plan to investigate methods such as pseudonymization and anonymization that will assist in securely sharing health data for research purposes while adhering to regulations that safeguard this data and the individuals it pertains to. However, apart from the benefits of those methods, we will also face several challenges and implications related to ensuring privacy protection while still allowing the extraction of useful information from this data.

The results of our thesis demonstrate the difficulties that accompany the utilization of health data in a way that maintains their usefulness for research purposes while keeping full compliance with GDPR regulations, thus protecting the subjects of the research. However, it becomes apparent that although there is no one-size-fits-all solution guaranteed to be optimal for all cases, a systematic approach oriented towards the specific requirements of each case is required be able to demonstrate that the appropriate safeguards have been taken regarding the sharing of such sensitive data.

Ευχαριστίες

Με την ολοκλήρωση της Μεταπτυχιακής Διατριβής μου θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες, αρχικά στον επιβλέποντα Καθηγητή μου Δρ. Κωνσταντίνο Λιμνιώτη για τη συνολική συμβολή του στην εκπόνηση της, τη συνεχή καθοδήγησή του καθ΄ όλη τη διάρκεια της και την άμεση στήριξή του τόσο κατά τη διαδικασία τις έρευνας όσο και στη συγγραφή της.

Επίσης θα ήθελα να ευχαριστήσω όλους τους οικείους μου, τη γυναικά μου, οικογένεια και φίλους οι οποίοι συνέβαλαν με τον δικό τους τρόπο με υπομονή και άμεση συμπαράσταση στην επιτυχή ολοκλήρωση του έργου μου.

Τέλος ένα μεγάλο ευχαριστώ στη μικρή μου κορούλα που έρχεται στο κόσμο, όχι μόνο επειδή με ενεργοποίησε ώστε να ολοκληρώσω ταχύτερα το έργο μου, αλλά επειδή απλά έρχεται και ήδη μου έχει αλλάξει τη ζωή.

Περιεχόμενα

Περίληψη ii
Summary iii
Ευχαριστίες iv
Περιεχόμενα v
Εισαγωγή 1
1.1 Αντικείμενο Διατριβής 2
1.1.1 Σκοπός Της Έρευνας 2
1.1.2 Αναγκαιότητα Και Σπουδαιότητα Της Έρευνας 3
1.2 Ερευνητικά Ερωτήματα 3
1.3 Μεθοδολογία της Έρευνας 4
1.4 Δομή Διατριβής 5
Θεωρητικό Υπόβαθρο 8
2.1 Προσωπικά Δεδομένα Και Ιδιωτικότητα 8
2.1.1 Προσωπικά Δεδομένα 9
2.1.2 Ιδιωτικότητα 10
2.2 Νομικό Πλαίσιο 10
2.2.1 Προσωπικά Δεδομένα Κατά Τον GDPR 11
2.2.2 Στόχος Του GDPR 12
2.2.3 Προστασία της Ιδιωτικότητας Πριν Τον GDPR 22
2.3 Κανονισμός Ευρωπαϊκού Χώρου Δεδομένων Υγείας 23
2.4 Ψευδωνυμοποίηση 24
2.4.1 Ορισμός Της Ψευδωνυμοποίησης 24
2.4.2 Αναγκαιότητα και Οφέλη Της Ψευδωνυμοποίησης 25
2.4.3 Αναγνωριστικά Γνωρίσματα 26
2.4.4 Μοντέλα Και Τεχνικές Επιθέσεων 26
2.4.5 Τεχνικές Ψευδωνυμοποίησης 28
2.4.6 Πολιτικές Ψευδωνυμοποίησης 32
2.4.7 Ανάκτηση Δεδομένων 33
2.5 Ανωνυμοποίηση 34
2.5.1 Ορισμός Της Ανωνυμοποίησης 35
2.5.2 Τύποι Γνωρισμάτων 35
2.5.3 Αποκάλυψη Πληροφοριών 37

2.5.4 Τύποι Και Μοντέλα Επιθέσεων 38
2.5.5 Μέθοδοι Ανωνυμοποίησης 40
2.5.6 Χρησιμότητα Έναντι Απώλειας 41
2.5.7 Μοντέλα Ανωνυμοποίησης 41
2.6 Σύνοψη 44
Βιβλιογραφική Ανασκόπηση 45
3.1 Εφαρμογές Της Ψευδωνυμοποίησης 46
3.1.1 Προηγμένες Τεχνικές Ψευδωνυμοποίησης 46
3.1.2 Ψευδωνυμοποίηση Στην Υγεία 50
3.1.3 Άλλες Περιπτώσεις Εφαρμογής Της Ψευδωνυμοποίησης 57
3.2 Εφαρμογές Ανωνυμοποίησης 61
3.2.1 Ανωνυμοποίησης Δεδομένων Υγείας 62
3.3 Αξιοποίηση Και Προστασία Δεδομένων Υγείας 64
3.3.1 Κανονισμός Ευρωπαϊκού Χώρου Δεδομένων Υγείας 64
3.3.2 Νόμος HIPAA 67
Συλλογή Δεδομένων 69
4.1 Είδη Δεδομένων 69
4.1.1 Πραγματικά Δεδομένα Υγείας 69
4.1.2 Υπάρχοντα Μη Πραγματικά Δεδομένα Υγείας 71
4.1.3 Δημιουργία Ρεαλιστικών Δεδομένων Υγείας 72
4.2 Συλλογή Δεδομένων 72
4.2.1 Επιλογή Μεθόδου 73
4.2.2 Ποιότητα Και Μορφή Δεδομένων 74
4.2.3 Διαθέσιμα Εργαλεία Παραγωγής Δεδομένων 75
4.2.3 Επιλογή Εργαλείου 77
Μελέτη Περίπτωσης 78
5.1 Μελέτη 78
5.1.1 Ανάγκη Αντιμετώπιση Της 79
5.1.2 Αναγκαιότητα Προστασίας Των Δεδομένων 79
5.2 Παραγωγή Των Δεδομένων 80
5.2.1 Διαμόρφωση Των Δεδομένων 80
5.2.2 Δημογραφικά Στοιχεία 81
5.2.2 Ευαίσθητα Πεδία 83
5.2.3 Μη Εμπιστευτικά Πεδία 85

5.2.4 Πρόσθετα πεδία αναφορικά με τη νοσηλεία ... 85
5.2.5 Πίνακας Γνωρισμάτων ... 86
Ψευδωνυμοποίηση - Προσέγγιση ... 88
6.1 Επιλογή Μεθόδου Ψευδωνυμοποίησης ... 88
6.1.1 Διερεύνηση Μεθόδων ... 89
6.1.2 Επιλογή Μεθόδου ... 91
6.2 Εφαρμογή Της Ψευδωνυμοποίησης ... 92
6.2.1 Ανάλυση Υλοποίησης ... 93
6.2.2 Υλοποίηση Της Ψευδωνυμοποίησης ... 94
6.3 Σύνοψη ... 97
Εφαρμογή τεχνικών ανωνυμοποίησης - Προσέγγιση ... 99
7.1 Μοντέλα Απορρήτου Και Μέθοδοι Ανωνυμοποίησης ... 99
7.1.1 Μοντέλα Απορρήτου ... 100
7.1.2 Μέθοδοι Ανωνυμοποίησης ... 100
7.2 Προσέγγιση ... 101
7.2.1 Κατηγοριοποίηση Γνωρισμάτων ... 101
7.2.2 Διαδικασία Ανωνυμοποίησης ... 102
7.2.3 Ανάλυση Αποτελεσμάτων ... 103
Επιλογή Εργαλείου ... 104
8.1 ARX ... 104
8.1.1 Προσέγγιση ... 105
8.1.2 Μοντέλα Απορρήτου ... 105
8.1.3 Μέθοδοι Ανωνυμοποίησης ... 106
8.1.4 Μοντέλα Ποιότητας/Χρησιμότητας ... 106
8.1.5 Ανάλυση Χρησιμότητας Και Ρίσκου ... 106
8.2 Amnesia ... 107
8.2.1 Μέθοδοι Ανωνυμοποίησης ... 107
8.2.2 Μοντέλα Απορρήτου ... 107
8.2.3 Ανάλυση ... 108
8.3 PyCanon ... 108
8.3.1 Προσέγγιση ... 108
8.3.2 Μοντέλα Απορρήτου ... 108
8.3.3 Ανάλυση ... 108
8.4 Σύνοψη ... 109

Ανωνυμοποίηση - Εφαρμογή **110**
9.1 Προετοιμασία Δεδομένων 110
9.1.1 Εισαγωγή Δεδομένων 111
9.1.2 Δημιουργία Ιεραρχιών 113
9.1.3 Διατήρηση Χρησιμότητας Δεδομένων 115
9.2 Ανωνυμοποίηση Δεδομένων 117
9.2.1 Προσεγγίσεις Ανωνυμοποίησης 117
9.3 Ανάλυση Αποτελεσμάτων 137
9.3.1 Ανάλυση Χρησιμότητας 138
9.3.2 Ανάλυση Ρίσκου 143
9.3.3 Επιλογή Προσέγγισης 145
9.4 Σύνοψη 148
Επίλογος **150**
10.1 Ανασκόπηση 150
10.2 Συμπεράσματα 152
10.3 Προκλήσεις Και Περιορισμοί 153
10.4 Θέματα Προς Μελλοντική Έρευνα 154
Βιβλιογραφία 156
Δεδομένα Υγείας **1**
Α.1 Δημιουργία Ρεαλιστικών δεδομένων 1
Ψευδωνυμοποίηση Δεδομένων **1**
Β.1 Κώδικας Ψευδωνυμοποίησης 1
Αποτελέσματα Ανωνυμοποίησης **1**
Γ.1 Γνωρίσματα Δεδομένων Υγείας 1
Γ.1 Προσεγγίσεις Ανωνυμοποίησης 2

Κεφάλαιο 1
Εισαγωγή

Σήμερα, τα δεδομένα υγείας αποτελούν μία από τις πιο ευαίσθητες κατηγορίες δεδομένων. Οι επιστήμονες, οι ερευνητές και οι επαγγελματίες της υγείας συλλέγουν συνεχώς δεδομένα υγείας από ασθενείς, με σκοπό τη παροχή της βέλτιστης φροντίδας υγείας καθώς και την εκτέλεση ιατρικών ερευνών. Σε καθημερινή βάση μεγάλος όγκος δεδομένων υγείας διανέμεται ανά το κόσμο μέσω διαφόρων ιδρυμάτων -δεδομένα τα οποία αφορούν τους ίδιους τους ασθενείς και πρέπει να φυλαχθούν και να διανεμηθούν με ασφάλεια μεταξύ των ενδιαφερομένων. Ωστόσο από τα δεδομένα αυτά, πέραν της αναγκαιότητάς τους για σκοπούς παροχής υπηρεσιών υγείας αλλά και για τον καθορισμό πολιτικών αναφορικά με τη Δημόσια Υγεία, προκύπτουν εξαιρετικά ωφέλιμα ευρήματα που μπορούν να χρησιμοποιηθούν για ερευνητικούς σκοπούς, για βελτίωση των Συστημάτων διαχείρισης Υγείας, πρόληψη για πιθανότητα εμφάνισης επιδημιών και άλλους σκοπούς. Όμως στη περίπτωση χρήσης τους πέραν του προβλεπόμενου προκύπτουν κάποιοι περιορισμοί όσον αφορά τον τρόπο εκμετάλλευσής τους, οι οποίοι πηγάζουν από το γεγονός ότι θεμελιώδη δικαιώματα και ελευθερίες μπορεί να θίγονται κατά τη χρήση των δεδομένων αυτών. Στην Ευρώπη, ο Κανονισμός Προστασίας Προσωπικών Δεδομένων (GDPR) ορίζει τα δεδομένα αυτά ως ευαίσθητα προσωπικά και η χρήση τους δεν μπορεί να γίνει αλόγιστα, αλλά ακολουθώντας κάποια πρότυπα και κάποιους κανόνες καθ' όλη την αξιοποίησή τους. Κανόνες που πρέπει να τηρηθούν με αυστηρότητα, καθώς η λανθασμένη χρήση τους μπορεί να έχει

καταστροφικές επιπτώσεις τόσο σε οικονομικό επίπεδο όσο και στη καταστροφή τους και σε κάποιες περιπτώσεις να έχουν αντίκτυπο και στους ίδιους τους ασθενείς [1].

Η Ευρωπαϊκή Ένωση παρουσίασε πολύ πρόσφατα μία πρόταση Κανονισμού αναφορικά με τα δεδομένα υγείας και τον διαμοιρασμό τους μεταξύ εμπλεκόμενων φορέων. Αν και η πρόταση αυτή δεν είναι οριστική, διαφαίνεται ότι επέρχονται πλατφόρμες για ασφαλή ανταλλαγή δεδομένων, με ανάγκη ύπαρξης κατάλληλων εχέγγυων για την προάσπιση θεμελιωδών δικαιωμάτων, όπως η ιδιωτικότητα και η προστασία προσωπικών δεδομένων. Αυτό ταυτόχρονα εισάγει την ανάγκη διερεύνησης αποτελεσματικών τεχνολογιών που να παρέχουν αυτά τα εχέγγυα [2].

1.1 Αντικείμενο Διατριβής

Αντικείμενο της παρούσας διατριβής είναι η μελέτη των τεχνολογιών καθώς και τεχνικών που χρησιμοποιούνται και μπορούμε να εκμεταλλευτούμε για ασφαλή διαμοιρασμό των δεδομένων υγείας μεταξύ οργανισμών και εμπλεκόμενων φορέων χωρίς να απειλείται η ιδιωτικότητα και άλλα ανθρώπινα δικαιώματα, παράγοντας παράλληλα χρήσιμη για τους σκοπούς της πληροφορία. Επίσης θα εξεταστούν οι προκλήσεις που θα συναντήσουμε κατά τη προσπάθεια αυτή και πως μπορούν να αντιμετωπιστούν σήμερα παρέχοντας έτσι ένα βέλτιστο αποτέλεσμα.

1.1.1 Σκοπός Της Έρευνας

Μέσω της έρευνας που θα πραγματοποιηθεί στοχεύουμε στην εύρεση τρόπων διάθεσης των δεδομένων υγείας για ερευνητικούς σκοπούς με τρόπο τέτοιο που θα παρέχουν χρήσιμη πληροφορία για τις ανάγκες της έρευνας αυτής, χωρίς όμως να θίγονται τα θεμελιώδη ανθρώπινα δικαιώματα [3]. Η διαδικασία αυτή θα πρέπει να παραμένει σε πλήρη συμμόρφωση με τους κανονισμούς των νομοθεσιών και πιο συγκεκριμένα του GDPR.

Πιο συγκεκριμένα θα πρέπει τα δεδομένα αυτά να μπορούν να διατεθούν, χωρίς όμως να αποκαλύπτονται περισσότερα στοιχεία από όσα είναι απολύτως απαραίτητα για τους ερευνητικούς σκοπούς και ταυτόχρονα να διατηρήσουν χρήσιμη για την επίτευξη των σκοπών αυτών πληροφορία. Πάνω σε αυτό το πεδίο θα ερευνηθούν και θα χρησιμοποιηθούν τεχνικές ψευδωνυμοποίησης και ανωνυμοποίησης έχοντας ως βάση αναφοράς τις απαιτήσεις που ορίζει ο GDPR. Με την έρευνα που θα ακολουθήσει δεν στοχεύουμε στο να καλύψουμε πλήρως όλες τις δυνατές προσεγγίσεις της ανωνυμοποίησης και της ψευδωνυμοποίησης καθώς αυτό θα

ξεπερνούσε κατά πολύ το πλαίσιο του στόχου μας, αλλά σκοπεύουμε να επικεντρωθούμε σε ένα συγκεκριμένο πεδίο δεδομένων όπως αυτά της υγείας.

1.1.2 Αναγκαιότητα Και Σπουδαιότητα Της Έρευνας

Έχει καταστεί πλέον σαφές ότι υπάρχει μεγάλη αναγκαιότητα διανομής των δεδομένων υγείας καθώς μπορούν να αποδειχτούν πάρα πολύ ωφέλιμα, με γνώμονα πάντα να μη διαρρεύσουν περισσότερες πληροφορίες από τις ελάχιστα αναγκαίες και να τηρούνται οι κανονισμοί απορρήτου όπως ορίζονται από τον GDPR.

Με τα ευρήματά μας στοχεύουμε να βοηθήσουμε την επιστημονική κοινότητα και φορείς διαχείρισης δεδομένων υγείας με κατάλληλες προσεγγίσεις εφαρμογής της ανωνυμοποίησης και της ψευδωνυμοποίησης τέτοιου είδους δεδομένων, καθώς και τους κινδύνους που ελλοχεύουν κατά τη δημοσίευση τους, όπως και τις συνέπειες που μπορούν να προκύψουν κατά την αλόγιστη χρήση τους.

1.2 Ερευνητικά Ερωτήματα

Λαμβάνοντας υπόψη το κίνητρο και τις ανάγκες της παρούσας έρευνας για το διαμοιρασμό χρήσιμων δεδομένων υγείας, παρουσιάζοντας έτσι μία πλήρη έρευνα των τεχνικών καθώς και των προκλήσεων που θα αντιμετωπίσουμε, υπό τις αυστηρές απαιτήσεις του GDPR, γεννιέται ένα καίριο ερώτημα:

Πώς θα μπορούσαν να εφαρμοστούν οι υπάρχουσες προσεγγίσεις ανωνυμοποίησης και ψευδωνυμοποίησης σε δεδομένα υγείας ακολουθώντας μία επαρκή συμμόρφωση με τις απαιτήσεις των κανονισμών και των νομοθεσιών, προστατεύοντας την ιδιωτικότητα και τα προσωπικά δεδομένα;

Αναλύοντας την ανησυχία αυτή προκύπτουν τέσσερα βασικά ερευνητικά ερωτήματα που βοηθούν να επικεντρωθούμε σε συγκεκριμένες πτυχές της έρευνας αυτής:

1. Θα μπορούσαν να λειτουργήσουν αποδοτικά οι τεχνικές ανωνυμοποίησης και ψευδωνυμοποίησης των δεδομένων υγείας;

2. Τι ειδικότερες απαιτήσεις προκύπτουν για τη ψευδωνυμοποίηση των δεδομένων; Μήπως κάθε τεχνική ψευδωνυμοποίησης πρέπει να συνδυάζεται και με τεχνικές ανωνυμοποίησης;

3. Η εφαρμογή τέτοιων τεχνικών διατηρεί τα δεδομένα σε μορφή τέτοια ώστε να παράγουν χρήσιμη, για τους εκάστοτε επιδιωκόμενους σκοπούς, πληροφορία;

4. Θα καλύπτουν τη συμμόρφωση με βάση τον GDPR;

1.3 Μεθοδολογία της Έρευνας

Στη προσπάθεια να προσεγγίσουμε αποτελεσματικότερα την έρευνά μας, αρχικά εξετάζουμε τις υπάρχουσες νομοθεσίες περί της προστασίας των προσωπικών δεδομένων και της ιδιωτικότητας, όπως και προσεγγίσεις ψευδωνυμοποίησης και ανωνυμοποίησης, το πώς έχουν χρησιμοποιηθεί έως τώρα σε προηγούμενες έρευνες και το πως έχουν εξελιχθεί με τα χρόνια. Θα εξετάσουμε κατά πόσο τα δεδομένα ανωνυμοποιούνται επαρκώς και αν το όφελος των ευρημάτων υπερβαίνει της απώλειας πληροφορίας, παράγοντας έτσι χρήσιμα προς μελέτη αποτελέσματα. Η όλη διαδικασία θα πρέπει να έρχεται σε πλήρη συμμόρφωση με τον GDPR.

Το επόμενο βήμα είναι να εστιάσουμε σε μία μελέτη περίπτωσης, «συλλέγοντας» τα δεδομένα αυτά για τους ερευνητικούς μας σκοπούς προσεγγίζοντας έτσι ένα ρεαλιστικό αποτέλεσμα. Θα εξετάσουμε τα οφέλη αλλά και τα μειονεκτήματα διάφορων τεχνικών συλλογής ρεαλιστικών δεδομένων όπως η χρήση πραγματικών δεδομένων υγείας, η παραγωγή ρεαλιστικών μη πραγματικών δεδομένων με κατάλληλα εργαλεία και η χρήση ήδη υπαρχόντων μη πραγματικών δεδομένων. Θα δούμε τα ρίσκα που εμφανίζονται κατά τη χρήση πραγματικών δεδομένων υγείας καθώς και τις προκλήσεις που εμφανίζονται κατά τη συλλογή τους, αλλά και την ανάγκη δημιουργίας ρεαλιστικών δεδομένων προς έρευνα. Ο όγκος των δεδομένων σε μία προσπάθεια προσέγγισης μίας αποτελεσματικής έρευνας θα πρέπει να είναι ικανοποιητικά μεγάλος καλύπτοντας έτσι ένα ρεαλιστικό σενάριο.

Στη συνέχεια έχοντας συλλέξει τα δεδομένα υγείας που χρειαζόμαστε θα διερευνηθεί σε τι βαθμό μία επιτυχής ψευδωνυμοποίηση θα πρέπει ταυτόχρονα να συνδυαστεί με τεχνικές ανωνυμοποίησης. Θα εξεταστούν κάποιες βασικές προσεγγίσεις ψευδωνυμοποίησης δεδομένων

και τα οφέλη που μπορούν να μας δώσουν, υιοθετώντας τεχνικές που έχουν προταθεί σε πρόσφατες αναφορές του Ευρωπαϊκού Οργανισμού Κυβερνοασφάλειας (ENISA).

Στο σημείο αυτό γεννιέται το ερώτημα κατά πόσο μια πολύ καλή προσέγγιση ψευδωνυμοποίησης μπορεί να αποτρέψει πλήρως την αποκάλυψη ευαίσθητης πληροφορίας και καταπάτησης της ιδιωτικότητας καθώς και τη δύναμη που έχουν κάποιες μη αναγνωριστικές χαρακτηριστικές πληροφορίες συνδυαστικά, θέτοντας και την ανωνυμοποίηση των δεδομένων απολύτως απαραίτητη. Έτσι θα εξεταστούν και τεχνικές ανωνυμοποίησης γενικεύοντας και αποκρύπτοντας κάποιες πληροφορίες, ευελπιστώντας να πετύχουμε τον αποτελεσματικότερο δυνατό συνδυασμό τους.

Συνεχίζοντας θα ερευνήσουμε κάποια από τα υπάρχοντα ελεύθερα διαθέσιμα εργαλεία ανωνυμοποίησης για να προχωρήσουμε στη διαδικασία αυτή, εξετάζοντας τις βασικές λειτουργίες που μας προσφέρει το κάθε εργαλείο, σε μία προσπάθεια να καλύψει κατά το βέλτιστο τις ανάγκες μας και τους στόχους που έχουμε θέσει.

Έχοντας έτσι συλλέξει τα δεδομένα μας, έχουμε επιλέξει τις μεθόδους ανωνυμοποίησης και ψευδωνυμοποίησης που θα εκμεταλλευτούμε και τα εργαλεία που θα χρησιμοποιήσουμε, μπορούμε να ξεκινήσουμε την υλοποίηση του έργου μας θέτοντας σε εφαρμογή πλέον τα έως τώρα ευρήματα και τις προκλήσεις που συναντήσαμε κατά τη διαδικασία της έρευνας. Αναμένουμε πως διαφορετικές προσεγγίσεις θα δώσουν διαφορετικά αποτελέσματα, αναζητώντας έτσι τη χρυσή τομή μεταξύ της απώλειας πληροφορίας στα δεδομένα μας έναντι του ρίσκου αναγνώρισης κάποιου προσώπου μέσα σε αυτά. Κατά την ανάλυση των δεδομένων αυτών, κρίνεται σημαντικό να καταγράφονται τα αποτελέσματα καθ' όλη τη διαδικασία, τόσο για τη σύγκριση μεταξύ διαφορετικών προσεγγίσεων όσο και για το πως οι τεχνικές αυτές συμμορφώνονται με τους κανονισμούς του GDPR.

Τέλος θα καταγράψουμε τα εμπειρικά δεδομένα που πήραμε και θα εξετάσουμε την αποτελεσματικότητα της έρευνας μας. Θα είμαστε σε θέση να έχουμε μία πλήρη εικόνα της διαδικασίας και των αποτελεσμάτων που λάβαμε, το κόστος έναντι του οφέλους και της απώλειας δεδομένων έναντι της προσπάθειας αποφυγής της αναγνώρισης και της ταυτοποίησης ευαίσθητων δεδομένων, όπως αυτών της υγείας, με φυσικά πρόσωπα.

1.4 Δομή Διατριβής

Το κεφάλαιο 2 της παρούσας Μεταπτυχιακής Διατριβής αποτελεί το θεωρητικό υπόβαθρο της έρευνας που θα ακολουθήσουμε και γίνεται μία εκτενής αναφορά στην ιδιωτικότητα ως θεμελιώδες δικαίωμα όπως και τα προσωπικά δεδομένα, τις διαδικασίες ανωνυμοποίησης και ψευδωνυμοποίηση όπως τους τύπους, τις μεθόδους και τις «βασικές» τεχνικές που χρησιμοποιούνται καθώς και αναφορές στους κανονισμούς του GDPR σε σχέση με τα προσωπικά δεδομένα, τους περιορισμούς που έχει θέσει και τους τρόπους διαχείρισης ευαίσθητων δεδομένων.

Στο κεφάλαιο 3 που είναι η βιβλιογραφική ανασκόπηση θα εστιάσουμε περισσότερο σε παλαιότερες έρευνες ψευδωνυμοποίησης και ανωνυμοποίησης σχετικά με τα δεδομένα υγείας και πώς τα έχουν διαχειριστεί και προσαρμόσει για τις ανάγκες έρευνας. Επίσης θα εξετάσουμε βαθύτερα κάποιες πιο προηγμένες πρακτικές ψευδωνυμοποίησης, ως προσπάθειες βελτίωσης των «κλασικών» εκδοχών τους, καθώς και περαιτέρω κανονισμούς σχετικά με την ασφαλή αξιοποίηση των δεδομένων υγείας, όπως η πρόταση για τον Ευρωπαϊκό Χώρο Δεδομένων Υγείας καθώς και το Νόμο περί Φορητότητας και Λογοδοσίας για την Ασφάλιση Υγείας.

Το κεφάλαιο 4 περιγράφει την ανάπτυξη του μεθοδολογικού πλαισίου το οποίο αξιοποιήσαμε για την έρευνά μας. Ειδικότερα, θα αναλυθούν οι τρόποι με τους οποίους μπορούμε να δημιουργήσουμε ή να συλλέξουμε δεδομένα υγείας, τα πλεονεκτήματα και τα μειονεκτήματα που έχει κάθε τεχνική και η επιλογή της μεθόδου συλλογής των δεδομένων υγείας προς έρευνα. Θα εξεταστούν τα εργαλεία που χρησιμοποιήθηκαν δημιουργώντας έτσι ένα σύνολο ρεαλιστικών δεδομένων υγείας.

Το κεφάλαιο 5 αφορά τη μελέτη περίπτωσης πάνω στην οποία θα εφαρμόσουμε την έρευνά μας αναλύοντας περαιτέρω και τα δεδομένα που δημιουργήσαμε. Θα αναλυθούν επίσης οι πληροφορίες που συμπεριλάβαμε στα δεδομένα αυτά ώστε να δημιουργήσουμε ένα ρεαλιστικό σενάριο ενός συνόλου νοσοκομειακών ασθενών.

Στο κεφάλαιο 6 θα προσεγγίσουμε και θα εφαρμόσουμε τις τεχνικές της ψευδωνυμοποίησης επιλέγοντας μία κατάλληλη μέθοδο που θα ψευδωνυμοποιήσει τα δεδομένα μας με αποτελεσματικότερο τρόπο με στόχο την ασφαλή διανομή τους εντός του νοσοκομείου από τα ενδιαφερόμενα τμήματα και κλινικές.

Στο κεφάλαιο 7 θα εστιάσουμε στην γενικότερη προσέγγιση της ανωνυμοποίησης που θα εφαρμόσουμε σε επόμενο κεφάλαιο. Θα οριστούν στόχοι καθώς και τεχνικές που θα

ακολουθηθούν κατά τη διαδικασία της ανωνυμοποίησης ώστε να καταστήσουν όσο το δυνατό δυσκολότερη τη πιθανότητα αναγνώρισης κάποιου προσώπου μέσα στο σύνολο των δεδομένων μας. Επίσης θα εξετάσουμε ορισμένες προκλήσεις που προκύπτουν σε μία διαδικασία ανωνυμοποίησης, ωστόσο πρακτικά οι δυσκολίες αυτές θα εμφανιστούν κατά την εφαρμογή της ανωνυμοποίησης σε επόμενο κεφάλαιο.

Στο κεφάλαιο 8 θα αναλυθούν εργαλεία διαθέσιμα για ανωνυμοποίηση επιλέγοντας έτσι τα καταλληλότερα εξ΄ αυτών που θα εξυπηρετήσουν τις ανάγκες μας κατά τη διαδικασία βάσει των χαρακτηριστικών που προσφέρουν και της φιλικότητας προς το χρήστη.

Το κεφάλαιο 9 αφορά την υλοποίηση της ανωνυμοποίησης έχοντας πλέον επιλέξει από τα προηγούμενα κεφάλαια τις κατάλληλες μεθόδους. Στο σημείο αυτό θα εφαρμοστούν πρακτικά οι διάφορες προσεγγίσεις με τα κατάλληλα εργαλεία ώστε να λάβουμε τα αποτελέσματα που στη συνέχεια θα παρουσιάσουμε.

Στο κεφάλαιο 10, το οποίο αποτελεί τον επίλογο της παρούσας Μεταπτυχιακής Διατριβής, θα αναφερθούν συνοπτικά τα γενικά συμπεράσματα της έρευνας που πραγματοποιήθηκε καθώς και θέματα προς μελλοντική έρευνα. Επίσης θα γίνει αναφορά σε πιθανούς περιορισμούς που αντιμετωπίσαμε κατά την διεξαγωγή της έρευνάς μας.

Κεφάλαιο 2
Θεωρητικό Υπόβαθρο

Το παρόν κεφάλαιο αποτελεί θεωρητικό υπόβαθρο της παρούσας Μεταπτυχιακής Διατριβής και ορίζει μία βάση για βαθύτερη κατανόηση και ανάλυση εννοιών καθώς και κανονισμών που θα μελετηθούν σε μεγάλο βαθμό κατά την έρευνά μας. Έννοιες όπως τα θεμελιώδη ανθρώπινα δικαιώματα της ιδιωτικότητας και της προστασίας των προσωπικών δεδομένων φυσικών προσώπων, κανονισμοί που έχει ορίσει ο Γενικός Κανονισμός Προστασίας Δεδομένων (GDPR) της Ευρωπαϊκής Ένωσης για τα δεδομένα αυτά, καθώς και έννοιες και τεχνικές όπως η ψευδωνυμοποίηση και η ανωνυμοποίηση συνεισφέροντας έτσι σε μία ασφαλή αξιοποίηση των προσωπικών δεδομένων.

2.1 Προσωπικά Δεδομένα Και Ιδιωτικότητα

Στην ενότητα αυτή γίνεται μία εισαγωγή ώστε να εξοικειωθούμε με τα προσωπικά δεδομένα που θα μελετήσουμε, το πόσο κρίσιμα είναι και το πώς χρίζουν ιδιαίτερης μεταχείρισης προστατεύοντας έτσι θεμελιώδη ανθρώπινα δικαιώματα όπως η ιδιωτικότητα φυσικών

προσώπων. Η ανάγκη προστασίας τους κρίνεται επιτακτική από τον GDPR και πρέπει να ακολουθείται πιστά σε κάθε είδους επεξεργασίας και εκμετάλλευσής τους.

2.1.1 Προσωπικά Δεδομένα

Ως προσωπικά δεδομένα ορίζουμε οποιαδήποτε πληροφορία σχετίζεται με ένα ταυτοποιημένο ή ταυτοποιήσιμο φυσικό πρόσωπο [4]. Το πρόσωπο αυτό αποκαλείται και υποκείμενο των δεδομένων. Σύμφωνα με το άρθρο 8 του χάρτη θεμελιωδών δικαιωμάτων της Ευρωπαϊκής Ένωσης, αναφέρονται ρητά τα παρακάτω [3],

1. Κάθε πρόσωπο έχει δικαίωμα στη προστασία των δεδομένων προσωπικού χαρακτήρα που το αφορούν.

2. Η επεξεργασία αυτών των δεδομένων πρέπει να γίνεται νομίμως, για καθορισμένους σκοπούς και με βάση τη συγκατάθεση του ενδιαφερομένου ή για άλλους θεμιτούς λόγους που προβλέπονται από το νόμο. Κάθε πρόσωπο έχει δικαίωμα να έχει πρόσβαση στα συλλέγοντα δεδομένα που το αφορούν και να επιτυγχάνει τη διόρθωσή του.

3. Ο σεβασμός των κανόνων αυτών υπόκειται στον έλεγχο ανεξάρτητης αρχής.

Τα δεδομένα αυτά μπορεί να περιλαμβάνουν το ονοματεπώνυμο, διεύθυνση, τηλέφωνο, email, ημερομηνία γέννησης, IP διεύθυνση, αριθμό κοινωνικής ασφάλισης, βιομετρικά δεδομένα, ωστόσο δεν περιορίζονται μόνο σε αυτά. Τα προσωπικά δεδομένο μπορεί να περιλαμβάνουν και άλλες πιο ευαίσθητες πληροφορίες όπως η φυλή, η εθνικότητα, οι πολιτικές απόψεις, οι θρησκευτικές πεποιθήσεις, οι πληροφορίες για την υγεία ή ο σεξουαλικός προσανατολισμός ενός ατόμου.

Στο πλαίσιο των νόμων που έχουν οριστεί περί της προστασίας των προσωπικών δεδομένων, τόσο από τον Γενικό Κανονισμό Προστασίας δεδομένων (GDPR) της Ευρωπαϊκής Ένωσης [5] όσο το νόμο περί απορρήτου των καταναλωτών της Καλιφόρνια (CCPA) [6], τα προσωπικά δεδομένα ορίζονται ευρέως ώστε να περιλαμβάνουν οποιεσδήποτε πληροφορίες μπορούν είτε έμμεσα είτε άμεσα να ταυτοποιήσουν ένα άτομο. Οι νόμοι αυτοί ορίστηκαν ώστε να παρέχουν στα φυσικά πρόσωπα δικαιώματα επί των προσωπικών τους δεδομένων, όσον αφορά τη πρόσβαση σε αυτά από τρίτους, την επεξεργασία και τη διαγραφή τους, καθώς και του δικαιώματος ελέγχου του τρόπου κοινής χρήσης των δεδομένων τους.

2.1.2 Ιδιωτικότητα

Η ιδιωτικότητα αφορά το δικαίωμα ενός φυσικού προσώπου να ελέγχει ή να περιορίζει την πρόσβαση στις προσωπικές του πληροφορίες ή δεδομένα διατηρώντας ένα επίπεδο εμπιστευτικότητας και αυτονομίας στη προσωπική του ζωή [7]. Η ιδιωτικότητα μπορεί να θεωρηθεί ως θεμελιώδες ανθρώπινο δικαίωμα, καθώς επιτρέπει στα άτομα να διατηρούν την προσωπική τους ακεραιότητα, αξιοπρέπεια και ταυτότητα χωρίς αδικαιολόγητη παρέμβαση από άλλους ή από θεσμούς. Επίσης κατά το άρθρο 7 του χάρτη θεμελιωδών δικαιωμάτων της Ευρωπαϊκής Ένωσης [3], όσον αφορά την ιδιωτικότητα αναφέρεται ότι, κάθε πρόσωπο έχει δικαίωμα στο σεβασμό της ιδιωτικής και οικογενειακής του ζωής, της κατοικίας του και των επικοινωνιών του.

Στη ψηφιακή εποχή, η ιδιωτικότητα έχει καταστεί ως μία επιτακτική ανησυχία, καθώς καθημερινά συλλέγονται πολλές πληροφορίες που αφορούν προσωπικά δεδομένα ατόμων, είτε αυτές υποβάλλονται σε επεξεργασία, είτε μοιράζονται από διάφορες οντότητες, όπως κυβερνήσεις, εταιρείες καθώς και διαδικτυακές πλατφόρμες. Τα δεδομένα αυτά μπορεί να περιλαμβάνουν ευαίσθητες πληροφορίες, όπως δεδομένα υγείας, τοποθεσίας ή θρησκευτικών πεποιθήσεων, και η κακή χρήση ή ο κακός χειρισμός τους μπορεί να οδηγήσει σε διάφορες επιπτώσεις, όπως κλοπή της ταυτότητας κάποιου προσώπου, κοινωνικές διακρίσεις ή και δυσφήμιση.

Η προστασία της ιδιωτικής ζωής μπορεί να επιτευχθεί με διάφορα νομικά, τεχνικά και κοινωνικά μέσα, όπως οι νόμοι περί προστασίας δεδομένων, η κρυπτογράφηση και η εκπαίδευση των χρηστών. Ωστόσο, η προστασία της ιδιωτικής ζωής πρέπει να εξισορροπείται με άλλα κοινωνικά συμφέροντα, όπως η δημόσια καθώς και η εθνική ασφάλεια ή η δημόσια υγεία και πρέπει να εντάσσεται σε συγκεκριμένα κοινωνικά, πολιτιστικά και πολιτικά πρότυπα και αξίες [8].

2.2 Νομικό Πλαίσιο

Το 2018 συστήθηκε νέος Κανονισμός από την Ευρωπαϊκή Ένωση, ως Γενικός Κανονισμός Προστασίας Δεδομένων (GDPR) και τέθηκε σε ισχύ το Μάιο του 2018 [5]. Ο Κανονισμός αυτός αφορά τη προστασία των προσωπικών δεδομένων των πολιτών της Ε.Ε. και έχει άμεση εφαρμογή σε όλα τα Κράτη – Μέλη της Ε.Ε. Ο Κανονισμός ισχύει για όλες τις επιχειρήσεις και τους οργανισμούς που συλλέγουν, επεξεργάζονται και αποθηκεύουν προσωπικά δεδομένα των

πολιτών της ΕΕ, ακόμα και αν αυτές δεν εδρεύουν στην ΕΕ. Η οδηγία αυτή εφαρμόστηκε άμεσα σε όλα τα κράτη μέλη, χωρίς να χρειάζεται εθνική νομοθεσία στοχεύοντας στην ενίσχυση των δικαιωμάτων των πολιτών.

2.2.1 Προσωπικά Δεδομένα Κατά Τον GDPR

Σύμφωνα με τον GDPR, τα προσωπικά δεδομένα όπως αναφέραμε και παραπάνω ορίζονται ως οποιαδήποτε πληροφορία από φυσική, βιολογική, ψυχολογική, οικονομική, πολιτιστική ή κοινωνική άποψη και αφορά ένα φυσικό πρόσωπο, το οποίο μπορεί να αναγνωριστεί άμεσα ή έμμεσα από αυτήν την πληροφορία.

Πρακτικά πληροφορίες που χαρακτηρίζουν ένα φυσικό πρόσωπο και ορίζονται ως προσωπικά δεδομένα είναι:

- Όνομα και Επώνυμο
- Διεύθυνση (ταχυδρομική και ηλεκτρονική - email)
- Τηλέφωνο
- Ενδιαφέροντα και απόψεις
- Εικόνα (φωτογραφία)

Επιπλέον πληροφορίες που μπορούν να μας αναγνωρίσουν και παρουσιάζονται ως προσωπικά δεδομένα κατά τον GDPR είναι το ψευδώνυμο σε οποιαδήποτε διαδικτυακή υπηρεσία, η IP διεύθυνση καθώς και αρκετά άλλα.

Υπάρχουν όμως και κάποια άλλα προσωπικά δεδομένα τα οποία χρήζουν ιδιαίτερης προσοχής και μεγαλύτερης προστασίας, καθώς εμπίπτουν σε ιδιαίτερες ανησυχίες όπως η παραβίαση της ιδιωτικότητας και του απορρήτου. Τα δεδομένα αυτά χαρακτηρίζονται ως ευαίσθητα προσωπικά δεδομένα ή ειδικές κατηγορίες προσωπικών δεδομένων. Στις κατηγορίες αυτές έχουμε τα παρακάτω:

- Φυλετική ή εθνική προέλευση

- Πολιτικά φρονήματα

- Θρησκευτικές πεποιθήσεις

- Συμμετοχή σε συνδικαλιστική οργάνωση

- Υγεία

- Κοινωνική πρόνοια

- Ερωτική ζωή

- Γενετικά δεδομένα (δεδομένα που έχουν προκύψει από ανάλυση DNA, RNA)

- Βιομετρικά δεδομένα, εφόσον χρησιμοποιούνται για το σκοπό της αδιαμφισβήτητης ταυτοποίησης ενός προσώπου,

- Ποινική καταδίκη

2.2.2 Στόχος Του GDPR

Ο στόχος του GDPR είναι πολύ σαφής σε μία προσπάθεια να προστατεύσει τα δικαιώματα και την ιδιωτικότητα των πολιτών της Ευρωπαϊκής Ένωσης σχετικά με τα προσωπικά τους δεδομένα. Ο κανονισμός έχει σχεδιαστεί για να διασφαλίσει ότι οι επιχειρήσεις και άλλοι φορείς επεξεργασίας προσωπικών δεδομένων σέβονται τα δικαιώματα των ανθρώπων να διαχειρίζονται τα δεδομένα τους με ασφάλεια και διαφάνεια.

Οι βασικοί του στόχοι περιλαμβάνουν τα εξής:

- Προστασία της ιδιωτικής ζωής και των δικαιωμάτων των πολιτών της ΕΕ σχετικά με τα προσωπικά δεδομένα.

- Εξασφάλιση ότι οι επιχειρήσεις και οι φορείς συλλέγουν και επεξεργάζονται τα προσωπικά δεδομένα με σαφείς και επαρκείς σκοπούς και μόνον όταν υπάρχει νομική βάση για την επεξεργασία αυτών των δεδομένων.

- Ενίσχυση της ασφάλειας και της προστασίας των προσωπικών δεδομένων από κακόβουλες δραστηριότητες και από απώλειες, κλοπές και καταστροφές δεδομένων.

- Δημιουργία ενός ενιαίου ευρωπαϊκού ρυθμιστικού πλαισίου για τη διαχείριση των προσωπικών δεδομένων, με σκοπό να διασφαλιστεί ότι οι επιχειρήσεις και άλλοι φορείς επεξεργασίας προσωπικών δεδομένων τηρούν τους ίδιους κανόνες και πρότυπα σε όλη την ΕΕ.

- Ενίσχυση της εμπιστοσύνης των καταναλωτών στην οικονομία του ψηφιακού εμπορίου, με τη διασφάλιση ότι οι επιχειρήσεις και άλλοι φορείς επεξεργασίας προσωπικών δεδομένων χρησιμοποιούν τα δεδομένα με ηθικό και νόμιμο τρόπο.

Βάση των ανωτέρω και σύμφωνα με το άρθρο 3 του GDPR το οποίο αναφέρεται στο πεδίο εφαρμογής του κανονισμού, ορίζει τα κριτήρια που καθορίζουν αν ο κανονισμός ισχύει για μία επεξεργασία προσωπικών δεδομένων.

Σύμφωνα με το άρθρο αυτό, ο κανονισμός εφαρμόζεται:

1. στην επεξεργασία δεδομένων προσωπικού χαρακτήρα στο πλαίσιο των δραστηριοτήτων μιας εγκατάστασης ενός υπευθύνου επεξεργασίας ή εκτελούντος την επεξεργασία στην Ένωση, ανεξάρτητα από το κατά πόσο η επεξεργασία πραγματοποιείται εντός της Ένωσης.

2. στην επεξεργασία δεδομένων προσωπικού χαρακτήρα υποκειμένων των δεδομένων που βρίσκονται στην Ένωση από υπεύθυνο επεξεργασίας ή εκτελούντα την επεξεργασία μη εγκατεστημένο στην Ένωση, εάν οι δραστηριότητες επεξεργασίας σχετίζονται με:

   α) την προσφορά αγαθών ή υπηρεσιών στα εν λόγω υποκείμενα των δεδομένων στην Ένωση, ανεξαρτήτως εάν απαιτείται πληρωμή από τα υποκείμενα των δεδομένων, ή

   β) την παρακολούθηση της συμπεριφοράς τους, στον βαθμό που η συμπεριφορά αυτή λαμβάνει χώρα εντός της Ένωσης.

3. για την επεξεργασία δεδομένων προσωπικού χαρακτήρα από υπεύθυνο επεξεργασίας μη εγκατεστημένο στην Ένωση, αλλά σε τόπο όπου εφαρμόζεται το δίκαιο κράτους μέλους δυνάμει του δημόσιου διεθνούς δικαίου.

Το άρθρο 4 του GDPR αποτελείται από ορισμούς που χρησιμοποιούνται στον κανονισμό και παρέχουν σημαντικές πληροφορίες που θα αναλυθούν στη πορεία της έρευνάς μας. Χαρακτηρίζοντας τους βασικότερους ορισμούς που θα μας απασχολήσουν στη συνέχεια της έρευνάς μας έχουμε τα ακόλουθα.

1. **Επεξεργασία**: κάθε πράξη ή σειρά πράξεων που πραγματοποιείται με ή χωρίς τη χρήση αυτοματοποιημένων μέσων, σε δεδομένα προσωπικού χαρακτήρα ή σε σύνολα δεδομένων προσωπικού χαρακτήρα, όπως η συλλογή, η καταχώριση, η οργάνωση, η διάρθρωση, η αποθήκευση, η προσαρμογή ή η μεταβολή, η ανάκτηση, η αναζήτηση πληροφοριών, η χρήση, η κοινολόγηση με διαβίβαση, η διάδοση ή κάθε άλλη μορφή διάθεσης, η συσχέτιση ή ο συνδυασμός, ο περιορισμός, η διαγραφή ή η καταστροφή.

2. **Υπεύθυνος της επεξεργασίας**: κάθε φυσικό ή νομικό πρόσωπο, δημόσιο φορέα, υπηρεσία ή άλλο φορέα που καθορίζει τους σκοπούς και τον τρόπο επεξεργασίας των προσωπικών δεδομένων. Αυτός ο φορέας είναι υπεύθυνος για την τήρηση των όρων του GDPR και για την προστασία των δεδομένων των υποκειμένων δεδομένων.

3. **Εκτελών την επεξεργασία**: κάθε φυσικό ή νομικό πρόσωπο, δημόσια αρχή, υπηρεσία ή άλλος φορέας που επεξεργάζεται τα δεδομένα προσωπικού χαρακτήρα για λογαριασμό του υπεύθυνου της επεξεργασίας.

4. **Ψευδωνυμοποίηση**: η επεξεργασία δεδομένων προσωπικού χαρακτήρα κατά τρόπο τέτοιο ώστε τα δεδομένα να μη μπορούν πλέον να αποδοθούν σε συγκεκριμένο υποκείμενο των δεδομένων χωρίς τη χρήση συμπληρωματικών πληροφοριών, εφόσον οι εν λόγω συμπληρωματικές πληροφορίες διατηρούνται χωριστά και υπόκεινται σε τεχνικά και οργανωτικά μέτρα προκειμένου να διασφαλιστεί ότι δεν μπορούν να αποδοθούν σε ταυτοποιημένο ή ταυτοποιήσιμο φυσικό πρόσωπο.

5. **Παραβίαση δεδομένων προσωπικού χαρακτήρα**: η παραβίαση της ασφάλειας που οδηγεί σε τυχαία ή παράνομη καταστροφή, απώλεια, μεταβολή, άνευ άδειας κοινολόγηση ή πρόσβαση δεδομένων προσωπικού χαρακτήρα που διαβιβάστηκαν, αποθηκεύτηκαν ή υποβλήθηκαν κατ' άλλο τρόπο σε επεξεργασία.

6. **Δεδομένα που αφορούν την υγεία**: δεδομένα προσωπικού χαρακτήρα τα οποία σχετίζονται με τη σωματική ή ψυχική υγεία ενός φυσικού προσώπου, περιλαμβανομένης

της παροχής υπηρεσιών υγειονομικής φροντίδας, και τα οποία αποκαλύπτουν πληροφορίες σχετικά με την κατάσταση της υγείας του.

Συνεχίζοντας με το άρθρο 5 που περιλαμβάνει τις αρχές που διέπουν την επεξεργασία δεδομένων προσωπικού χαρακτήρα [9], έχουμε τα παρακάτω.

Τα δεδομένα προσωπικού χαρακτήρα:

1. Υποβάλλονται σε σύννομη και θεμιτή επεξεργασία με διαφανή τρόπο σε σχέση με το υποκείμενο των δεδομένων («νομιμότητα, αντικειμενικότητα και διαφάνεια»).

2. Συλλέγονται για καθορισμένους, ρητούς και νόμιμους σκοπούς και δεν υποβάλλονται σε περαιτέρω επεξεργασία κατά τρόπο ασύμβατο προς τους σκοπούς αυτούς η περαιτέρω επεξεργασία για σκοπούς αρχειοθέτησης προς το δημόσιο συμφέρον ή σκοπούς επιστημονικής ή ιστορικής έρευνας ή στατιστικούς σκοπούς δεν θεωρείται ασύμβατη με τους αρχικούς σκοπούς σύμφωνα με το άρθρο 89 παράγραφος 1 («περιορισμός του σκοπού»).

3. Είναι κατάλληλα, συναφή και περιορίζονται στο αναγκαίο για τους σκοπούς για τους οποίους υποβάλλονται σε επεξεργασία («ελαχιστοποίηση των δεδομένων»).

4. Είναι ακριβή και, όταν είναι αναγκαίο, επικαιροποιούνται· πρέπει να λαμβάνονται όλα τα εύλογα μέτρα για την άμεση διαγραφή ή διόρθωση δεδομένων προσωπικού χαρακτήρα τα οποία είναι ανακριβή, σε σχέση με τους σκοπούς της επεξεργασίας («ακρίβεια»).

5. Διατηρούνται υπό μορφή που επιτρέπει την ταυτοποίηση των υποκειμένων των δεδομένων μόνο για το διάστημα που απαιτείται για τους σκοπούς της επεξεργασίας των δεδομένων προσωπικού χαρακτήρα· τα δεδομένα προσωπικού χαρακτήρα μπορούν να αποθηκεύονται για μεγαλύτερα διαστήματα, εφόσον τα δεδομένα προσωπικού χαρακτήρα θα υποβάλλονται σε επεξεργασία μόνο για σκοπούς αρχειοθέτησης προς το δημόσιο συμφέρον, για σκοπούς επιστημονικής ή ιστορικής έρευνας ή για στατιστικούς σκοπούς, σύμφωνα με το άρθρο 89 παράγραφος 1 και εφόσον εφαρμόζονται τα κατάλληλα τεχνικά και οργανωτικά μέτρα που απαιτεί ο παρών κανονισμός για τη διασφάλιση των δικαιωμάτων και ελευθεριών του υποκειμένου των δεδομένων («περιορισμός της περιόδου αποθήκευσης»).

6. Υποβάλλονται σε επεξεργασία κατά τρόπο που εγγυάται την ενδεδειγμένη ασφάλεια των δεδομένων προσωπικού χαρακτήρα, μεταξύ άλλων την προστασία τους από μη εξουσιοδοτημένη ή παράνομη επεξεργασία και τυχαία απώλεια, καταστροφή ή φθορά, με τη χρησιμοποίηση κατάλληλων τεχνικών ή οργανωτικών μέτρων («ακεραιότητα και εμπιστευτικότητα»).

Το άρθρο 6 αναφέρεται στη νομιμότητα της επεξεργασίας, ορίζοντας έτσι την ανάγκη να πληρείται τουλάχιστον μία από τις παρακάτω προϋποθέσεις, ώστε να είναι σύννομη η επεξεργασία.

1. το υποκείμενο των δεδομένων έχει συναινέσει στην επεξεργασία των δεδομένων προσωπικού χαρακτήρα του για έναν ή περισσότερους συγκεκριμένους σκοπούς,

2. η επεξεργασία είναι απαραίτητη για την εκτέλεση σύμβασης της οποίας το υποκείμενο των δεδομένων είναι συμβαλλόμενο μέρος ή για να ληφθούν μέτρα κατ' αίτηση του υποκειμένου των δεδομένων πριν από τη σύναψη σύμβασης,

3. η επεξεργασία είναι απαραίτητη για τη συμμόρφωση με έννομη υποχρέωση του υπευθύνου επεξεργασίας,

4. η επεξεργασία είναι απαραίτητη για τη διαφύλαξη ζωτικού συμφέροντος του υποκειμένου των δεδομένων ή άλλου φυσικού προσώπου,

5. η επεξεργασία είναι απαραίτητη για την εκπλήρωση καθήκοντος που εκτελείται προς το δημόσιο συμφέρον ή κατά την άσκηση δημόσιας εξουσίας που έχει ανατεθεί στον υπεύθυνο επεξεργασίας,

6. η επεξεργασία είναι απαραίτητη για τους σκοπούς των έννομων συμφερόντων που επιδιώκει ο υπεύθυνος επεξεργασίας ή τρίτος, εκτός εάν έναντι των συμφερόντων αυτών υπερισχύει το συμφέρον ή τα θεμελιώδη δικαιώματα και οι ελευθερίες του υποκειμένου των δεδομένων που επιβάλλουν την προστασία των δεδομένων προσωπικού χαρακτήρα, ιδίως εάν το υποκείμενο των δεδομένων είναι παιδί.

Το στοιχείο 6) του παραπάνω εδαφίου δεν εφαρμόζεται στην επεξεργασία που διενεργείται από δημόσιες αρχές κατά την άσκηση των καθηκόντων τους.

Το άρθρο 9 περιλαμβάνει την επεξεργασία ειδικών κατηγοριών δεδομένων προσωπικού χαρακτήρα. Κατά τη παράγραφο 1 του συγκεκριμένου άρθρου, απαγορεύεται η επεξεργασία δεδομένων προσωπικού χαρακτήρα που αποκαλύπτουν τη φυλετική ή εθνική καταγωγή, τα πολιτικά φρονήματα, τις θρησκευτικές ή φιλοσοφικές πεποιθήσεις ή τη συμμετοχή σε συνδικαλιστική οργάνωση, καθώς και η επεξεργασία γενετικών δεδομένων, βιομετρικών δεδομένων με σκοπό την αδιαμφισβήτητη ταυτοποίηση προσώπου, δεδομένων που αφορούν την υγεία ή δεδομένων που αφορούν τη σεξουαλική ζωή φυσικού προσώπου ή τον γενετήσιο προσανατολισμό.

Στη παράγραφο 2 τονίζεται ότι η απαγόρευση της παραγράφου 1 δεν εφαρμόζεται σε κάποιες περιπτώσεις και, άρα, υπό προϋποθέσεις μπορεί να είναι κατ' εξαίρεση επιτρεπτή η επεξεργασία και δεδομένων ειδικών κατηγοριών. Πιο συγκεκριμένα, κάποιες περιπτώσεις για τις οποίες είναι επιτρεπτή κατ' εξαίρεση η επεξεργασία δεδομένων ειδικών κατηγοριών είναι οι εξής:

α) το υποκείμενο των δεδομένων έχει παράσχει ρητή συγκατάθεση για την επεξεργασία αυτών των δεδομένων προσωπικού χαρακτήρα για έναν ή περισσότερους συγκεκριμένους σκοπούς, εκτός εάν το δίκαιο της Ένωσης ή κράτους μέλους προβλέπει ότι η απαγόρευση που αναφέρεται στην παράγραφο 1 δεν μπορεί να αρθεί από το υποκείμενο των δεδομένων,

β) επεξεργασία είναι απαραίτητη για σκοπούς προληπτικής ή επαγγελματικής ιατρικής, εκτίμησης της ικανότητας προς εργασία του εργαζομένου, ιατρικής διάγνωσης, παροχής υγειονομικής ή κοινωνικής περίθαλψης ή θεραπείας ή διαχείρισης υγειονομικών και κοινωνικών συστημάτων και υπηρεσιών βάσει του ενωσιακού δικαίου ή του δικαίου κράτους μέλους ή δυνάμει σύμβασης με επαγγελματία του τομέα της υγείας και με την επιφύλαξη των προϋποθέσεων και των εγγυήσεων που αναφέρονται στην παράγραφο 3,

γ) η επεξεργασία είναι απαραίτητη για λόγους δημόσιου συμφέροντος στον τομέα της δημόσιας υγείας, όπως η προστασία έναντι σοβαρών διασυνοριακών απειλών κατά της υγείας ή η διασφάλιση υψηλών προτύπων ποιότητας και ασφάλειας της υγειονομικής περίθαλψης και των φαρμάκων ή των ιατροτεχνολογικών προϊόντων, βάσει του δικαίου της Ένωσης ή του δικαίου κράτους μέλους, το οποίο προβλέπει κατάλληλα και συγκεκριμένα μέτρα για την προστασία των δικαιωμάτων και ελευθεριών του υποκειμένου των δεδομένων, ειδικότερα δε του επαγγελματικού απορρήτου,

δ) η επεξεργασία είναι απαραίτητη για σκοπούς αρχειοθέτησης προς το δημόσιο συμφέρον, για σκοπούς επιστημονικής ή ιστορικής έρευνας ή για στατιστικούς σκοπούς σύμφωνα με το άρθρο 89 παράγραφος 1 βάσει του δικαίου της Ένωσης ή κράτους μέλους, οι οποίοι είναι ανάλογοι προς τον επιδιωκόμενο στόχο, σέβονται την ουσία του δικαιώματος στην προστασία των δεδομένων και προβλέπουν κατάλληλα και συγκεκριμένα μέτρα για τη διασφάλιση των θεμελιωδών δικαιωμάτων και των συμφερόντων του υποκειμένου των δεδομένων.

Τέλος από τη παράγραφο 4 του άρθρου 9 έχουμε ότι, τα κράτη μέλη μπορούν να διατηρούν ή να θεσπίζουν περαιτέρω όρους, μεταξύ άλλων και περιορισμούς, όσον αφορά την επεξεργασία γενετικών δεδομένων, βιομετρικών δεδομένων ή δεδομένων που αφορούν την υγεία.

Από εκεί και ύστερα, ο GDPR παρέχει ένα σύνολο ειδικότερων υποχρεώσεων με τους οποίους είναι επιφορτισμένοι οι υπεύθυνοι επεξεργασίας. Μία τέτοια υποχρέωση, γνωστή με το όνομα «προστασία των δεδομένων ήδη από το σχεδιασμό», περιγράφεται στο άρθρο 25 του ΓΚΠΔ. Ειδικότερα, όπως αναφέρεται στο άρθρο αυτό, λαμβάνοντας υπόψη το επίπεδο της τεχνολογίας, το κόστος, τη φύση και το πεδίο εφαρμογής, το πλαίσιο και τους σκοπούς της επεξεργασίας, καθώς και τους κινδύνους διαφορετικής πιθανότητας και σοβαρότητας για τα δικαιώματα και τις ελευθερίες των φυσικών προσώπων που ενέχει η επεξεργασία, , ο υπεύθυνος επεξεργασίας, τόσο κατά τον καθορισμό των μέσων επεξεργασίας όσο και κατά την ίδια την επεξεργασία, εφαρμόζει κατάλληλα τεχνικά και οργανωτικά μέτρα, όπως η ψευδωνυμοποίηση, τα οποία αποσκοπούν στην αποτελεσματική εφαρμογή των αρχών προστασίας των δεδομένων, όπως η ελαχιστοποίηση των δεδομένων, και στην ενσωμάτωση των αναγκαίων εγγυήσεων στην επεξεργασία, ώστε να πληρούνται οι απαιτήσεις του παρόντος κανονισμού και να προστατεύονται τα δικαιώματα των υποκειμένων των δεδομένων.

Περαιτέρω, όπως επίσης αναφέρεται στο ίδιο άρθρο, ο υπεύθυνος επεξεργασίας εφαρμόζει κατάλληλα τεχνικά και οργανωτικά μέτρα για να διασφαλίζει ότι, εξ ορισμού, υφίστανται επεξεργασία μόνο τα δεδομένα προσωπικού χαρακτήρα που είναι απαραίτητα για τον εκάστοτε σκοπό της επεξεργασίας. Αυτή η υποχρέωση ισχύει για το εύρος των δεδομένων προσωπικού χαρακτήρα που συλλέγονται, τον βαθμό της επεξεργασίας τους, την περίοδο αποθήκευσης και την προσβασιμότητά τους. Ειδικότερα, τα εν λόγω μέτρα διασφαλίζουν ότι, εξ ορισμού, τα δεδομένα προσωπικού χαρακτήρα δεν καθίστανται προσβάσιμα χωρίς την παρέμβαση του φυσικού προσώπου σε αόριστο αριθμό φυσικών προσώπων.

Συνεχίζουμε με το άρθρο 32 που περιγράφει την ασφάλεια επεξεργασίας των προσωπικών δεδομένων. Πιο συγκεκριμένα αναφέρεται στο κατάλληλα τεχνικά και οργανωτικά μέτρα προκειμένου να διασφαλίζεται το κατάλληλο επίπεδο ασφάλειας έναντι των κινδύνων που πρέπει να εφαρμόζουν κατά την επεξεργασία των δεδομένων, ο υπεύθυνος επεξεργασίας και ο εκτελών την επεξεργασία, λαμβάνοντας υπόψη τις τελευταίες εξελίξεις, το κόστος εφαρμογής και τη φύση, το πεδίο εφαρμογής, το πλαίσιο και τους σκοπούς της επεξεργασίας, καθώς και τους κινδύνους διαφορετικής πιθανότητας επέλευσης και σοβαρότητας για τα δικαιώματα και τις ελευθερίες των φυσικών προσώπων. Κατά περίπτωση αναφέρονται τα παρακάτω.

1. Ψευδωνυμοποίηση και κρυπτογράφηση δεδομένων προσωπικού χαρακτήρα.

2. Δυνατότητα διασφάλισης του απορρήτου, της ακεραιότητας, της διαθεσιμότητας και της αξιοπιστίας των συστημάτων και των υπηρεσιών επεξεργασίας σε συνεχή βάση.

3. Δυνατότητα αποκατάστασης της διαθεσιμότητας και της πρόσβασης σε δεδομένα προσωπικού χαρακτήρα σε εύθετο χρόνο σε περίπτωση φυσικού ή τεχνικού συμβάντος.

4. Διαδικασία για την τακτική δοκιμή, εκτίμηση και αξιολόγηση της αποτελεσματικότητας των τεχνικών και των οργανωτικών μέτρων για τη διασφάλιση της ασφάλειας της επεξεργασίας.

Κατά την εκτίμηση του ενδεδειγμένου επιπέδου ασφάλειας λαμβάνονται ιδίως υπόψη οι κίνδυνοι που απορρέουν από την επεξεργασία, ιδίως από τυχαία ή παράνομη καταστροφή, απώλεια, αλλοίωση, άνευ αδείας κοινολόγηση ή προσπέλαση δεδομένων προσωπικού χαρακτήρα που διαβιβάστηκαν, αποθηκεύτηκαν ή υποβλήθηκαν κατ' άλλο τρόπο σε επεξεργασία.

Το άρθρο 35 του GDPR αναφέρει την εκτίμηση αντικτύπου σχετικά με τη προστασία δεδομένων. Σύμφωνα με το άρθρο έχουμε τα εξής.

1. Όταν ένα είδος επεξεργασίας, ιδίως με χρήση νέων τεχνολογιών και συνεκτιμώντας τη φύση, το πεδίο εφαρμογής, το πλαίσιο και τους σκοπούς της επεξεργασίας, ενδέχεται να επιφέρει υψηλό κίνδυνο για τα δικαιώματα και τις ελευθερίες των φυσικών προσώπων, ο υπεύθυνος επεξεργασίας διενεργεί, πριν από την επεξεργασία, εκτίμηση των επιπτώσεων των σχεδιαζόμενων πράξεων επεξεργασίας στην προστασία δεδομένων προσωπικού

χαρακτήρα. Σε μία εκτίμηση μπορεί να εξετάζεται ένα σύνολο παρόμοιων πράξεων επεξεργασίας οι οποίες ενέχουν παρόμοιους υψηλούς κινδύνους.

2. Ο υπεύθυνος επεξεργασίας ζητεί τη γνώμη του υπευθύνου προστασίας δεδομένων, εφόσον έχει οριστεί, κατά τη διενέργεια εκτίμησης αντικτύπου σχετικά με την προστασία δεδομένων.

3. Η αναφερόμενη στην παράγραφο 1 εκτίμηση αντικτύπου σχετικά με την προστασία δεδομένων απαιτείται ιδίως στην περίπτωση:

   α) συστηματικής και εκτενούς αξιολόγησης προσωπικών πτυχών σχετικά με φυσικά πρόσωπα, η οποία βασίζεται σε αυτοματοποιημένη επεξεργασία, περιλαμβανομένης της κατάρτισης προφίλ, και στην οποία βασίζονται αποφάσεις που παράγουν έννομα αποτελέσματα σχετικά με το φυσικό πρόσωπο ή ομοίως επηρεάζουν σημαντικά το φυσικό πρόσωπο,

   β) μεγάλης κλίμακας επεξεργασίας των ειδικών κατηγοριών δεδομένων που αναφέρονται στο άρθρο 9 παράγραφος 1 ή δεδομένων προσωπικού χαρακτήρα που αφορούν ποινικές καταδίκες και αδικήματα ή

   γ) συστηματικής παρακολούθησης δημοσίως προσβάσιμου χώρου σε μεγάλη κλίμακα.

4. Η εκτίμηση περιέχει τουλάχιστον:

   α) συστηματική περιγραφή των προβλεπόμενων πράξεων επεξεργασίας και των σκοπών της επεξεργασίας, περιλαμβανομένου, κατά περίπτωση, του έννομου συμφέροντος που επιδιώκει ο υπεύθυνος επεξεργασίας,

   β) εκτίμηση της αναγκαιότητας και της αναλογικότητας των πράξεων επεξεργασίας σε συνάρτηση με τους σκοπούς,

   γ) εκτίμηση των κινδύνων για τα δικαιώματα και τις ελευθερίες των υποκειμένων των δεδομένων που αναφέρονται στην παράγραφο 1 και

   δ) τα προβλεπόμενα μέτρα αντιμετώπισης των κινδύνων, περιλαμβανομένων των εγγυήσεων, των μέτρων και μηχανισμών ασφάλειας, ώστε να διασφαλίζεται η προστασία

των δεδομένων προσωπικού χαρακτήρα και να αποδεικνύεται η συμμόρφωση προς τον παρόντα κανονισμό, λαμβάνοντας υπόψη τα δικαιώματα και τα έννομα συμφέροντα των υποκειμένων των δεδομένων και άλλων ενδιαφερόμενων προσώπων.

Ουσιαστικά, η εν λόγω απαίτηση υποχρεώνει τους υπευθύνους επεξεργασίας, εφόσον η επεξεργασία που διενεργούν ενέχει πολλούς κινδύνους για δικαιώματα και ελευθερίες προσώπων, να κάνουν μία συστηματική αξιολόγηση των κινδύνων και να λάβουν τεκμηριωμένες αποφάσεις για τα μέτρα αντιμετώπισής τους που πρέπει να ληφθούν. Σημειώνεται ότι η επεξεργασία δεδομένων υγείας μεγάλης κλίμακας, όπως είναι αυτή που εξετάζεται στο πλαίσιο της παρούσας διατριβής, εμπίπτει στην εν λόγω περίπτωση.

Τέλος τα άρθρα 37 - 39 αναφέρονται στον ορισμό Υπεύθυνου Προστασίας Δεδομένων (DPO), τη θέση του, καθώς και τα καθήκοντά του. Ειδικότερα σε περιπτώσεις όπου τα δεδομένα μας αφορούν ευαίσθητες προσωπικές πληροφορίες ασθενών όπως νοσοκομεία, φορείς της υγείας αλλά και γενικότερα δημόσιους φορείς και αρχές (πλην των δικαστηρίων που ενεργούν στο πλαίσιο της δικαιοδοτικής τους αρμοδιότητας), ο ορισμός του ΥΠΔ κρίνεται απαραίτητος.

Πιο συγκεκριμένα, υπό συγκεκριμένες προϋποθέσεις, ο υπεύθυνος επεξεργασίας των δεδομένων και ο εκτελών την επεξεργασία πλέον υποχρεούνται να ορίζουν ΥΠΔ.

Ο ΥΠΔ βοηθά ως προς τη διευκόλυνση του υπεύθυνου επεξεργασίας και του εκτελούντος την επεξεργασία των δεδομένων στη συμμόρφωση με τις διατάξεις του GDPR, τηρώντας ένα ρόλο συμβουλευτικό (όχι αποφασιστικό) μη φέροντας οποιαδήποτε ευθύνη για τη μη τήρηση των κανονισμών του GDPR. Επίσης μεσολαβεί μεταξύ των ενδιαφερομένων (εποπτικές αρχές, υποκείμενα των δεδομένων), ενεργώντας ως σημείο επικοινωνίας για ζητήματα που αφορούν την επεξεργασία, καθώς και ενημερώνεται, παρακολουθεί τη συμμόρφωση με τον παρόντα κανονισμό, τις διατάξεις του κράτους ή της Ένωσης σχετικά με τη προστασία δεδομένων προσωπικού χαρακτήρα και τις πολιτικές που εφαρμόζουν πάνω σε αυτή ο υπεύθυνος επεξεργασίας και ο εκτελών την επεξεργασία των δεδομένων.

Αντίθετα υπεύθυνος να διασφαλίζει και να μπορεί να αποδείξει ότι η επεξεργασία των δεδομένων διενεργείται με βάσει την διατάξεις που έχει ορίσει ο GDPR είναι ο υπεύθυνος επεξεργασίας ή ο εκτελών την επεξεργασία.

2.2.3 Προστασία της Ιδιωτικότητας Πριν Τον GDPR

Η προστασία της ιδιωτικότητας έχει αναγνωριστεί ως δικαίωμα ήδη από το 1950 στην Ευρωπαϊκή Σύμβαση των Δικαιωμάτων του Ανθρώπου (ΕΣΔΑ) από το Συμβούλιο της Ευρώπης [10]. Σύμφωνα με το άρθρο 8 της σύμβασης αυτής, κάθε πρόσωπο δικαιούται σεβασμό στην ιδιωτική και οικογενειακή του ζωή, τη κατοικία του και την αλληλογραφία του. Στη συνέχεια αναφέρει ρητά ότι, δεν επιτρέπεται να υπάρξει επέμβαση δημόσιας αρχής, ως προς την άσκηση του δικαιώματος αυτού, εκτός αν αυτό προβλέπεται από το νόμο και είναι αναγκαίο για την εθνική ή τη δημόσια ασφάλεια σε μία δημοκρατική κοινωνία.

Το 1981 το Συμβούλιο της Ευρώπης υπέγραψε την Σύμβαση 108 για την προστασία του ατόμου από την αυτοματοποιημένη επεξεργασία προσωπικών δεδομένων [11]. Στοχεύει στο να ενισχυθεί η προστασία των δεδομένων προσωπικού χαρακτήρα σε παγκόσμιο επίπεδο. Αναφέρεται στις πληροφορίες που αποκαλύπτουν τα πολιτικά φρονήματα, τη φυλετική προέλευση, θρησκευτικές πεποιθήσεις, καθώς και πληροφορίες που αφορούν την υγεία ή τη σεξουαλική ζωή και ποινικές καταδίκες, δεν μπορούν να επεξεργάζονται με αυτοματοποιημένες διαδικασίες, αν το εσωτερικό δίκαιο δεν προβλέπει κατάλληλες εγγυήσεις.

Το 1995 εκδόθηκε η οδηγία 95/46/ΕΚ για τη προστασία των φυσικών προσώπων για τη προστασία των προσωπικών τους δεδομένων και την ελεύθερη διακίνηση των δεδομένων τους [12]. Θέσπισε ένα κανονιστικό πλαίσιο που αποσκοπεί στην εγκαθίδρυση μιας ισορροπίας μεταξύ ενός υψηλού επιπέδου προστασίας της ιδιωτικής ζωής των προσώπων και της ελεύθερης κυκλοφορίας των δεδομένων προσωπικού χαρακτήρα ανά την Ευρωπαϊκή Ένωση (ΕΕ). Προς το σκοπό αυτό, η οδηγία όρισε τα όρια για τη συλλογή και τη χρησιμοποίηση των δεδομένων προσωπικού χαρακτήρα και ζητά τη δημιουργία, σε κάθε κράτος μέλος, ενός ανεξάρτητου εθνικού οργανισμού επιφορτισμένου με την εποπτεία οποιασδήποτε δραστηριότητας συνδέεται με την επεξεργασία δεδομένων προσωπικού χαρακτήρα.

Η Οδηγία 95/46/ΕΚ είχε ενσωματωθεί στις εθνικές νομοθεσίες των Κρατών – Μελών της ΕΕ, συμπεριλαμβανομένων φυσικά της Κύπρου και της Ελλάδας. Ωστόσο, μετά την θέση σε εφαρμογή του GDPR, έχει πλέον καταργηθεί.

Άξιον αναφοράς είναι ότι, ειδικά για την Ελλάδα, η προστασία προσωπικών δεδομένων ως ατομικό δικαίωμα έχει πλέον προβλεφθεί και στο Σύνταγμα με την αναθεώρηση του 2001 [13]. Συγκεκριμένα κατά το άρθρο 9Α του συντάγματος, καθένας έχει δικαίωμα προστασίας από τη συλλογή, την επεξεργασία και τη χρήση, ιδίως με ηλεκτρονικά μέσα, των προσωπικών του δεδομένων, όπως ορίζει ο νόμος. Η προστασία των προσωπικών δεδομένων διασφαλίζεται από ανεξάρτητη αρχή, που συγκροτείται και λειτουργεί, όπως ορίζει ο νόμος.

Αντίστοιχα για τη Κύπρο, το άρθρο 15 του Συντάγματος της Κυπριακής Δημοκρατίας αναφέρει ότι, έκαστος έχει το δικαίωμα σεβασμού ως προς την τη ιδιωτική και οικογενειακή του ζωή. Δεν χωράει επέμβαση κατά την άσκηση του δικαιώματος αυτού, παρά μόνο αν είναι σύμφωνος αυτός ή αν είναι αναγκαίο για το συμφέρον, την ασφάλειας της Δημοκρατίας ή της συνταγματικής τάξεων ή δημόσιας ασφάλειας, της δημόσιας υγείας [14].

2.3 Κανονισμός Ευρωπαϊκού Χώρου Δεδομένων Υγείας

Στη προσπάθεια της πλήρους εκμετάλλευσης των δυνατοτήτων των δεδομένων υγείας, από την Ευρωπαϊκή Επιτροπή, το Μάη του 2022 προτάθηκε ένας νέος Κανονισμός αναφορικά με τα δεδομένα αυτά, ως Ευρωπαϊκός Χώρος Δεδομένων Υγείας (EHDS) [15]. Η πρόταση του Κανονισμού αυτού έχει ως στόχο τα εξής.

1. Την υποστήριξη ώστε τα υποκείμενα των δεδομένων υγείας να μπορούν να έχουν τον έλεγχο των δεδομένων τους.

2. Την υποστήριξη ώστε η χρήση των δεδομένων υγείας να εξυπηρετεί καλύτερα την παροχή υγειονομικής περίθαλψης καθώς και τις ανάγκες έρευνας και την καινοτομία.

3. Τη δυνατότητα της ΕΕ να αξιοποιεί πλήρως τις δυνατότητες των δεδομένων υγείας μέσω ενός ασφαλούς διαμοιρασμού τους και να μπορεί να επαναχρησιμοποιήσει τα δεδομένα αυτά.

Σε επόμενο κεφάλαιο θα αναφερθούμε αναλυτικότερα στον Κανονισμό αυτό, καθώς ακόμα και αν δεν έχει τεθεί σε πλήρη εφαρμογή ακόμα, παρουσιάζει ιδιαίτερο ενδιαφέρον όσον αφορά τα δεδομένα υγείας αλλά και γενικότερα ολόκληρο τον τομέα της υγειονομικής περίθαλψης.

2.4 Ψευδωνυμοποίηση

Παρατηρώντας τους ρυθμούς ανάπτυξης και υιοθέτησης νέων τεχνολογιών τις τελευταίες δεκαετίες σε διάφορους τομείς να εξελίσσεται ραγδαία, εμφανίζεται και η επιρροή τους σε προσωπικά δεδομένα, όπως για παράδειγμα στον τομέα της υγείας για τη βελτίωση της υγειονομικής περίθαλψης. Έτσι με την ενσωμάτωση των τεχνολογιών αυτών σε διάφορους τομείς δημιουργούνται νέες προκλήσεις σχετικά με τα προσωπικά δεδομένα και την ασφάλειά τους στο διαδίκτυο, καθώς οι ανάγκες διαμοιρασμού τους αυξάνονται. Ως εκ τούτου είναι σημαντικό η επεξεργασία και η συλλογή των δεδομένων αυτών, αρχικά να πραγματοποιείται με τα ελάχιστα αναγκαία δεδομένα για τους σκοπούς, καθώς και να αξιοποιούνται κατάλληλα οργανωτικά και τεχνικά μέτρα για τη προστασία του απορρήτου και της ιδιωτικότητας των υποκειμένων.

2.4.1 Ορισμός Της Ψευδωνυμοποίησης

Η ψευδωνυμοποίηση είναι μία τεχνική επεξεργασίας δεδομένων η οποία περιλαμβάνει την αντικατάσταση στοιχείων με κάποιο ψευδώνυμο ή κάποιο κωδικό. Τα στοιχεία αυτά συνήθως αποτελούν προσωπική ταυτοποίηση και στοχεύουμε στο να μην αποκαλύψουμε τη πραγματική τους ταυτότητα [16]. Σκοπός της τεχνικής αυτής είναι η προστασία του απορρήτου και της ασφάλειας των προσωπικών δεδομένων καθιστώντας δυσκολότερη τη πρόσβαση σε αυτά ή τη χρήση τους από μη εξουσιοδοτημένες οντότητες. Η ψευδωνυμοποίηση χρησιμοποιείται συχνά σε καταστάσεις όπου απαιτείται επεξεργασία ή κοινή χρήση προσωπικών ή και ευαίσθητων δεδομένων, αλλά οι αυστηροί κανονισμοί περί απορρήτου ή οι νόμοι περί προστασίας δεδομένων εμποδίζουν τη χρήση τους με μη κρυπτογραφημένο τρόπο. Χαρακτηριστικό παράδειγμα είναι και ο τομέας της επεξεργασίας δεδομένων υγείας για ερευνητικούς σκοπούς, όπου τα δεδομένα ασθενών, όπως τα ιατρικά αρχεία, θα πρέπει να είναι ψευδωνυμοποιημένα για την αποτροπή μη εξουσιοδοτημένης πρόσβασης ή κακής χρήσης τους.

Σύμφωνα με τον Γενικό Κανονισμό Προστασίας Δεδομένων (GDPR) της Ευρωπαϊκής Ένωσης, η ψευδωνυμοποίηση ορίζεται ως «η επεξεργασία προσωπικών δεδομένων με τέτοιο τρόπο ώστε τα δεδομένα αυτά να μην μπορούν πλέον να αποδοθούν σε συγκεκριμένο υποκείμενο χωρίς τη χρήση πρόσθετων πληροφοριών, υπό την προϋπόθεση ότι αυτές οι πρόσθετες πληροφορίες διατηρούνται χωριστά και υπόκεινται σε τεχνικά και οργανωτικά μέτρα για να διασφαλιστεί ότι τα προσωπικά δεδομένα δεν αποδίδονται σε αναγνωρισμένο ή αναγνωρίσιμο φυσικό πρόσωπο.»

[16]. Ο GDPR αποτελεί και το πρώτο νομικό κείμενο στο οποίο γίνεται αναφορά στην έννοια της ψευδωνυμοποίησης.

Τα ψευδωνυμοποιημένα δεδομένα μπορούν να συνδεθούν ξανά με το άτομο εάν το ψευδώνυμο συνδυάζεται με πρόσθετες πληροφορίες, όπως ένα κλειδί ψευδωνυμοποίησης (που αποτελεί μία πληροφορία που επιτρέπει την αντιστροφή της ψευδωνυμοποίησης) ή άλλα δεδομένα αναγνώρισης. Επιπλέον, ενώ η ψευδωνυμοποίηση μπορεί να βοηθήσει στην προστασία των προσωπικών δεδομένων, δεν είναι αλάνθαστη και εξακολουθεί να υπάρχει κίνδυνος επαναπροσδιορισμού. Ως εκ τούτου, είναι σημαντικό να λαμβάνονται υπόψη και άλλα μέτρα ασφαλείας, όπως η κρυπτογράφηση και οι έλεγχοι πρόσβασης, σε συνδυασμό με την ψευδωνυμοποίηση για να διασφαλιστεί το υψηλότερο επίπεδο προστασίας των δεδομένων αυτών.

2.4.2 Αναγκαιότητα και Οφέλη Της Ψευδωνυμοποίησης

Αναλύοντας τα παραπάνω αντιλαμβανόμαστε τη μεγάλη αναγκαιότητα που προκύπτει για προστασία των φυσικών προσώπων σε περιπτώσεις επεξεργασίας ή κοινής χρήσης των δεδομένων τους [17]. Από τα κυριότερα πλεονεκτήματα της ψευδωνυμοποίησης, όταν υλοποιείται σωστά, είναι να αποκρύψει την ταυτότητα της οντότητας που περιγράφει μέσα σε συγκεκριμένα σύνολα δεδομένων, έτσι ώστε τα αποτελέσματα να μη μπορούν να συνδεθούν άμεσα με αυτή από τρίτους. Αντικαθιστώντας λοιπόν τα αναγνωριστικά γνωρίσματα με ψευδώνυμα, προσφέρεται μεγαλύτερη ασφάλεια καθώς και ένα επίπεδο δυσκολίας από τρίτους σε περιπτώσεις όπως η διαρροή των δεδομένων. Έτσι, κάτω από έναν πολύ καλό σχεδιασμό ψευδωνυμοποίησης ο υπεύθυνος επεξεργασίας των δεδομένων είναι και ο μόνος που μπορεί να επαναπροσδιορίσει τα αναγνωριστικά στοιχεία, χωρίς τη χρήση άλλων πληροφοριών.

Η ψευδωνυμοποίηση μπορεί να χρησιμοποιηθεί σε αρκετές περιπτώσεις, όπως της υγειονομικής περίθαλψης, της έρευνας, καθώς και σε πολλές άλλες περιπτώσεις όπου υπάρχει η ανάγκη προστασίας της ιδιωτικότητας και της αποτροπής ταυτοποίησης φυσικών προσώπων. Εξάλλου, στην πράξη, πολλές φορές καθίσταται αναγκαία για τη συμμόρφωση με νομικές απαιτήσεις προστασίας δεδομένων, όπως ιδίως με την αρχή της ελαχιστοποίησης των δεδομένων που είδαμε νωρίτερα ότι προβλέπεται στο άρθρο 5 του GDPR – ενώ, εξάλλου, δεν είναι τυχαίο ότι ο GDPR την αναφέρει ρητώς, ως ενδεικτικό μέτρο προστασίας που πρέπει να λαμβάνεται υπόψη, τόσο για την προστασία των δεδομένων ήδη από το σχεδιασμό (άρθρο 25 του GDPR) όσο και για την ασφάλεια της επεξεργασίας (άρθρο 32 του GDPR).

Επομένως μπορούμε να πλέον να αντιληφθούμε τα οφέλη της ψευδωνυμοποίησης πέραν της συμμόρφωσης με τη νομοθεσία και του κανονισμούς , όπως την απόκρυψη της ταυτότητας του υποκειμένου των δεδομένων από τρίτες μη εξουσιοδοτημένες οντότητας (πέραν του υπεύθυνου επεξεργασίας τους και όσους έχει ορίσει) προστατεύοντας έτσι την ασφάλεια και την ιδιωτικότητα των φυσικών προσώπων καθώς και άλλων γνωρισμάτων τα οποία θέλει να αποκρύψει.

2.4.3 Αναγνωριστικά Γνωρίσματα

Όπως έχουμε παρατηρήσει μέχρι τώρα, όλη η διαδικασία της ψευδωνυμοποίησης πραγματοποιείται γύρω από τα αναγνωριστικά πεδία ή γνωρίσματα μέσα σε ένα σύνολο δεδομένων. Έχουμε κατανοήσει πλέον ότι πρόκειται για πληροφορίες οι οποίες συνδέονται άμεσα με τα φυσικά πρόσωπα και μάλιστα μπορούν να το ταυτοποιήσουν χωρίς περεταίρω πληροφορίες, άρα μιλάμε για μοναδικές τιμές οι οποίες χαρακτηρίζουν άμεσα ένα φυσικό πρόσωπο, όπως ονοματεπώνυμο, Αριθμός Κοινωνικής Ασφάλισης, Αριθμός Φορολογικού Μητρώου κλπ.

Επίσης υπάρχουν και δεδομένα σύνθετου τύπου που μπορούν να οριστούν ως αναγνωριστικά (φωτογραφίες, δακτυλικά αποτυπώματα κλπ.), καθώς και αυτά μπορούν να τα ταυτοποιήσουν κάποιο φυσικό πρόσωπο. Επιπλέον ο συνδυασμός άλλων δεδομένων (συνδυασμός ΤΚ, φύλου, ηλικίας) μπορεί επίσης να ταυτοποιήσει κάποιο φυσικό πρόσωπο μέσα σε ένα σύνολο δεδομένων. Ως εκ τούτου όταν εξετάζουμε κατά πόσο κάποια πληροφορία μπορεί να χαρακτηριστεί ως αναγνωριστικό ή όχι, θα πρέπει να λαμβάνουμε υπόψη τόσο την άμεση όσο και την έμμεση αναγνώριση και ταυτοποίηση του προσώπου αυτού [16].

Καταλήγοντας μπορούμε να επιβεβαιώσουμε ότι ψευδωνυμοποιώντας ένα μόνο αναγνωριστικό μέσα σε ένα σύνολο δεδομένων, δεν αποτελεί απαραίτητα μία καλή πρακτική και θα πρέπει να εξετάσουμε και άλλες μεθόδους.

2.4.4 Μοντέλα Και Τεχνικές Επιθέσεων

Ένας από τους κυριότερους στόχους της ψευδωνυμοποίησης όπως αναφέραμε και παραπάνω, είναι η αντικατάσταση των αναγνωριστικών στοιχείων μέσα σε ένα σύνολο δεδομένων με απώτερο σκοπό να προστατέψουμε το υποκείμενο των δεδομένων αυτών και να μετριάσουμε τη

σύνδεση των ψευδωνυμοποιημένων δεδομένων με τα αρχικά σε περίπτωση επίθεσης ή διαρροής των δεδομένων αυτών [18].

Οι επιθέσεις αυτές μπορεί να είναι εσωτερικές, όπως για παράδειγμα κάποια οντότητα εντός του οργανισμού, που έχει τη κατάλληλη πρόσβαση στα δεδομένα αυτά και στόχο έχει να βλάψει τον οργανισμό ή να εκθέσει κάποιο φυσικό πρόσωπο. Στη περίπτωση αυτή πρέπει να είναι σαφώς ορισμένα τα δικαιώματα των χρηστών που μπορούν να έχουν πρόσβαση στα δεδομένα αυτά. Ως εσωτερικές οντότητες συμπεριλαμβάνονται και τρίτοι συνεργάτες εκτός του οργανισμού που έχουν δικαιοδοσία πρόσβασης στα δεδομένα.

Από την άλλη έχουμε και τις εξωτερικές επιθέσεις. Εδώ έχουμε να κάνουμε με εξωτερικούς παράγοντες, όπως για παράδειγμα οντότητες που υπό φυσιολογικές συνθήκες δεν θα έπρεπε να έχουν καν πρόσβαση στα δεδομένα αυτά. Τα κίνητρα μπορεί να είναι τα ίδια με παραπάνω, όπως η αναγνώριση κάποιου προσώπου ή ο πλήρης επαναπροσδιορισμός του αρχικού πίνακα δεδομένων αναζητώντας για το μυστικό κλειδί της ψευδωνυμοποίησης, η προσπάθεια να βλάψουν τον οργανισμό ή να αποκαλύψουν κάποια δεδομένα για δικούς τους σκοπούς.

Το μυστικό κλειδί αποτελεί το μέσο με το οποίο επιτεύχθηκε η ψευδωνυμοποίηση και αυτό είναι που στοχεύουν συνήθως οι επιτιθέμενοι. Πρόκειται είτε για κάποιον πίνακα αντιστοίχισης του αρχικού γνωρίσματος με το ψευδώνυμο, είτε για κάποιο πραγματικό κλειδί που χρησιμοποιήθηκε βάσει συγκεκριμένων τεχνικών. Αν οι επιτιθέμενοι πετύχουν το σκοπό τους είναι σε θέση να ανακτήσουν πλήρως όλο τον αρχικό πίνακα.

Όσον αφορά τις τεχνικές τώρα των επιθέσεων στα ψευδωνυμοποιημένα δεδομένα, έχουμε τρείς κύριες τεχνικές και είναι η επίθεση βάσει λεξικού (dictionary attack), η επίθεση εξαντλητικής αναζήτησης (brute force attack) και η επίθεση με εικασία (guesswork) [18].

Οι επιθέσεις brute force αποτελούν επιθέσεις στα ψευδώνυμα με πολλούς συνδυασμούς χαρακτήρων ή ASCII κωδικών ώστε να αποκαλύψουμε κάποιο ψευδώνυμο σε αρχικό γνώρισμα. Αποτελεί μία διαδικασία αρκετά χρονοβόρα καθώς και απαιτητική. Είναι αρκετά λειτουργική επίθεση σε μικρού μεγέθους σύνολα δεδομένων, ωστόσο σε μεγάλες βάσεις αποτελεί μεγάλη πρόκληση. Στις επιθέσεις τέτοιου τύπου, ο επιτιθέμενος θα πρέπει να γνωρίζει τη συνάρτηση της ψευδωνυμοποίησης για να πετύχει το σκοπό.

1. Στη περίπτωση των επιθέσεων λεξικού, ο επιτιθέμενος μπορεί να χρησιμοποιήσει μια ήδη υπάρχουσα λίστα ή βάση με ψευδώνυμα και αντιστοιχίες σε γνωρίσματα, όπως για παράδειγμα ASCII αντιστοιχίες. Κάθε φορά που κάνει απόπειρα κάποιας επίθεσης μπορεί να ανατρέξει στη βάση αυτή με τη πιθανότητα να αποκαλύψει κάποιο ψευδώνυμο. Για τη δημιουργία του λεξικού αυτού έχουν προηγηθεί σε παλαιότερες περιπτώσεις επιθέσεις brute force δημιουργώντας τη λίστα αυτή.

2. Η επίθεση με εικασία προϋποθέτει ότι ο επιτιθέμενος γνωρίζει κάποιες πληροφορίες σχετικά με το σύνολο των δεδομένων ή το υποκείμενο που βρίσκεται μέσα σε αυτό, για να μπορέσει να επιτύχει την αντιστοίχιση. Επίσης αυτό επηρεάζεται από τη τεχνική αλλά και τη πολιτική ψευδωνυμοποίησης που χρησιμοποιήθηκε, καθώς και τη συχνότητα εμφάνισης του ίδιου αναγνωριστικού μέσα στον πίνακα. Η γνώση της τεχνικής που χρησιμοποιήθηκε δεν είναι απαραίτητη, καθώς και δεν τον επηρεάζει ούτε το μέγεθος του πίνακα.

2.4.5 Τεχνικές Ψευδωνυμοποίησης

Όπως αναφέραμε παραπάνω, η ψευδωνυμοποίηση αποτελεί αντικατάσταση των αναγνωριστικών στοιχείων με ψευδώνυμα σε μία προσπάθεια απόκρυψης των αναγνωριστικών γνωρισμάτων. Για την επίτευξη της υπάρχουν διαφορετικές τεχνικές, η κάθε μία αποτελούμενη από συγκεκριμένα χαρακτηριστικά και προσεγγίσεις, προτερήματα καθώς και περιορισμούς [18] [17] [16].

Παρακάτω παρουσιάζονται οι βασικότερες από τις «κλασικές» τεχνικές με τις οποίες μπορούμε να πετύχουμε τον σκοπό αυτό, ωστόσο υπάρχουν και άλλες πιο προηγμένες τεχνικές, κάποιες από τις οποίες θα εξεταστούν σε επόμενο κεφάλαιο.

1. Αρχικά έχουμε τον κοινό μετρητή. Πρόκειται για την απλούστερη τεχνική ψευδωνυμοποίησης. Στη περίπτωση αυτή τα αναγνωριστικά που θέλουμε να ψευδωνυμοποιήσουμε αντικαθίστανται από ένα απλό νούμερο που ακολουθεί την ιδέα ενός μετρητή. Μπορεί να ξεκινήσει από το 0 και σε κάθε επόμενο αναγνωριστικό ο μετρητής αυξάνεται κατά ένα. Ως εκ τούτου γνωρίζουμε πως δεν πρόκειται να επαναληφθεί ποτέ το ίδιο ψευδώνυμο.

Σε ένα σύστημα με δεδομένα μικρού μεγέθους μπορεί να λειτουργήσει αποτελεσματικά και λόγω της απλότητάς τους μπορεί να φανεί αρκετά χρήσιμο, ωστόσο δεν συμβαίνει το ίδιο σε μεγάλες ή και περίπλοκες βάσεις δεδομένων, καθώς μπορούν να προκύψουν θέματα αρίθμησης και προκύπτει η ανάγκη αποθήκευσης ενός ακόμα πίνακα αντιστοίχισης των ψευδωνύμων. Τα ψευδώνυμα που παράγει δεν συνδέονται σε καμία περίπτωση με το αρχικό αναγνωριστικό, παρ' όλα αυτά μιλάμε για μία ακολουθία που μπορεί να δώσει πληροφορίες για τη σειρά των δεδομένων μέσα στο σύνολο, όπως για παράδειγμα αν αυτά είναι ταξινομημένα κατά αλφαβητική σειρά.

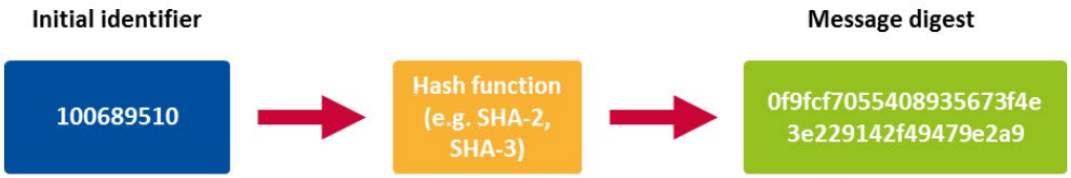
2. Στη συνέχεια έχουμε τη Γεννήτρια τυχαίων αριθμών (RNG). Πρόκειται για ένα μηχανισμό που παράγει τυχαίους αριθμούς ίσους με όσους χρειαζόμαστε για τα αναγνωριστικά μας. Μοιάζει να μοιράζεται αρκετά χαρακτηριστικά με την μέθοδο του μετρητή, ωστόσο διαφέρει στην αρίθμηση καθώς παράγει τυχαίους αριθμούς που θα αντικαταστήσουν τα αναγνωριστικά και όχι ακολουθία αριθμών.

   Καθίσταται η ίδια ανάγκη για ξεχωριστό πίνακα συσχέτισης των ψευδωνύμων με το αναγνωριστικό και όπως και στον μετρητή τα ψευδώνυμα είναι πλήρως διαχωρισμένα με το αναγνωριστικό, πέραν της σύνδεσής τους στον πίνακα αντιστοίχισης.

   Ένα μειονέκτημα της μεθόδου αυτής είναι ότι παράγοντας τυχαίους αριθμούς ελλοχεύει ο κίνδυνος της «σύγκρουσης», της πιθανότητας δηλαδή να παραχθούν για δύο διαφορετικά αναγνωριστικά (δηλαδή για δύο διαφορετικά πρόσωπα) δύο όμοια ψευδώνυμα (δηλαδή ο ίδιος αριθμός και για τα δύο πρόσωπα), κάτι που θα προκαλούσε σύγχυση και θα έθετε τον αλγόριθμό ως μη λειτουργικό, καθώς κατά κανόνα είναι σημαντικό τα ψευδώνυμα να είναι μοναδικά. Σε τέτοιες περιπτώσεις η γεννήτρια δεν είναι εντελώς τυχαία αλλά ψευδο-τυχαία δεσμεύοντας έτσι πως κάθε ψευδώνυμο που θα παραχθεί θα είναι μοναδικό.

3. Προχωρώντας σε πιο περίπλοκες τεχνικές ψευδωνυμοποίησης έχουμε αρχικά τη κρυπτογραφική συνάρτηση κατακερματισμού (hash function). Πρόκειται για συναρτήσεις οι οποίες είναι ειδικά διαμορφωμένες έτσι ώστε να λαμβάνουν μία είσοδο ανεξαρτήτου μεγέθους (για παράδειγμα το μήνυμα ή το αναγνωριστικό) και να εξάγουν ένα ψευδώνυμο σταθερού πάντα μεγέθους ανεξάρτητο της εισόδου, ανάλογα τη συνάρτηση. Το ψευδώνυμο αυτό αν και με πρώτη ματιά δείχνει τυχαίο, στη

πραγματικότητα δεν είναι και κάθε φορά η ίδια είσοδος θα δίνει πάντα ακριβώς την ίδια έξοδο, όσο χρησιμοποιείται ο ίδιο αλγόριθμος.

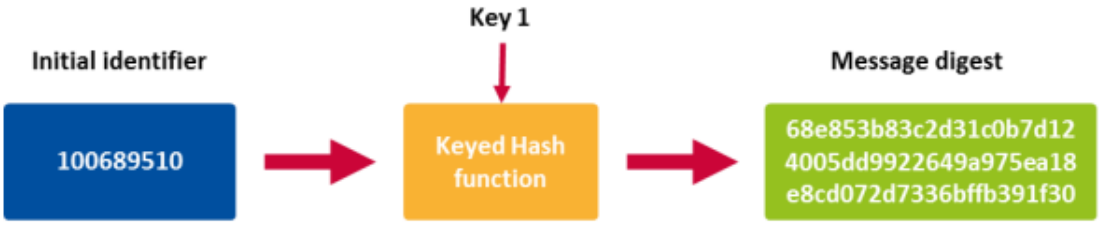


Εικόνα 2.1: Ψευδωνυμοποίηση με συνάρτηση Κατακερματισμού [18]

Προτερήματα της μεθόδου αυτής είναι, ότι αρχικά οι συναρτήσεις κατακερματισμού είναι μιας κατεύθυνσης (one-way), εννοώντας ότι δεν υπάρχει άμεσος τρόπος που να μπορούμε από το ψευδώνυμο να παράξουμε ξανά το αρχικό αναγνωριστικό διότι είναι ειδικά διαμορφωμένες για αυτό το σκοπό. Επίσης επιλέγοντας μία καλή συνάρτηση κατακερματισμού λύνουμε και το πρόβλημα της σύγκρουσης, καθώς είναι εξαιρετικά δύσκολο να προκύψει το ίδιο ψευδώνυμο από δύο διαφορετικές εισόδους.

Αν και δείχνει ως αρκετά ασφαλής προσέγγιση, έχει αρκετούς κινδύνους και αδυναμίες συγκριτικά με τις προηγούμενες καθώς και έμμεσους τρόπους να ταυτοποιήσουμε το αρχικό αναγνωριστικό, οι οποίες θα μελετηθούν αναλυτικότερα παρακάτω καθώς και τρόποι να ξεπεράσουμε τους κινδύνους αυτούς.

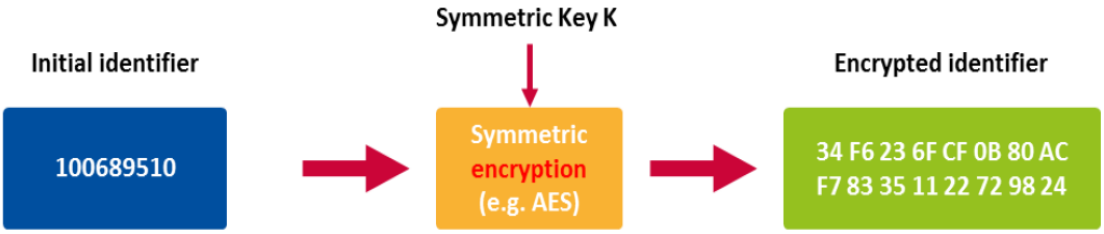
4. Στη συνέχεια έχουμε τη τεχνική του Message Authentication Code (MAC) η οποία είναι παρόμοια με τις συναρτήσεις κατακερματισμού με τη διαφορά ότι στη διαδικασία αυτή υπεισέρχεται και ένα μυστικό κλειδί ώστε να παραχθεί το ψευδώνυμο. Αν και στη παραπάνω μέθοδο υπάρχει έμμεσος όπως αναφέραμε τρόπος να αναγνωρίσουμε το αρχικό μήνυμα, στη περίπτωση αυτή δεν ισχύει το ίδιο αν δεν γνωρίζουμε το μυστικό κλειδί.



Εικόνα 2.2: Ψευδωνυμοποίηση με κώδικας αυθεντικοποίησης μηνύματος [18]

Η μέθοδος αυτή παρέχει πολύ μεγάλη ασφάλεια έναντι στον κίνδυνο αντιστοίχισης του ψευδωνύμου με το αναγνωριστικό, όσο το κλειδί δεν είναι γνωστό. Σε περίπτωση διαρροής του κλειδιού παύει να ισχύει οποιαδήποτε ασφάλεια, προκύπτοντας έτσι η ανάγκη περιπλοκότητας τους κλειδιού καθώς και διασφάλισής του. Επίσης στη περίπτωση που το κλειδί χαθεί, δεν είμαστε πλέον σε θέση να επαναπροσδιορίσουμε το αρχικό μήνυμα.

5. Κλείνοντας με τις κυριότερες τεχνικές ψευδωνυμοποίησης έχουμε τη συμμετρική κρυπτογράφηση των αναγνωριστικών του υποκειμένου. Αποτελεί επίσης μία αρκετά αποτελεσματική μέθοδο για τη δημιουργία ψευδωνύμων.



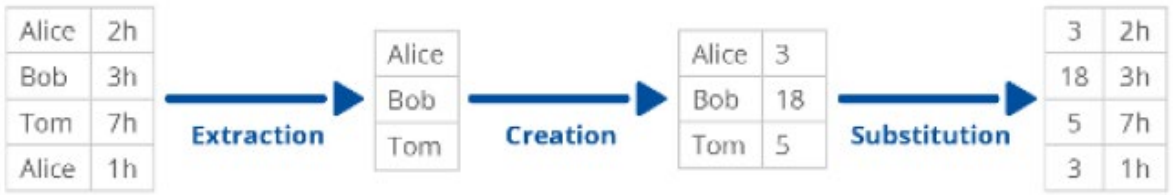
Εικόνα 2.3: Ψευδωνυμοποίηση Με Συμμετρική Κρυπτογράφηση [18]

Ως πρότυπο συμμετρικής κρυπτογράφησης συνήθως χρησιμοποιείται ο AES, ακολουθώντας γενικότερα την χρήση αλγορίθμων τμημάτων, έχοντας ως πρόταση τη χρήση κλειδιού 256 bits . Αφού έχουμε ψευδωνυμοποιήσει το αναγνωριστικό σε ένα κρυπτογραφημένο μήνυμα, απαιτείται το μυστικό κλειδί για αποκρυπτογραφηθεί. Στη περίπτωση αυτή μόνο όσοι έχουν πρόσβαση στο κλειδί μπορούν να επαναπροσδιορίσουν το αναγνωριστικό. Ένα σημαντικό μειονέκτημα σε αυτό είναι ότι αν το κλειδί καταστραφεί ή χαθεί, δεν υπάρχει τρόπος επαναπροσδιορισμού των αρχικών αναγνωριστικών πεδίων.

2.4.6 Πολιτικές Ψευδωνυμοποίησης

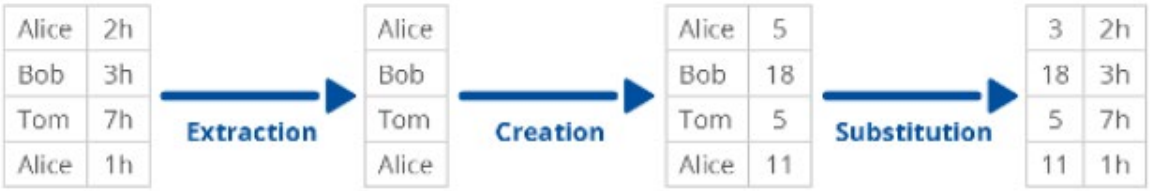
Στο σημείο αυτό θα πρέπει να αναφερθούν και κάποιες πολιτικές που ακολουθεί μία επιτυχημένη ψευδωνυμοποίηση. Οι πολιτικές αυτές ορίζονται ανάλογα τις ανάγκες που έχει θέσει ο υπεύθυνος επεξεργασίας των δεδομένων από το σχεδιασμό [18].

Αρχικά έχουμε τη ντετερμινιστική ψευδωνυμοποίηση. Στη περίπτωση αυτή, σε κάθε σύνολο ή βάση δεδομένων όπου εμφανίζεται ένα αναγνωριστικό με το ίδιο όνομα περισσότερες από μία φορές, ψευδωνυμοποιείται πάντα με το ίδιο ψευδώνυμο, είτε εμπεριέχεται στο ίδιο σύνολο δεδομένων, είτε πρόκειται για διαφορετικές βάσεις.



Εικόνα 2.4: Ντετερμινιστική ψευδωνυμοποίηση [18]

Έπειτα ακολουθεί η τυχαία ψευδωνυμοποίηση μέσα σε ένα έγγραφο. Σε αντίθεση με τη παραπάνω πολιτική, ίδια αναγνωριστικά αντικαθίστανται από διαφορετικά ψευδώνυμα μέσα σε ένα σύνολο δεδομένων. Ως αποτέλεσμα έχουμε ότι τη δεύτερη φορά που θα εμφανιστεί το ίδιο αρχικό αναγνωριστικό θα αντικατασταθεί με διαφορετικό ψευδώνυμο.



Εικόνα 2.5: Ψευδωνυμοποίηση μέσα σε έγγραφο [18]

Τέλος η πλήρως τυχαιοποιημένη ψευδωνυμοποίηση, η οποία αποτελεί κάποια εξέλιξη της προηγούμενης πολιτικής, ωστόσο αφορά κυρίως δεδομένα σε διαφορετικές βάσεις, ενώ δεν

προσφέρει τίποτα διαφορετικό στο ίδιο σύνολο δεδομένων, ωστόσο κατά τη παρούσα διατριβή ξεπερνά το πεδίο της έρευνάς μας.

2.4.7 Ανάκτηση Δεδομένων

Έως τώρα έχουμε εξετάσει διάφορες πτυχές της ψευδωνυμοποίησης καθώς και τρόπους υλοποίησής της. Ωστόσο στοχεύοντας στη λειτουργικότητα της όλης διαδικασίας θα πρέπει να υπάρχει και ένας μηχανισμός ανάκτησης. Ένα χαρακτηριστικό παράδειγμα ανάγκης του μηχανισμού αυτού είναι να ανατρέξουμε τα πραγματικά αναγνωριστικά δεδομένα, ή η ανάγκη επικύρωσης της σωστής ψευδωνυμοποίησης. Μια άλλη περίπτωση είναι η ενημέρωση των εμπλεκομένων υποκειμένων βάση των κανονισμών του GDPR σε περίπτωση κάποιας διαρροής των δεδομένων τους [18].

Ο μηχανισμός αυτός μπορεί να διαφέρει ανά περιπτώσεις ανάλογα τη πολυπλοκότητα της ψευδωνυμοποίησης, παρ' όλα αυτά ως βάση έχει μια κεντρική ιδέα. Αρχικά έχουμε ένα αναγνωριστικό, ένα κλειδί(τρόπος ψευδωνυμοποίησης) και ένα ψευδώνυμο το οποίο προέκυψε από τα προηγούμενα δύο. Στις περισσότερες περιπτώσεις ψευδωνυμοποίησης, για να επιτύχουμε ανάκτηση των αρχικών γνωρισμάτων, χρειάζεται από το σχεδιασμό της ψευδωνυμοποίησης ένας πίνακας αντιστοίχισης. Ο πίνακας αυτός φυλάσσεται ξεχωριστά από τους υπόλοιπους για λόγους ασφάλειας και αποφυγής ταυτοποίησης από τρίτους περιλαμβάνοντας σε κάποιες περιπτώσεις την αντιστοίχιση των γνωρισμάτων και των ψευδωνύμων, λειτουργώντας έτσι ως κλειδί της ψευδωνυμοποίησης. Στη περίπτωση όπως της κρυπτογράφησης ως τεχνική ψευδωνυμοποίησης δεν απαιτείται τέτοιος πίνακας, παρά μόνο ο κρυπτογραφικός αλγόριθμος και το κλειδί του.

Method	Recovery based on pseudonym
Counter	Mapping table
Random Number Generator	Mapping table
Hash function	Mapping table
Message Auth. Codes	Mapping table
Encryption	Decryption

Πίνακας 2.6: Κλασικές Μέθοδοι Ψευδωνυμοποίησης [18]

2.5 Ανωνυμοποίηση

Στη προηγούμενη ενότητα συναντήσαμε τρόπους και μεθόδους που μπορούμε να μετριάσουμε την αποκάλυψη προσωπικών δεδομένων και γενικότερα επιλεγμένων πληροφοριών σε τρίτες μη εξουσιοδοτημένες οντότητες και πιο συγκεκριμένα τη ψευδωνυμοποίηση. Η μέθοδος αυτή, αν και δείχνει αρκετά αξιόπιστη και σε αρκετές περιπτώσεις αναγκαία, ενδεχομένως να μην μπορεί πάντα να προστατέψει πλήρως την ιδιωτικότητα των υποκειμένων μέσα σε ένα σύνολο δεδομένων από μόνη της. Εξάλλου, όπως ρητά αναφέρει και ο GDPR, τα ψευδωνυμοποιημένα δεδομένα θα πρέπει να θεωρούνται προσωπικά δεδομένα – ειδικότερα, στην αιτιολογική σκέψη 26 του GDPR, αναφέρεται ότι «τα δεδομένα προσωπικού χαρακτήρα που έχουν υποστεί ψευδωνυμοποίηση, η οποία θα μπορούσε να αποδοθεί σε φυσικό πρόσωπο με τη χρήση συμπληρωματικών πληροφοριών, θα πρέπει να θεωρούνται πληροφορίες σχετικά με ταυτοποιήσιμο φυσικό πρόσωπο». Ενδεχομένως σε κάποιες περιπτώσεις, από τη φύση της επεξεργασίας, να είναι απαραίτητο τα δεδομένα να μην επιτρέπουν κατά κανένα τρόπο την αναγνώριση των προσώπων στα οποία αναφέρονται από τον οποιονδήποτε, το οποίο σημαίνει ότι θα έχουμε ανώνυμα δεδομένα. Τα ανώνυμα δεδομένα δεν θεωρούνται προσωπικά δεδομένα. Ειδικότερα, πάλι στην αιτιολογική σκέψη 26 του GDPR, αναφέρεται ότι «για να κριθεί κατά πόσον ένα φυσικό πρόσωπο είναι ταυτοποιήσιμο, θα πρέπει να λαμβάνονται υπόψη όλα τα μέσα τα οποία είναι ευλόγως πιθανό ότι θα χρησιμοποιηθούν, όπως για παράδειγμα ο διαχωρισμός του, είτε από τον υπεύθυνο επεξεργασίας είτε από τρίτο για την άμεση ή έμμεση εξακρίβωση της ταυτότητας του φυσικού προσώπου. Για να διαπιστωθεί κατά πόσον κάποια μέσα είναι ευλόγως πιθανό ότι θα χρησιμοποιηθούν για την εξακρίβωση της ταυτότητας του φυσικού προσώπου, θα πρέπει να λαμβάνονται υπόψη όλοι οι αντικειμενικοί παράγοντες, όπως τα έξοδα και ο χρόνος που απαιτούνται για την ταυτοποίηση, λαμβανομένων υπόψη της τεχνολογίας που είναι διαθέσιμη κατά τον χρόνο της επεξεργασίας και των εξελίξεων της τεχνολογίας. Οι αρχές της προστασίας δεδομένων δεν θα πρέπει συνεπώς να εφαρμόζονται σε ανώνυμες πληροφορίες, δηλαδή πληροφορίες που δεν σχετίζονται προς ταυτοποιημένο ή ταυτοποιήσιμο φυσικό πρόσωπο ή σε δεδομένα προσωπικού χαρακτήρα που έχουν καταστεί ανώνυμα κατά τρόπο ώστε η ταυτότητα του υποκειμένου των δεδομένων να μην μπορεί ή να μην μπορεί πλέον να εξακριβωθεί».

Στη προσπάθεια να χαρακτηρίσουμε τα προσωπικά δεδομένα ως ανώνυμα όπως αναφέραμε και παραπάνω βάση του GDPR, θα πρέπει πρώτα να εξασφαλίσουμε ότι δεν μπορούν να ταυτοποιήσουν το υποκείμενο με κανέναν άμεσο ή έμμεσο τρόπο, λαμβάνοντας υπόψη κάθε δυνατό μέσο που ευλόγως μπορεί να θεωρήσει ότι δύναται να χρησιμοποιήσει ένας οποιοδήποτε

τρίτος. Με γνώμονα λοιπόν τη πλήρη συμμόρφωση με τους κανονισμούς αυτούς, αλλά και τη προστασία των προσωπικών δεδομένων φυσικών προσώπων ή ευαίσθητων πληροφοριών, τα δεδομένα θα πρέπει να περνούν από περαιτέρω επεξεργασία πριν αξιοποιηθούν από τρίτες οντότητες, εξασφαλίζοντας την ακεραιότητά τους και προστατεύοντας τα υποκείμενα των δεδομένων από διαρροές. Η διαδικασία αυτή είναι η ανωνυμοποίηση.

2.5.1 Ορισμός Της Ανωνυμοποίησης

Η ανωνυμοποίηση δεδομένων πρόκειται για επεξεργασία με την οποία τα προσωπικά δεδομένα τροποποιούνται με τέτοιο τρόπο, ώστε να μη μπορούν πλέον να αποδοθούν σε φυσικό πρόσωπο με άμεσο ή έμμεσο τρόπο [19]. Αποτελεί πολλές φορές διαδικασία επιτακτική σε περιπτώσεις κοινής χρήσης και γενικότερα διαμοιρασμού των προσωπικών δεδομένων στοχεύοντας στη μείωση του ρίσκου αποκάλυψης ή σύνδεσης ευαίσθητων πληροφοριών με το υποκείμενο που περιγράφουν. Με το τρόπο αυτό επιτρέπει την ανάλυση και την επεξεργασία προσωπικών δεδομένων χωρίς τη πιθανότητα ταυτοποίησής τους με κάποιο άτομο. Αυτό μπορεί να επιτευχθεί με διάφορους τρόπους όπως η απόκρυψη κάποιων γνωρισμάτων ή η τροποποίηση και η γενίκευση κάποιων άλλων μέσα σε ένα σύνολο δεδομένων. Στόχος της είναι να διατηρηθεί η χρησιμότητα των δεδομένων αυτών για ανάλυση και έρευνα, ενώ ταυτόχρονα να προστατεύεται η ιδιωτικότητα των προσώπων που αναφέρονται στα δεδομένα αυτά.

2.5.2 Τύποι Γνωρισμάτων

Όπως έχουμε κατανοήσει έως τώρα η διαδικασία της ανωνυμοποίησης στοχεύει στη προστασία των προσωπικών δεδομένων από τη σύνδεση αυτών ή πρόσθετων πληροφοριών μιας εγγραφής με το φυσικό πρόσωπο που περιγράφουν. Για το λόγο αυτό θα πρέπει να είμαστε σε θέση να ξεχωρίσουμε τους διαφορετικούς τύπους γνωρισμάτων καθώς και των ιδιοτήτων του κάθε ενός και πώς αυτό μπορεί να ταυτοποιήσει το υποκείμενο ή όχι. Έτσι λοιπόν ο κάθε τύπος γνωρίσματος υφίσταται διαφορετική επεξεργασία κατά τη διαδικασία της ανωνυμοποίησης ανάλογη των ιδιοτήτων του γνωρίσματος και της πληροφορίας που μπορεί να μας δώσει [20] [21].

Κάποια από αυτά τα γνωρίσματα μπορούν να ταυτοποιήσουν άμεσα κάποιο φυσικό πρόσωπο, άλλα απαιτούν συνδυασμό τους ώστε να αναγνωρίσουν κάποιον με έμμεσο τρόπο ενώ άλλα αποτελούν ευαίσθητες εμπιστευτικές πληροφορίες που πρέπει να προστατευτούν. Τα γνωρίσματα αυτά χωρίζονται σε ξεχωριστές κατηγορίες και αντιμετωπίζονται ανάλογα.

1. Αναγνωριστικά (Identifier):

   Αφορά δεδομένα τα οποία μπορούν να αναγνωρίσουν άμεσα μία οντότητα και μπορεί να χρησιμοποιήσει ο επιτιθέμενος για να ταυτοποιήσει απευθείας κάποιο πρόσωπο. Τέτοια δεδομένα είναι τα ονοματεπώνυμα, αριθμοί κοινωνικής ασφάλισης κλπ.

2. Ψευδο – Αναγνωριστικά (Quasi-Identifiers):

   Επίσης πρόκειται για γνωρίσματα που μπορούν να ταυτοποιήσουν κάποιο πρόσωπο, ωστόσο μόνο συνδυαστικά με άλλα ψευδο-αναγνωριστικά. Ένα παράδειγμα ταυτοποίησης με τη χρήση ψευδο-αναγνωριστικών είναι ο συνδυασμός ηλικίας, φύλου, ταχυδρομικού κώδικα. Παρακάτω θα δούμε τεχνικές που μπορούμε να μετριάσουμε επιθέσεις σε αυτό το συνδυασμό.

3. Εμπιστευτικά – Ευαίσθητα γνωρίσματα (Confidential Attributes):

   Πρόκειται για πληροφορίες άκρως ευαίσθητες που αφορούν κάποια οντότητα και πρέπει να προστατευτούν. Στη περίπτωση που τέτοιες πληροφορίες εκτεθούν δημόσια ταυτοποιώντας κάποιο φυσικό πρόσωπο θα μπορούσαν να έχουν τεράστιες συνέπειες. Από μόνες τους σαν πληροφορίες δεν αποτελούν κάποιον κίνδυνο σε αρκετές περιπτώσεις, ωστόσο συνδυαστικά με ονόματα ή άλλες οντότητες αποτελούν καταπάτηση ευαίσθητων προσωπικών δεδομένων. Ένα παράδειγμα είναι η υγεία ενός ασθενούς, που μπορεί να πρόκειται για κάποια νόσο της οποίας είναι φορέας και δεν θέλει να αποκαλυφτεί.

4. Μη εμπιστευτικά γνωρίσματα(Non-confidential Attributes):

   Πρόκειται για επιπλέον πληροφορίες οι οποίες δεν είναι εμπιστευτικές, αλλά ούτε έχουν και καμία απολύτως χρήση αναγνώρισης είτε με άμεσο αλλά ούτε και με έμμεσο τρόπο.

Παρά τις τέσσερις κατηγορίες γνωρισμάτων που εξετάσαμε παραπάνω, η ανωνυμοποίηση ουσιαστικά εστιάζει μόνο σε δύο από αυτές, τα εμπιστευτικά/ευαίσθητα γνωρίσματα και τα ψευδο-αναγνωριστικά. Όσο για τα αναγνωριστικά και τα μη εμπιστευτικά γνωρίσματα, τα πρώτα αφαιρούνται εντελώς κατά τη διαδικασία της ανωνυμοποίησης (άρα, υπ' αυτήν την έννοια, επίσης έχουν κρίσιμο ρόλο για την ανωνυμοποίηση – απλά είναι τετριμμένη η διαχείρισή τους, αφού απλά απαλείφονται), ενώ τα δεύτερα δεν αποτελούν κανένα κίνδυνο.

2.5.3 Αποκάλυψη Πληροφοριών

Η ανωνυμοποίηση όπως έχουμε δει έως τώρα έχει στόχο να αποκρύψει ευαίσθητες πληροφορίες και πιο συγκεκριμένα τη σύνδεσή τους με φυσικά πρόσωπα. Αποτελεί μία διαδικασία προστασίας των υποκειμένων έναντι της αποκάλυψης εμπιστευτικών πληροφοριών, καθώς και εξαγωγής συμπερασμάτων σχετικά με πληροφορίες που τους αφορούν και πρέπει να προστατευτούν.

Υπάρχουν δύο τύποι αποκάλυψης πληροφοριών και αυτοί είναι η αποκάλυψη της ταυτότητας και η αποκάλυψη των γνωρισμάτων [22].

1. Αποκάλυψη Ταυτότητας (Identity Disclosure):

   Αποτελεί την αποκάλυψη της ταυτότητας μιας οντότητας μέσα σε ένα σύνολο δεδομένων. Ουσιαστικά μέσω αυτής ταυτοποιείται κάποιο φυσικό πρόσωπο καθώς και οι πληροφορίες που το αφορούν μέσα σε ένα σύνολο. Οι πληροφορίες μπορεί να είναι ευαίσθητες όπως για παράδειγμα ευαίσθητα προσωπικά δεδομένα ενός ασθενή που θέλει να αποκρύψει (κατάσταση της υγείας του). Ως εκ τούτου ο επιτιθέμενος γνωρίζει ακριβώς παρατηρώντας μία εγγραφή, σε ποιο πρόσωπο ανήκει και τί ευαίσθητες πληροφορίες περιλαμβάνει.

   Για παράδειγμα, αν σε ένα σύνολο δεδομένων μπορέσει κάποιος να εξάγει το συμπέρασμα ότι μία συγκεκριμένη εγγραφή αντιστοιχεί σε ένα συγκεκριμένο ταυτοποιήσιμο πρόσωπο Χ, τότε έχουμε αποκάλυψη ταυτότητας.

2. Αποκάλυψη Γνωρίσματος (Attribute Disclosure):

   Η αποκάλυψη γνωρισμάτων δεν είναι απαραίτητο να ταυτοποιήσει κάποιο φυσικό πρόσωπο σε αντίθεση με τη προηγούμενη περίπτωση. Μπορεί η εγγραφή να μην είναι ταυτοποιήσιμη ως φυσικό πρόσωπο, ωστόσο στοχεύει γενικότερα στη αποκάλυψη ευαίσθητων πληροφοριών. Μπορεί για παράδειγμα να συνδυάσει ψευδο-αναγνωριστικά και να καταλήξει σε ένα συμπέρασμα συσχέτισής τους με κάποια συγκεκριμένη ευαίσθητη πληροφορία. Γνωρίζοντας ότι κάποιο πρόσωπο πληροί αυτά τα ψευδο-αναγνωριστικά θα μπορούσε να εκτεθεί στη ευαίσθητη πληροφορία αυτή.

   Για παράδειγμα, αν σε ένα σύνολο δεδομένων δεν μπορεί κάποιος να εξάγει το συμπέρασμα σε ποια συγκεκριμένη εγγραφή αντιστοιχεί ένα συγκεκριμένο

ταυτοποιήσιμο πρόσωπο Χ, αλλά παρόλα αυτά είναι σε θέση να εξάγει μία πληροφορία για το Χ την οποία δεν γνώριζε (π.χ. ότι νόσησε από κάποια ασθένεια), τότε έχουμε αποκάλυψη γνωρίσματος.

2.5.4 Τύποι Και Μοντέλα Επιθέσεων

Καθώς ασχολούμαστε με ανωνυμοποιημένα δεδομένα, θα πρέπει να λάβουμε υπόψη και διάφορους τύπους επιθέσεων οι οποίο μπορούν να επαναπροσδιορίσουν τα υποκείμενα των δεδομένων αυτών ή να εξάγουν εμπιστευτικές πληροφορίες μέσα από ένα σύνολο ανωνυμοποιημένων δεδομένων. Παρακάτω περιγράφονται κάποιοι από τους κυριότερους τύπους αυτούς [23].

1. Επιθέσεις συσχέτισης (Linkage Attacks) :

   Οι επιθέσεις αυτές περιλαμβάνουν τον συνδυασμό των ανωνυμοποιημένων δεδομένων με εξωτερικούς πίνακες ή βάσεις δεδομένων, καθώς και εξωτερικών πληροφοριών για να μπορέσουν να ταυτοποιήσουν μία οντότητα. Αυτό μπορεί να επιτευχθεί συνδυάζοντας ψευδο-αναγνωριστικά μεταξύ διαφορετικών συνόλων ή βάσεων δεδομένων.

2. Επιθέσεις εξαγωγής συμπεράσματος (Attribute Inference Attacks):

   Στη περίπτωση αυτή οι επιθέσεις αποσκοπούν στο να εξάγουν ευαίσθητα ή εμπιστευτικά χαρακτηριστικά ατόμων, βασιζόμενα στα διαθέσιμα ανώνυμα δεδομένα προκειμένου να καταλήξουν σε κάποιο συμπέρασμα. Ουσιαστικά προσπαθούν να συμπεράνουν ευαίσθητες πληροφορίες αναλύοντας το μοτίβο ή τις ιδιότητες των δεδομένων και των γνωρισμάτων που έχουν πρόσβαση, χωρίς να χρειάζεται να αναγνωρίσουν επακριβώς σε ποιόν ανήκει η εκάστοτε εγγραφή στον ανωνυμοποιημένο πίνακα.

3. Επιθέσεις ομοιογένειας (Homogeneity Attacks):

   Οι επιθέσεις αυτές εκμεταλλεύονται το γεγονός ότι τα ανώνυμα σύνολα δεδομένων μπορεί να παρουσιάζουν ομοιογένεια σε ορισμένα χαρακτηριστικά. Οι επιτιθέμενοι μπορούν να εκμεταλλευτούν συγκεκριμένα μοτίβα ή χαρακτηριστικά που μοιράζεται μία ομάδα

ατόμων, εντοπίζοντας ευαίσθητες πληροφορίες και έτσι να εξάγουν συμπεράσματα ή να ταυτοποιήσουν πρόσωπα.

Επίσης εξετάζοντας κάποια σενάρια επιθέσεων σε σύνολα ανώνυμων δεδομένων έχουμε και ορισμένα μοντέλα επιθέσεων που αφορούν τον τρόπο που ο επιτιθέμενος μπορεί να προσεγγίσει την επίθεσή του και κατά συνέπεια να ταυτοποιήσει αναγνωριστικά/ φυσικά πρόσωπα μέσα στα δεδομένα αυτά [20] [24].

1. Αρχικά έχουμε το prosecutor μοντέλο [21] στο οποίο ο επιτιθέμενος στοχεύει μία συγκεκριμένη οντότητα μέσα στο σύνολο δεδομένων, γνωρίζοντας με σιγουριά ότι η οντότητα αυτή βρίσκεται μέσα στον ανώνυμο πίνακα, καθώς επίσης και πληροφορίες για την οντότητα αυτή όπως ονοματεπώνυμο, διεύθυνση κλπ. Από τη στιγμή που η εγγραφή αυτή περιέχει κάποιες εμπιστευτικές πληροφορίες, μπορεί να τις ανακαλύψει ο επιτιθέμενος ωστόσο μόνο για τη συγκεκριμένη οντότητα καθιστώντας σαφές τη δύναμη που έχουν τα ψευδο-αναγνωριστικά συνδυάζοντάς τα.

2. Στη συνέχεια έχουμε το μοντέλο journalist [20] [21]. Στη περίπτωση αυτή, ο επιτιθέμενος στοχεύει κάποιες συγκεκριμένες οντότητες εικάζοντας ότι μπορεί να βρίσκονται μέσα στο ανώνυμο σύνολο δεδομένων χωρίς να έχει υπόβαθρο για τα συγκεκριμένα άτομα αν βρίσκονται μέσα στα δεδομένα. Έτσι προσπαθεί να συνδυάσει διάφορα ψευδο-αναγνωριστικά σε μία προσπάθεια να ταυτοποιήσει κάποιες οντότητες.

3. Τέλος έχουμε και το μοντέλο marketer [20] στο οποίο ο επιτιθέμενος δεν στοχεύει καμία συγκεκριμένη οντότητα. Ουσιαστικά απλά προσπαθεί να ταυτοποιήσει όσο περισσότερα φυσικά πρόσωπα μπορεί μέσα από τα ανώνυμα δεδομένα.

Στο σημείο αυτό είναι σημαντικό να σημειωθεί ότι η αποτελεσματικότητα της κάθε επίθεσης εξαρτάται από παράγοντες όπως η ποιότητα των τεχνικών ανωνυμοποίησης που χρησιμοποιήθηκαν (οι οποίες παρατίθενται στη συνέχεια), το πλήθος καθώς και η μοναδικότητα των πρόσθετων πληροφοριών που εμπεριέχονται στις εγγραφές, ακόμα και το επίπεδο της γενίκευσης ή της κατάργησης γνωρισμάτων που εφαρμόστηκε στα δεδομένα κατά τη διαδικασία της ανωνυμοποίησης. Ως εκ τούτου κατά το σχεδιασμό αλλά και τη διαδικασία της ανωνυμοποίησης οι μέθοδοι και οι τεχνικές αυτής θα πρέπει να αξιολογούνται και να αξιοποιούνται ανάλογα, ώστε να επιτυγχάνεται ο ιδανικότερος μετριασμός των κινδύνων τέτοιων επιθέσεων και η διασφάλιση της προστασίας της ιδιωτικότητας των υποκειμένων.

2.5.5 Μέθοδοι Ανωνυμοποίησης

Καθώς η ανωνυμοποίηση δεδομένων αποτελεί διαδικασία μείωσης του ρίσκου αποκάλυψης ευαίσθητων πληροφοριών ή πληροφοριών που χρήζουν κάποιας προστασίας, κρατώντας έτσι ακόμη χρήσιμα τα δεδομένα προς επεξεργασία και ανάλυση, προκύπτουν κάποιες μέθοδοι που βοηθούν στην υλοποίησής της [25].

Κάποιες από τις κυριότερες μεθόδους ανωνυμοποίησης ακολουθούν παρακάτω.

1. Η χρήση μάσκας (masking) ή αλλιώς κάλυψη γνωρισμάτων αποτελεί μια μέθοδο αλλαγής ή κατάργησης συγκεκριμένων γνωρισμάτων ή μέρος τους σε μία προσπάθεια ανωνυμοποίησης. Μπορεί να εφαρμοστεί σε αναγνωριστικά, ψευδο-αναγνωριστικά καθώς και εμπιστευτικά γνωρίσματα. Στόχος έχει τη μείωση του ρίσκου διαρροής κάποια εγγραφής μέσα σε ένα σύνολο δεδομένων. Σε κάποιες περιπτώσεις με τη μέθοδο αυτή η πρόσβαση σε δεδομένα γίνεται με τροποποιημένες τιμές συγκριτικά με τα πραγματικά δεδομένα. Παραδείγματα είναι η τροποποίηση των αναγνωριστικών, η κατάργησή τους ή η πλήρης αντικατάστασή τους με ψευδώνυμα, όπως επίσης η αντικατάσταση των τελευταίων ψηφίων ενός αριθμού με (*) όπως ο Τ.Κ (13***).

2. Τα συνθετικά δεδομένα (synthetic data) αποτελούν δεδομένα που δημιουργήθηκαν μέσω κάποιο αλγορίθμου προσομοιάζοντας τα πραγματικά από κάποιο μοτίβο, ωστόσο δεν έχουν καμία σύνδεση με αυτά. Ως αποτέλεσμα αυτού ωστόσο πιθανό να έχουμε λιγότερο ακριβή δεδομένα σε σχέση με τα πραγματικά, οπότε θα πρέπει πάντα να εξετάζεται η ανάγκη της έρευνας.

3. Η γενίκευση (generalization) περιλαμβάνει διαδικασία αντικατάστασης κάποιας αριθμητικής συνήθως τιμής με μία πιο γενικευμένη εκδοχή της, όπως για παράδειγμα η ηλικία μπορεί να αντικατασταθεί με ένα εύρος ηλικιών. Επίσης κάποια ψευδο-αναγνωριστικά όπως οι ημερομηνίες μπορούν να αντικατασταθούν με περιόδους προσπαθώντας να προστατεύσουμε την αποκάλυψη κάποια εγγραφής.

4. Η προσθήκη θορύβου αποτελεί και αυτή μία μέθοδο ανωνυμοποίησης και περιλαμβάνει κάποια μικρή αλλοίωση, γνωστή και ως προσθήκη θορύβου, κυρίως αριθμητικών τιμών γνωρισμάτων ως μέρος τη διαδικασίας για την απόκρυψη της ταυτότητας των εγγραφών μέσα σε ένα σύνολο. Αυτό μπορεί για παράδειγμα να επιτευχθεί με τη στρογγυλοποίηση

κάπου αριθμού ή μικρή σχετικά αλλοίωση τιμών χωρίς όπως να επηρεάσει σε μεγάλο βαθμό το αποτέλεσμα.

2.5.6 Χρησιμότητα Έναντι Απώλειας

Ένα πολύ σημαντικό κομμάτι που θα πρέπει να λάβουμε σοβαρά υπόψη και να εξετάσουμε σχολαστικά όταν προσπαθούμε να ανωνυμοποιήσουμε προσωπικά δεδομένα είναι η χρησιμότητα τους έναντι της απώλειας πληροφορίας που θα επιφέρει η διαδικασία [22]. Η ανωνυμοποίηση όπως είδαμε μετριάζει σε μεγάλο βαθμό τη σύνδεσή των πληροφοριών με φυσικά πρόσωπα, προστατεύοντας έτσι τη ταυτότητα των προσώπων αυτών, ωστόσο αυτό έχει ως συνέπεια την απώλεια πρόσθετης πληροφορίας και σε κάποιες περιπτώσεις της πληροφορίας που εξετάζεται. Θα πρέπει λοιπόν, για να ορίσουμε μία διαδικασία ανωνυμοποίησης ως αποτελεσματική, αρχικά να μη μπορεί να συνδέσει τα ψευδο-αναγνωριστικά ή τα ευαίσθητα γνωρίσματα με τα αναγνωριστικά των υποκειμένων, αλλά σε τέτοιο βαθμό που να μπορεί να μας δώσει χρήσιμα δεδομένα και αποτελέσματα για τις ανάγκες μας. Για παράδειγμα, σε ένα σενάριο εξέτασης επιρροής κάποιας ασθένειας με βάση τις ηλικιακές ομάδες, το εύρος των ηλικιών δεν θα πρέπει να είναι πολύ μεγάλο, καθώς αυτό θα αντικρουόταν με τις ανάγκες τις έρευνας αυτής, ωστόσο αν ήταν υπερβολικά μικρό θα μπορούσε να ταυτοποιήσει έμμεσα κάποιο πρόσωπο μιας εγγραφής στο σύνολο των δεδομένων προς εξέταση.

Θα πρέπει λοιπόν να είμαστε σε θέση να υπολογίσουμε το πώς αποτιμάται η χρήσιμη πληροφορία που χάνεται κατά τη διαδικασία της ανωνυμοποίησης και αν αυτή η απώλεια είναι ανεκτή σε σχέση με τη χρησιμότητα των ανωνυμοποιημένων δεδομένων αυτών. Υπάρχουν κάποιες τεχνικές μέτρησης της απώλειας πληροφορίας όπως η κανονικοποιημένη ποινή βεβαιότητας(NCP) [23] η οποία εφαρμόζεται κυρίως στη γενίκευση αριθμητικών τιμών. Η συνάρτηση αυτή μετρά το βάθος της γενίκευσης και όσο μικρότερη είναι η τιμή που θα λάβει, τόσο μικρότερη είναι η απώλεια της πληροφορίας. Επίσης υπάρχουν και οι R-U χάρτες (Risk-Utility Maps) [26] οι οποίοι αποτελούν δισδιάστατα γραφήματα μεταξύ του ρίσκου αποκάλυψης και χρήσιμης πληροφορίας παρουσιάζοντας μία σύγκριση μεταξύ διαφορετικών προσεγγίσεων ανωνυμοποίησης.

2.5.7 Μοντέλα Ανωνυμοποίησης

Τα βασικότερα μοντέλα ανωνυμοποίησης είναι τέσσερα και εφαρμόζονται ανάλογα στα γνωρίσματα παρέχοντας ένα επίπεδο προστασίας της ιδιωτικότητας σε ένα σύνολο δεδομένων.

Η k-ανωνυμία, η l-διαφορετικότητα, η t-εγγύτητα και η διαφορική ιδιωτικότητα (differential privacy). Υπάρχουν και άλλα μοντέλα πέραν των βασικών, ωστόσο κατά κύριο λόγο αποτελούν βελτιωμένες εκδοχές τους.

1. Αρχικά έχουμε τη k-ανωνυμία η οποία στοχεύει στο να εξασφαλιστεί ένα ελάχιστο επίπεδο ανωνυμίας στις ψευδο-αναγνωριστικές τιμές ενός συνόλου δεδομένων, βάσης δεδομένων ή πίνακα [27]. Σκοπός είναι για κάθε εγγραφή στο σύνολο αυτό να είναι «αξεχώριστη», με βάση τις τιμές των ψευδο-αναγνωριστικών, από τουλάχιστον k-1 άλλες εγγραφές προστατεύοντας έτσι την ταυτότητα του υποκειμένου καθιστώντας δύσκολη ή αδύνατη τη σύνδεση προσωπικών πληροφοριών με αυτό. Ουσιαστικά απαιτείται η κάθε εγγραφή να έχει όμοιες τιμές στα ψευδο-αναγνωριστικά της με τουλάχιστον k - 1 άλλες εγγραφές, όπου k είναι μια προκαθορισμένη τιμή κατωφλιού.

   Για την επίτευξη της k-ανωνυμίας, τα χαρακτηριστικά των δεδομένων που θα μπορούσαν έμμεσα να ταυτοποιήσουν ένα άτομο (quasi-identifiers) τροποποιούνται ή γενικεύονται έτσι ώστε κάθε εγγραφή να έχει τον ίδιο συνδυασμό ψευδο-αναγνωριστικών με τουλάχιστον k-1 άλλες εγγραφές. Κατά τη γενίκευση ένα χαρακτηριστικό παράδειγμα είναι η ηλικία όπου για ένα πρόσωπο με ηλικία 24 χρονών μπορούμε να την γενικεύσουμε σε ένα εύρος ηλικιών από (20-25).

   Το μοντέλο αυτό βοηθά στην προστασία από επιθέσεις αποκάλυψης της ταυτότητας, όπως αναφέραμε παραπάνω. Ωστόσο, δεν προστατεύει το υποκείμενο από όλους τους κινδύνους ως προς την ιδιωτικότητα, καθώς ο επιτιθέμενος είναι ακόμα σε θέση να συμπεράνει ευαίσθητες πληροφορίες από το σύνολο δεδομένων με βάση τα πρότυπα ή τις κατανομές των τιμών των δεδομένων.

   Το μοντέλο αυτό χρησιμοποιείται ευρέως στην ανταλλαγή και στην ανάλυση δεδομένων, όπου θέλουμε να διατηρήσουμε την ιδιωτικότητα. Ειδικότερα σε περιπτώσεις όπως η ανάλυση δεδομένων που αφορούν την υγειονομική περίθαλψη, όπου συλλέγονται συχνά ιδιαίτερα ευαίσθητες προσωπικές πληροφορίες.

2. Καθώς η k-ανωνυμία παρουσιάζει αρκετές αδυναμίες από μόνη της, η l-διαφορετικότητα έρχεται για να βελτιώσει κάποια από αυτά τα κενά προσφέροντας ακόμη μεγαλύτερη ασφάλεια στα ανωνυμοποιημένα δεδομένα από τη ταυτοποίηση κάποιας οντότητας μέσα σε αυτά [27]. Τροποποιώντας τα δεδομένα ώστε να έχουν τον ίδιο συνδυασμό των ψευδο-

αναγνωριστικών κατά k φορές δεν αποτελεί εγγύηση ότι δεν θα αποκαλυφθεί κάποια ταυτότητα. Σε αρκετές περιπτώσεις τα ευαίσθητα πεδία όπως κάποια ασθένεια μπορεί να είναι τα ίδια σε όλες τις k ισοδύναμες τιμές του ψευδο-αναγνωριστικού, οπότε γνωρίζοντας ότι κάποιο πρόσωπο βρίσκεται σε αυτόν τον πίνακα, όπως εγγραφές κάποιας κλινικής, ακόμα και να μη γνωρίζουν ποιο ακριβώς είναι το πρόσωπο αυτό μπορούμε γνωρίζοντας την ομάδα που ανήκει να διαπιστώσουμε ότι πάσχει από κάποια συγκεκριμένη ασθένεια. Έτσι η l-διαφορετικότητα φροντίζει ώστε σε κάθε κλάση k ισοδυναμίας ή ομάδας των ψευδο-αναγνωριστικών, να περιέχει τουλάχιστον l διαφορετικές τιμές του εμπιστευτικού/ευαίσθητου πεδίου, όπου l είναι η ελάχιστη απαίτηση ποικιλομορφίας των ευαίσθητων χαρακτηριστικών μέσα στην ομάδα. Με αυτό το τρόπο στοχεύει στο να αυξήσει το επίπεδο προστασίας έναντι της αποκάλυψης γνωρίσματος όπως είδαμε και παραπάνω δυσκολεύοντας έτσι τον επιτιθέμενο στο να βγάλει συμπεράσματα ή να εντοπίσει κάποια οντότητα μέσω των ευαίσθητων χαρακτηριστικών. Ωστόσο και σε αυτή τη περίπτωση μπορούν να εφαρμοστούν επιθέσεις εξαγωγής συμπεράσματος μέσω πρόσθετων πληροφοριών αποτελώντας κινδύνους.

3. Η t-εγγύτητα αποτελεί ένα ακόμα μοντέλο απορρήτου που εστιάζει στα ευαίσθητα πεδία και αποσκοπεί στην προστασία από την αποκάλυψη χαρακτηριστικών σε ανώνυμα σύνολα δεδομένων. Αντιμετωπίζει κάποιους από τους περιορισμούς που έχουν να δύο προηγούμενα μοντέλα διασφαλίζοντας ότι η κατανομή των ευαίσθητων γνωρισμάτων σε κάθε κλάση ισοδυναμίας των ανώνυμων εγγραφών, όπου ως κλάση ισοδυναμίας ορίζεται ένα σύνολο εγγραφών με όμοιες τιμές στα ψευδο-αναγνωριστικά (και, άρα, αξεχώριστες μεταξύ τους). είναι αρκετά κοντά σε σχέση με την αντίστοιχη κατανομή στο υπόλοιπο σύνολο δεδομένων [27]. Με το τρόπο αυτό μπορεί να παρέχει εγγυήσεις ότι ένας επιτιθέμενος δεν μπορεί να συμπεράνει με ακρίβεια και να αποδώσει τη τιμή ενός ευαίσθητου χαρακτηριστικού σε κάποια οντότητα συγκρίνοντας την κατανομή του χαρακτηριστικού αυτού στη ανωνυμοποιημένη κλάση. Οπότε ικανοποιώντας την t-εγγύτητα έχουμε ότι η απόσταση μεταξύ της κατανομής του ευαίσθητου γνωρίσματος σε μία κλάση ισοδυναμίας συγκριτικά με τα υπόλοιπες κλάσεις του πίνακα είναι μικρότερη ή ίση με t ,το οποίο αποτελεί κατώφλι.

   Με το τρόπο αυτό μπορούμε να μετριάσουμε τον κίνδυνο αποκάλυψης χαρακτηριστικών. Παρόλο που ο επιτιθέμενος ίσως να μπορεί να εντοπίσει ότι η εγγραφή μιας οντότητας βρίσκεται σε μία συγκεκριμένη ομάδα, δεν μπορεί να συμπεράνει με ακρίβεια το

ευαίσθητο πεδίο αυτής της οντότητας λόγω της ομοιότητας της κατανομής στην ομάδα και σε όλο το σύνολο των δεδομένων.

Το μοντέλο αυτό χρησιμοποιείται συνδυαστικά με τα δύο παραπάνω μοντέλα απορρήτου, παρέχοντας έτσι ισχυρότερη προστασία της ιδιωτικότητας και του απορρήτου σε ανώνυμα σύνολα δεδομένων.

2.6 Σύνοψη

Όπως είδαμε τα δεδομένα προσωπικού χαρακτήρα χρήζουν ιδιαίτερης προσοχής και διαφύλαξης από δυνητικούς κινδύνους που συνδέονται με τη κακή χρήση τους. Στο ψηφιακό κόσμο η ανάγκη για ισχυρά μέτρα προστασίας των δεδομένων αυτών και κατ' επέκταση της προστασίας της ιδιωτικότητας φυσικών προσώπων κρίνεται επιτακτική. Ο GDPR έχει ορίσει σαφείς κανονισμούς κατά το τρόπο χειρισμού τους και επεξεργασίας τους στην έρευνα, προστατεύοντας έτσι θεμελιώδη ανθρώπινα δικαιώματα και εστιάζει ιδιαίτερα σε συγκεκριμένες κατηγορίες τέτοιων δεδομένων όπως τα ευαίσθητα προσωπικά δεδομένα. Έτσι γίνεται κατανοητή η συμμόρφωση με τις κατευθυντήριες γραμμές του νόμου για τη διασφάλιση δεοντολογικών και νόμιμων ερευνητικών πρακτικών. Μέσα σε αυτές τις πρακτικές ανήκει και η ψευδωνυμοποίηση, η οποία εστιάζει σε διάφορες μεθόδους και στρατηγικές απόκρυψης των υποκειμένων των δεδομένων, αντικαθιστώντας τα αναγνωριστικά των υποκειμένων με ψευδώνυμα σε μία προσπάθεια να προστατέψει τη πραγματική τους ταυτότητα. Ωστόσο τα δεδομένα αυτά είναι έτσι διαμορφωμένα που η τεχνική αυτή δείχνει ελλιπής, απαιτώντας επιπλέον επεξεργασία των δεδομένων αυτών θέτοντάς τα πλέον ως ανώνυμα. Η ανωνυμοποίηση τους λοιπόν αποτελεί τεχνική η οποία προσπαθεί να «σπάσει» εντελώς τους δεσμούς μεταξύ των αναγνωριστικών των υποκειμένων και των πρόσθετων πληροφοριών που τα χαρακτηρίζουν, διατηρώντας παράλληλα τη χρηστικότητα τους, δίνοντας έμφαση στην ισορροπία μεταξύ της ιδιωτικότητας και της χρησιμότητας των δεδομένων σε ερευνητικά περιβάλλοντα.

Σε κάθε περίπτωση, η επιλογή της κατάλληλης τεχνικής για ψευδωνυμοποίηση ή/και ανωνυμοποίηση, προκειμένου να πληρούνται οι επιταγές του GDPR, δεν είναι πάντα προφανής: τουναντίον, κατά κανόνα συνιστά μία ιδιαίτερη πρόκληση για τον κάθε υπεύθυνο επεξεργασίας.

Κεφάλαιο 3
Βιβλιογραφική Ανασκόπηση

Το κεφάλαιο αυτό αφορά τη βιβλιογραφική ανασκόπηση της παρούσας Διατριβής αποτελώντας έτσι συνέχεια του θεωρητικού υποβάθρου που εξετάσαμε έως τώρα. Αποσκοπεί στη βαθύτερη πλαισίωση του έργου που εξετάζουμε δίνοντας μεγαλύτερη έμφαση στα δεδομένα της υγείας. Στο σημείο αυτό θα εξεταστούν διαφορετικές προσεγγίσεις που έχουν υλοποιηθεί έως τώρα τόσο στην ψευδωνυμοποίηση των δεδομένων υγείας, όσο και στην ανωνυμοποίηση, μελετώντας παλαιότερες έρευνες, καθώς και κάποιες προηγμένες τεχνικές ψευδωνυμοποίησης. Επίσης θα εμβαθύνουμε περισσότερο στη πρόταση της Ευρωπαϊκής Επιτροπής για το κανονισμό δημιουργίας του Ευρωπαϊκού Χώρου Δεδομένων για την Υγεία (EHDS), τις ανησυχίες που έχει εκφράσει για τη πρόταση αυτή η Ευρωπαϊκή Επιτροπή Προστασίας Δεδομένων ως προς τη συμμόρφωση με τους κανονισμούς το GDPR που αναφέραμε στο προηγούμενο κεφάλαιο, καθώς και άλλους κανονισμούς περί της προστασίας των δεδομένων και πληροφοριών υγείας όπως ο Νόμος περί φορητότητας και λογοδοσίας για την ασφάλιση της υγείας (HIPAA) σε μία προσπάθεια κατανόησης των κατευθυντήριων γραμμών που πρέπει να ακολουθούνται με συνέπεια από τους εμπλεκόμενους φορείς.

3.1 Εφαρμογές Της Ψευδωνυμοποίησης

Μέσα από μελέτες και αναφορές του ευρωπαϊκού οργανισμού ENISA, καθώς και προηγούμενες έρευνες που έχουν υλοποιηθεί για τη προστασία των δεδομένων υγείας, μπορούμε να εξετάσουμε αρκετές περιπτώσεις εφαρμογής της ψευδωνυμοποίησης σε προσωπικά δεδομένα, οι οποίες μπορούν να εφαρμοστούν σε πλήθος εφαρμογών, συμπεριλαμβανομένων περιπτώσεων που αφορούν ευαίσθητα προσωπικά δεδομένα υγείας [28]. Περιλαμβάνονται επίσης διαφορετικές προσεγγίσεις βασισμένες σε κάποιες προηγμένες τεχνικές εφαρμογής της ψευδωνυμοποίησης στοχεύοντας στον ασφαλέστερο διαμοιρασμό δεδομένων υγείας μεταξύ φορέων της υγειονομικής περίθαλψης ή και παροχής τους σε Ερευνητικά Κέντρα για ανάγκες έρευνας διατηρώντας τη χρησιμότητά τους. Παράλληλα προστατεύεται η ιδιωτικότητα των υποκειμένων των δεδομένων αυτών μη αποκαλύπτοντας περισσότερες πληροφορίες από τις ελάχιστα αναγκαίες.

3.1.1 Προηγμένες Τεχνικές Ψευδωνυμοποίησης

Στην υποενότητα αυτή θα περιγράψουμε ορισμένες από τις προηγμένες τεχνικές ψευδωνυμοποίησης που θα μπορούσαν να εφαρμοστούν σε πραγματικά σενάρια. Οι τεχνικές αυτές προσφέρουν πολύ μεγαλύτερη ασφάλεια συγκριτικά με τις «κλασικές» προσεγγίσεις

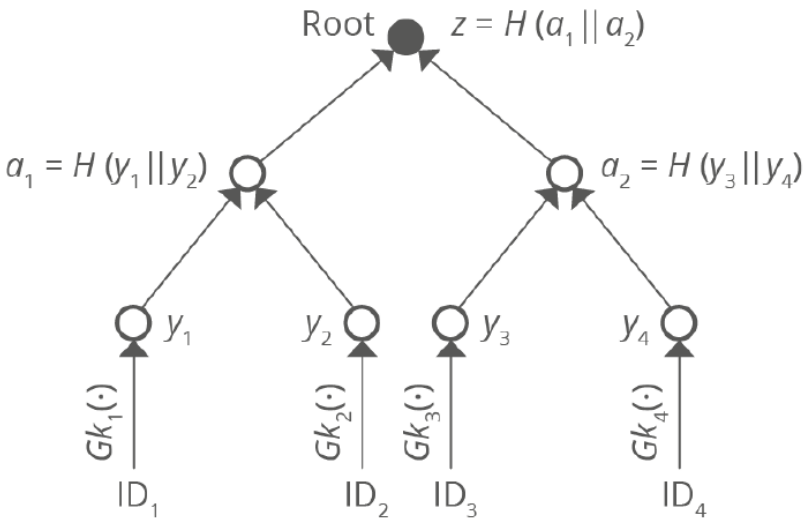
προστατεύοντας εγγυημένα τα δεδομένα από την άμεση αποκάλυψη τους και τη ταυτοποίηση φυσικών προσώπων [28].

1. Ασύμμετρη κρυπτογράφηση: Η ιδέα πίσω από την τεχνική αυτή εμπεριέχει δύο οντότητες, δημόσιου και ιδιωτικού κλειδιού, που συμμετέχουν κατά τη ψευδωνυμοποίηση προσφέροντας μεγαλύτερη ασφάλεια. Η πρώτη οντότητα μπορεί να δημιουργεί τα ψευδώνυμα κρυπτογραφώντας με το δημόσιο κλειδί της δεύτερης οντότητας, ενώ η δεύτερη χρειάζεται το ιδιωτικό κλειδί για να επαναπροσδιορίσει τα αναγνωριστικά. Σε αντίθεση με τη συμμετρική κρυπτογράφηση όπου χρησιμοποιείται το ίδιο κλειδί για κρυπτογράφηση και αποκρυπτογράφηση, στην ασύμμετρη οι δύο οντότητες δεν χρειάζεται να μοιράζονται το ίδιο μυστικό κλειδί. Παράδειγμα αποτελεί η περίπτωση του υπεύθυνου επεξεργασίας και του εκτελούντος την επεξεργασία των δεδομένων. Ο πρώτος μπορεί να διαθέσει το δημόσιο κλειδί του ώστε ο δεύτερος να μπορεί με αυτό να ψευδωνυμοποιήσει τα δεδομένα. Το μυστικό κλειδί δεν μοιράζεται, καθώς το έχει μόνο ο υπεύθυνος επεξεργασίας και είναι ο μόνος που μπορεί να επαναπροσδιορίσει τα αναγνωριστικά, χωρίς να υπάρχει άλλος τρόπος επαλήθευσης τους. Επίσης το δημόσιο κλειδί για μεγαλύτερη ασφάλεια και επαλήθευσή του μπορεί να εκδοθεί και από τρίτες έμπιστες οντότητες (TTP) σε έναν ή και περισσότερους υπευθύνους επεξεργασίας των δεδομένων.

   Κρίσιμος παράγοντας βέβαια για μία τέτοια υλοποίηση αποτελεί η χρήση πιθανοτικής κρυπτογράφησης: διαφορετικά, αν θεωρηθεί – όπως ισχύει στην κλασική κρυπτογράφηση δημοσίου κλειδιού – ότι το δημόσιο κλειδί μίας οντότητας δεν προστατεύεται και είναι εύκολα προσβάσιμο από όλους, τότε θα μπορούσε κάποιος επιτιθέμενος να δημιουργήσει, για όλα τα πιθανά αναγνωριστικά, όλα τα αντίστοιχα ψευδώνυμα και, ακολούθως να παρατηρήσει το ψευδώνυμο που δημιουργεί η πρώτη οντότητα, έχοντας τη δυνατότητα να το αντιστοιχίσει, βάσει των προ-υπολογισμένων ψευδωνύμων, το αντίστοιχο αναγνωριστικό. Με την πιθανοτική κρυπτογράφηση όμως, κάθε ψευδώνυμο που θα παράγεται για τον ίδιο χρήστη, και με το ίδιο κλειδί, θα είναι κάθε φορά διαφορετικό – άρα, προσφέρεται ιδίως για μη ντετερμινιστική ψευδωνυμοποίηση.

2. Ψευδωνυμοποίηση βασισμένη σε πολλαπλά αναγνωριστικά: Στη περίπτωση αυτή αντί της μετατροπής ενός αναγνωριστικού γνωρίσματος σε ψευδώνυμο, ένα προς ένα, θα μπορούσαμε να μετατρέψουμε το συνδυασμό μίας σειράς αναγνωριστικών σε ένα ψευδώνυμο. Τα γνωρίσματα θα μπορούσαν να αποτελούν όλα τον ίδιο τύπο ή συνδυασμό

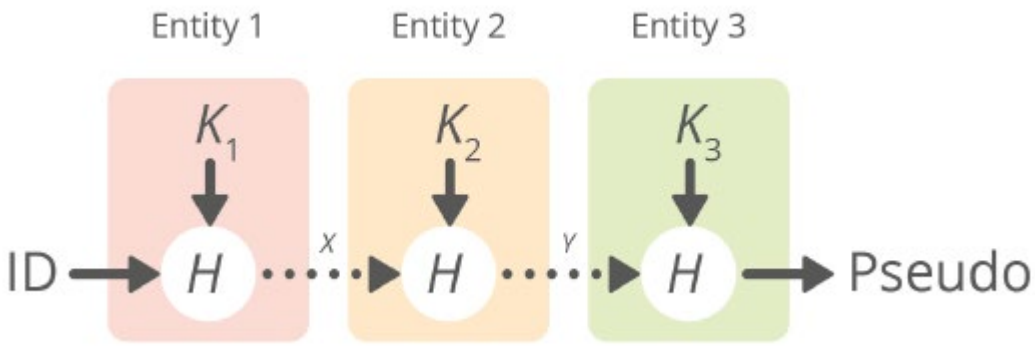
διαφορετικών τύπων γνωρισμάτων. Εκμεταλλευόμενοι τη τακτική αυτή μπορούμε να εφαρμόσουμε αρκετές από τις βασικές τεχνικές ψευδωνυμοποίησης όπως τις συναρτήσεις κατακερματισμού ή και τη χρήση κώδικα αυθεντικοποίησης μηνύματος. Πρακτικά φέρνοντας ως παράδειγμα τη χρήση συναρτήσεων κατακερματισμού, θα μπορούσαν να ψευδωνυμοποιηθούν τα γνωρίσματα βασισμένα στο δένδρο Merkle δημιουργώντας μία αλυσίδα κατακερματισμένων τιμών [29]. Αρχικά κατακερματίζονται τα γνωρίσματα ξεχωριστά στο κατώτατο επίπεδο και όσο ανεβαίνουμε επίπεδο κατακερματίζονται ως συνδυασμός των ήδη κατακερματισμένων τιμών που λάβαμε.



Εικόνα 3.1: Δέντρο Merkle [28]

Στο ανώτερο επίπεδο του δένδρου έχουμε τον κατακερματισμό όλων των τιμών ως μία μεμονωμένη ακολουθία.

3. Λειτουργία αλυσίδας: Αποτελεί τεχνική που εκμεταλλεύεται τις συναρτήσεις κατακερματισμού ως μία αλυσίδα. Αρχικά έχουμε μία οντότητα η οποία χρησιμοποιεί συνάρτηση hash με κάποιο κλειδί (k1) μετατρέποντας το αναγνωριστικό σε ακολουθία σταθερού μεγέθους H(1). Στη συνέχεια η ακολουθία αυτή κατακερματίζεται ξανά από δεύτερη οντότητα σε H(2) και η διαδικασία ακολουθείται μία ακόμη φορά από τρίτη οντότητα ώστε να δημιουργηθεί το ψευδώνυμο.



Εικόνα 3.2: Ψευδωνυμοποίηση - λειτουργία αλυσίδας [28]

Όπως βλέπουμε και στην εικόνα 3.2 κάθε επίπεδο κατακερματισμού αποτελείται από διαφορετική οντότητα, όπου υπεισέρχεται ένα ξεχωριστό κλειδί κάθε φορά και έτσι η κάθε μία από αυτές κρατάει και κάποιο μυστικό κατά τη δημιουργία του ψευδωνύμου. Ως εκ τούτου για να επαναπροσδιορίσουμε το αρχικό αναγνωριστικό, οι τρείς αυτές οντότητες θα πρέπει να συνεργαστούν προσφέροντας έτσι αρκετά ισχυρή ασφάλεια επαναπροσδιορισμού του αρχικού αναγνωριστικού γνωρίσματος. Σημαντικό ρόλο στη διαδικασία αυτή παίζει η εμπιστοσύνη των οντοτήτων αυτών μεταξύ τους δίδοντας έτσι τη πραγματική κατακερματισμένη τιμή η μία στην άλλη.

4. Απόδειξη μηδενικής γνώσης: Αποτελεί σενάριο στο οποίο συμμετέχουν δύο οντότητες, η μία που μπορεί να αποδείξει ότι γνωρίζει κάποιο μυστικό (prover) και η δεύτερη που επαληθεύει τη γνώση αυτή χωρίς όμως να γνωρίζει περί τίνος πρόκειται (verifier). Ουσιαστική ο prover έχει στη κατοχή του κάποιο μυστικό χωρίς να αποκαλύψει καμία πληροφορία για το μυστικό αυτό. Από την άλλη ο verifier θα πρέπει να επαληθεύσει, χωρίς καμία γνώση για το μυστικό αυτό, ότι ο prover λέει την αλήθεια. Για να επιτευχθεί αυτό ακολουθείται μία διαδραστική διαδικασία μεταξύ των δύο οντοτήτων. Ο verifier θα πρέπει να ρωτήσει πολλές φορές τον prover κάποιο ερώτημα και καθώς ο prover γνωρίζει όντος το μυστικό θα πρέπει να δίνει πάντα τη σωστή απάντηση. Με το τρόπο αυτό ο verifier επαληθεύει τη γνώση του prover χωρίς να έχει πρόσβαση σε κάποια πληροφορία του μυστικού αυτού [30]. Στα πλαίσια της ψευδωνυμοποίησης, μία οντότητα που συνδέεται με ένα ψευδώνυμο θα πρέπει να μπορεί να αποδείξει ότι είναι κάτοχος του ψευδωνύμου αυτού χωρίς να αποκαλύψει τη ταυτότητά του.

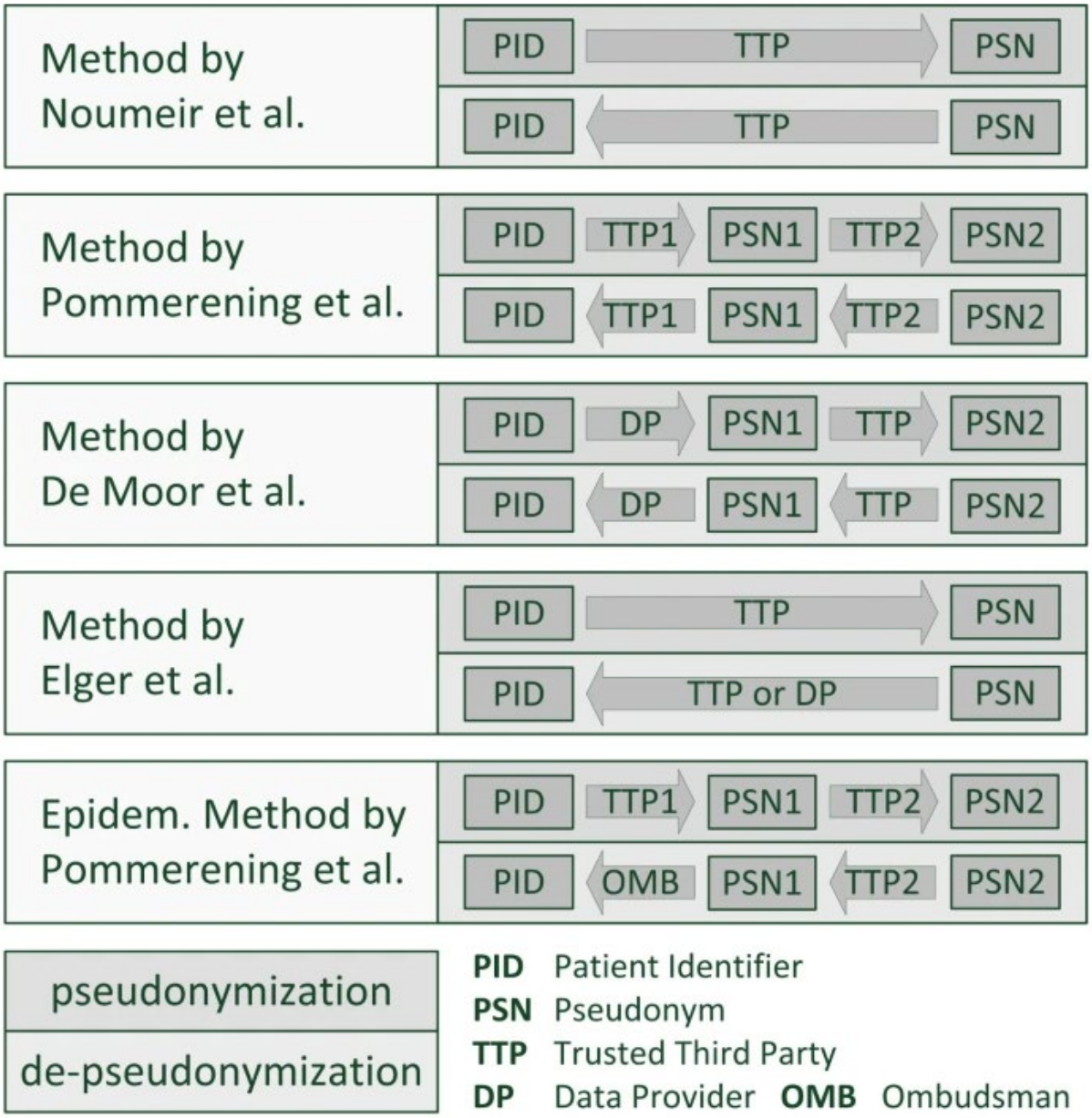
Υπάρχουν και άλλες πρακτικές προηγμένων τεχνικών ψευδωνυμοποίησης, ωστόσο η αναφορά όλων θα ξεπερνούσε το πεδίο εφαρμογής της έρευνάς μας.

3.1.2 Ψευδωνυμοποίηση Στην Υγεία

Αναφορικά με τη προστασία των ευαίσθητων προσωπικών δεδομένων στην υγεία, έχουν πραγματοποιηθεί αρκετές έρευνες τα τελευταία χρόνια, καθώς τα δεδομένα αυτά αποτελούν ένα σημαντικό πεδίο έρευνας, εξάγοντας πολύτιμα συμπεράσματα, καθώς έχουν ωφελήσει σε ένα μεγάλο βαθμό την ιατρική κοινότητα και γενικότερα έχουν συνεισφέρει πολύ σε ολόκληρη την υγειονομική περίθαλψη. Οι τεχνικές της ψευδωνυμοποίησης αποτελούν ένα σημαντικό κομμάτι στην διαφύλαξη της ιδιωτικότητας των υποκειμένων τέτοιων δεδομένων, παίζοντας καθοριστικό ρόλο στον ασφαλή διαμοιρασμό τους.

Έρευνα του 2013 η οποία μελετά το διαμοιρασμό δεδομένων υγείας, περιγράφει τεχνικές με τις οποίες μπορούν να διατεθούν τα δεδομένα αυτά για ερευνητικούς σκοπούς με ασφάλεια μέσω διαδικασιών ψευδωνυμοποίησης [31]. Επισημαίνει το ρίσκο και τα ηθικά και νομικά ζητήματα που μπορούν να προκύψουν κατά την επεξεργασία και διαμοιρασμό τέτοιου είδους δεδομένων σχετικά με την ιδιωτικότητα των ασθενών, αλλά και τα οφέλη που μπορούν να μας δώσουν. Οι ερευνητές, για τις ανάγκες αυτές και αφού έχουν εξετάσει προηγούμενες έρευνες ψευδωνυμοποίησης σε ιατρικά δεδομένα, παρουσιάζοντας τις τεχνικές που χρησιμοποιήθηκαν στις έρευνες αυτές, τονίζοντας τα οφέλη τους, προτείνουν και μία δική τους νέα εκδοχή, συνεισφέροντας σε μία ασφαλή ψευδωνυμοποίηση δεδομένων υγείας.

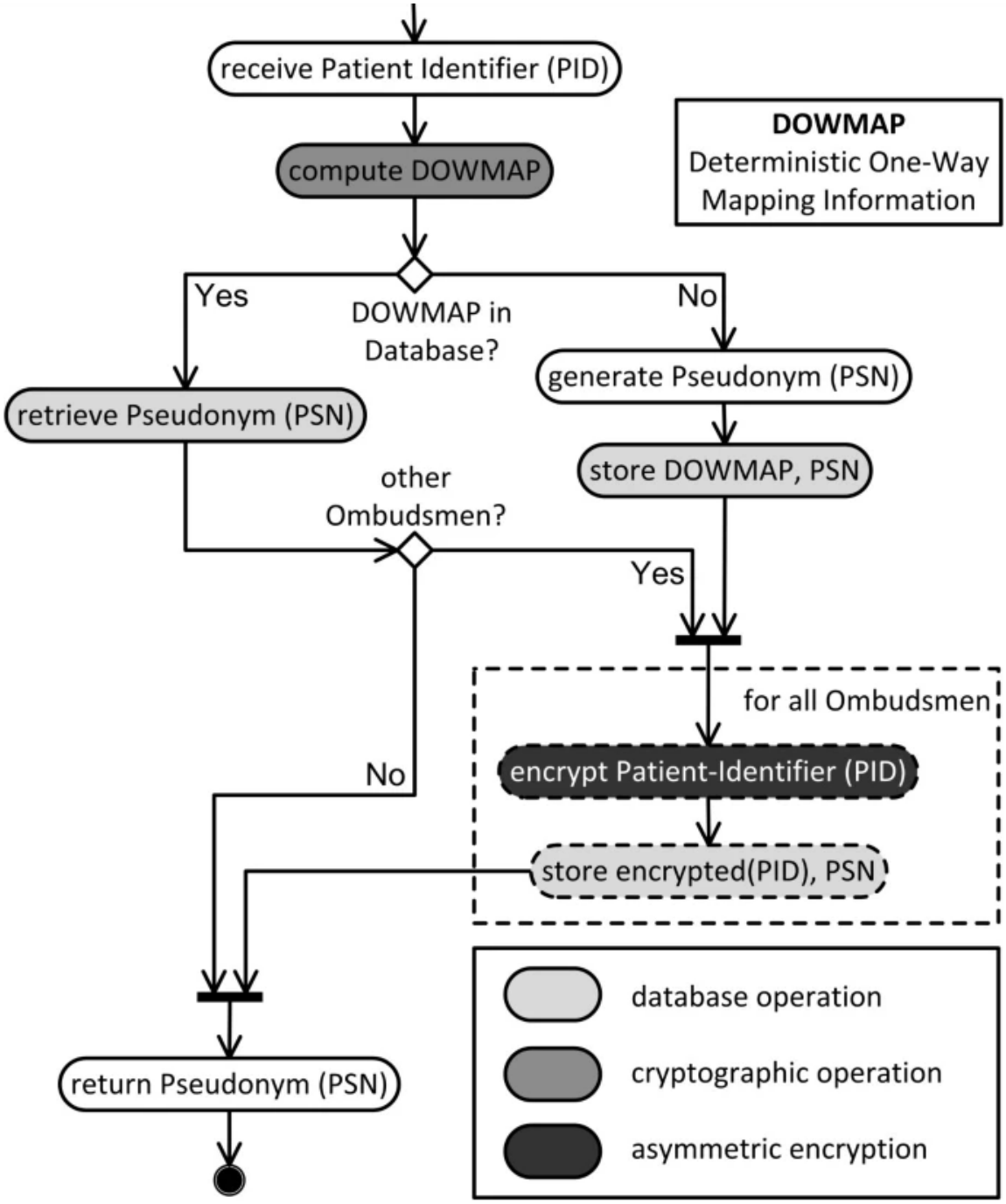
Κατά την έρευνά τους λοιπόν παρουσιάζουν κάποιες τεχνικές, όπου κυρίως χρησιμοποιείται συμμετρική ή και ασύμμετρη κρυπτογράφηση, ενώ σε κάποιες από τις μεθοδολογίες που εξετάζουν, τα αναγνωριστικά ψευδωνυμοποιούνται περισσότερες από μία φορές από διαφορετικές οντότητες, ενώ συμμετέχουν διαφορετικές ομάδες (διαμεσολαβητές, τρίτες έμπιστες οντότητες κλπ.) σε τέτοιες περιπτώσεις όπως φαίνεται στην εικόνα 3.3 παρακάτω.



Εικόνα 3.3: Μεθοδολογίες ψευδωνυμοποίησης δεδομένων υγείας [31]

Έχοντας λοιπόν εξετάσει τις μεθοδολογίες προηγούμενων ερευνών, οι ερευνητές προτείνουν το δικό τους μηχανισμό ψευδωνυμοποίησης των δεδομένων υγείας σε ερευνητικά έργα, όπου προτείνουν ένα διαχωρισμό καθηκόντων από-ψευδωνυμοποίησης, η οποία πραγματοποιείται από ειδικούς διαμεσολαβητές. Ξεκινώντας λοιπόν με τη ψευδωνυμοποίηση, η διαδικασία αρχικά υλοποιείται από μία έμπιστη τρίτη οντότητα η οποία λαμβάνει το αναγνωριστικό του ασθενή και δημιουργεί μία ντετερμινιστική χαρτογράφηση πληροφοριών μίας κατεύθυνσης όπως φαίνεται

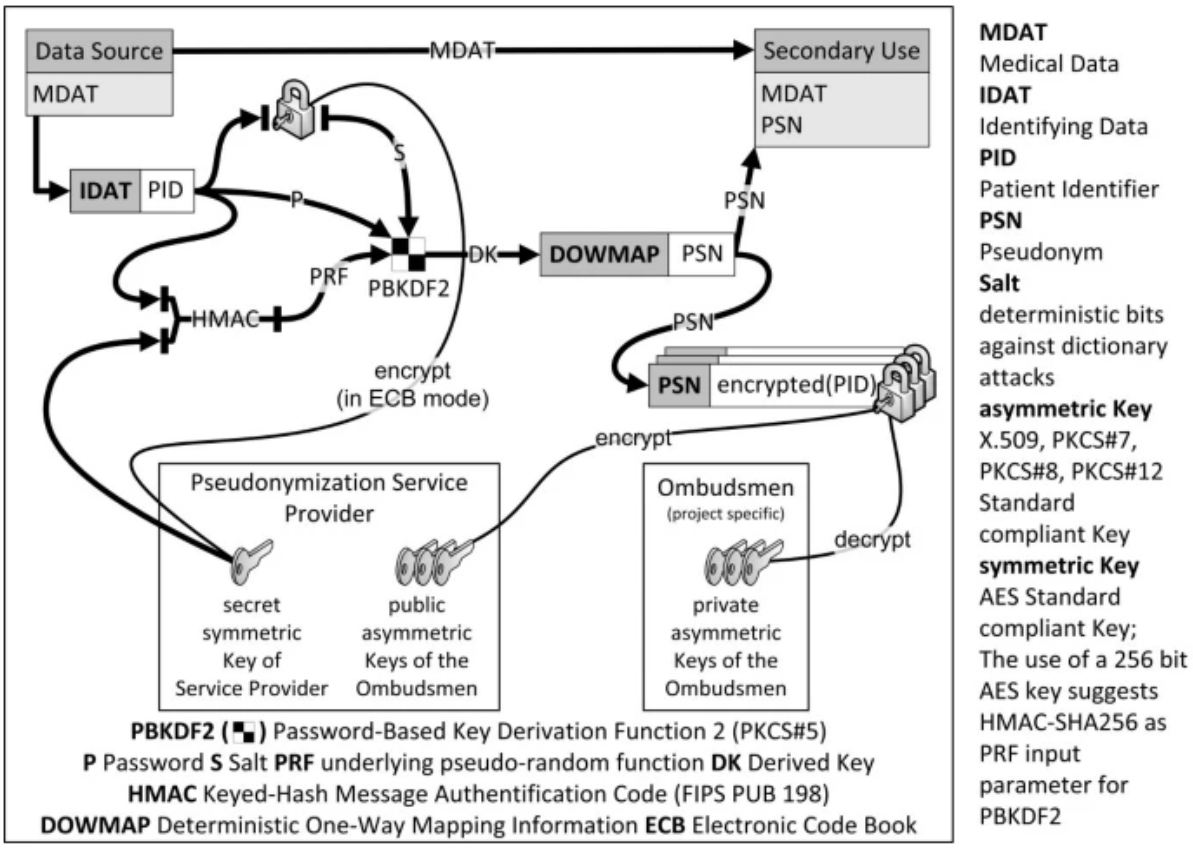
παρακάτω δημιουργώντας έτσι ένα μοναδικό ψευδώνυμο για τον ασθενή σε περίπτωση που δεν το βρει στη βάση, αλλιώς καλεί το ψευδώνυμο που βρήκε.



Εικόνα 3.4: Δημιουργία χαρτογράφησης ψευδωνύμου [31]

Εκεί γεννιέται ένας προβληματισμός που απασχολεί τους ερευνητές, καθώς κατά τη δημιουργία ενός ψευδωνύμου, προκειμένου να επιτευχθεί η ασφάλεια, απαιτείται να χρησιμοποιηθούν αλγόριθμοι τμήματος οι οποίοι δημιουργούν κάποια προβλήματα με βάση τις λειτουργίες τους [32]. Πέραν των αλγορίθμων σε λειτουργία ECB(Electronic Code Book) οι οποίοι δεν προτείνονται

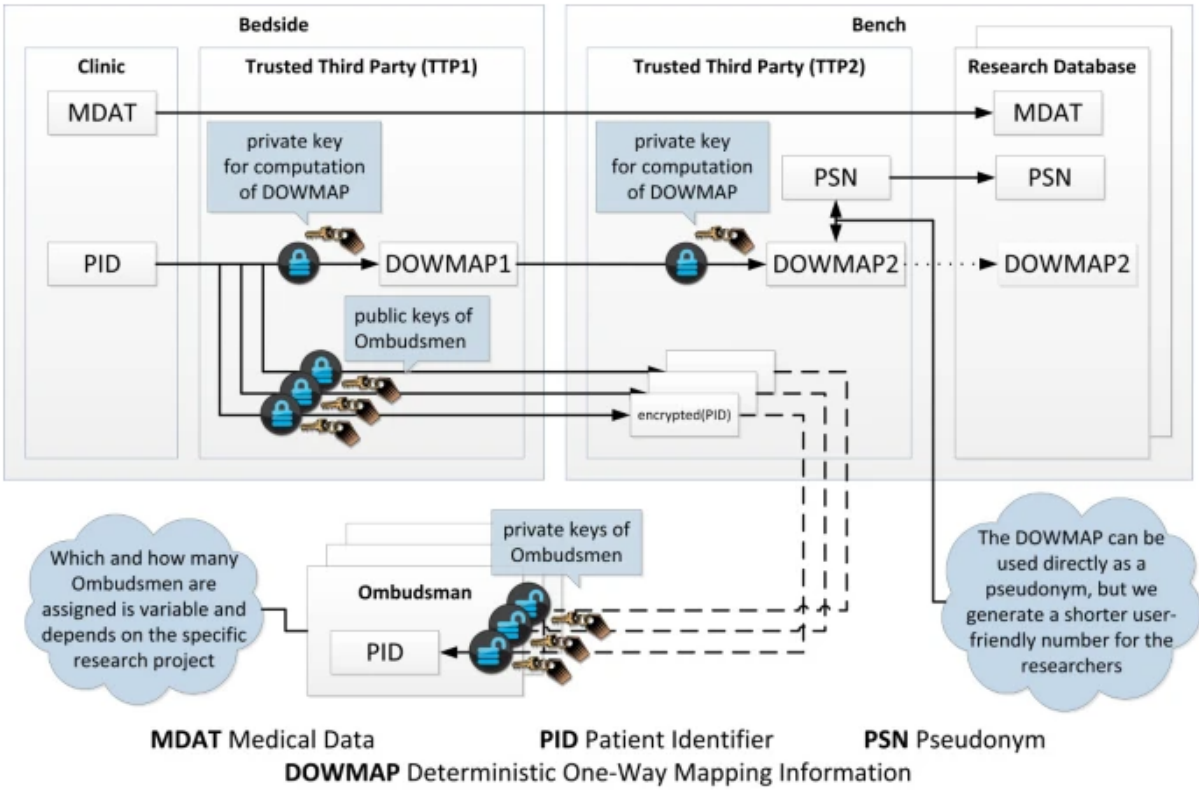
λόγω προβλημάτων ασφαλείας, όλες οι άλλες λειτουργίες δημιουργούν διαφορετικό κρυπτοκείμενο κάθε φορά που κρυπτογραφείται το ίδιο μήνυμα, ακόμα και με το ίδιο κλειδί. Ωστόσο αυτό που ζητούν οι ερευνητές είναι να παράγονται μοναδικά και πάντα ίδια ψευδώνυμα για τον ίδιο ασθενή. Έτσι προκειμένου να το πετύχουν αυτό, παρέχοντας παράλληλα ασφάλεια στη ψευδωνυμοποίηση, αποφασίζουν να χρησιμοποιήσουν μία ειδική κατηγορία του Κώδικα Αυθεντικοποίησης Μηνύματος (που εξετάσαμε σε προηγούμενο κεφάλαιο), τον Keyed-Hash MAC (HMAC). Όπως έχουμε αναφέρει οι MAC συναρτήσεις προσφέρουν μεγάλη ασφάλεια, καθώς πρόκειται για συναρτήσεις μίας κατεύθυνσης, δημιουργούν μοναδικά ψευδώνυμα και πάντα ίδια για την ίδια είσοδο, ενώ καθώς στη διαδικασία υπεισέρχεται και ένα κλειδί, τους παρέχει ακόμα μεγαλύτερη ασφάλεια, αυξάνοντας κατά πολύ το κόστος και τον χρόνο διαφόρων επικείμενων επιθέσεων στα ψευδωνυμοποιημένα δεδομένα.



Εικόνα 3.5: Παραγωγή ψευδωνύμου για το διαμοιρασμό δεδομένων υγείας [31]

Όπως βλέπουμε στην εικόνα 3.5 παραπάνω, για να αποσταλούν τα δεδομένα του ασθενούς, ακολουθείται μία πολύπλοκη διαδικασία ώστε να δημιουργηθεί το ψευδώνυμο του

αναγνωριστικού ασθενούς. Αρχικά δημιουργείται μία γεννήτρια στην οποία εισέρχεται το αναγνωριστικό σε τρείς διαφορετικές μορφές, 'plaintext', κρυπτογραφημένο με συμμετρικό κλειδί και μέσα από τη συνάρτηση HMAC, ώστε να παραχθεί το ψευδώνυμο που θα αποθηκευτεί στο 'DOWMAP'. Με το τρόπο αυτό, όταν τα δεδομένα σταλούν σε ερευνητικό κέντρο θα λάβουν ένα ψευδώνυμο αντικαθιστώντας το αρχικό αναγνωριστικό του ασθενούς. Στη συνέχεια όπως βλέπουμε κρυπτογραφείται το αναγνωριστικό με το δημόσιο ασύμμετρο κλειδί κάποιου διαμεσολαβητή, ο οποίος είναι και αυτός που συνδυαστικά μαζί με το ψευδώνυμο μπορεί να το αποκρυπτογραφήσει με το ιδιωτικό του κλειδί. Ο μηχανισμός που προτείνεται από τους ερευνητές για την παραγωγή του ψευδωνύμου, αλλά και την αποκρυπτογράφηση του αναγνωριστικού από ειδικούς διαμεσολαβητές, διαφαίνεται στην εικόνα 3.6.



Εικόνα 3.6: Μηχανισμός παραγωγής ψευδωνύμων [31]

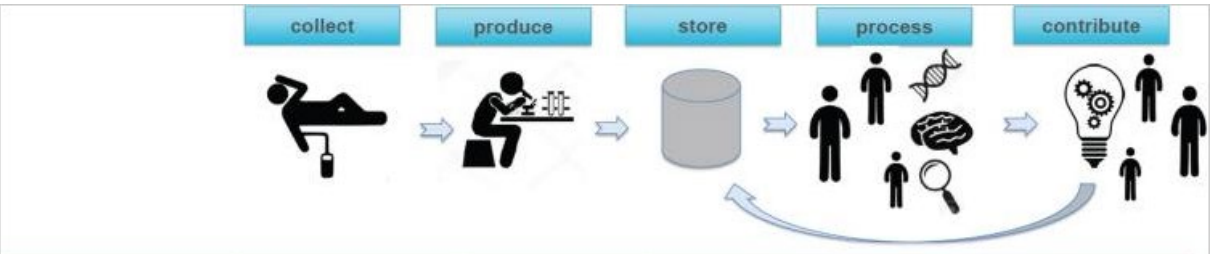
Σε μία άλλη μελέτη που δημοσιεύτηκε το 2021 [33], περιγράφει μία προηγμένη μορφή ψευδωνυμοποίησης που πραγματοποιήθηκε για την ασθένεια Parkinson, με το όνομα 'Personalized Parkinson Project' (PPP). Η μελέτη αυτή πραγματοποιήθηκε στην Ολλανδία, συλλέγοντας διάφορα βιοϊατρικά δεδομένα, ωστόσο τα δεδομένα αυτά συλλέγονται σε παγκόσμια κλίματα ώστε να χρησιμοποιηθούν σε έρευνα. Τα δεδομένα είναι ψευδωνυμοποιημένα, βάσει των κανονισμών που ορίζει ο GDPR, χρησιμοποιώντας τεχνικές

'πολυμορφικής' κρυπτογράφησης για τη ψευδωνυμοποίηση, αποδίδοντας σε κάθε συμμετέχουσα ερευνητική ομάδα τα δικά της «τοπικά» ψευδώνυμα για τα δεδομένα αυτά. Καθώς το σύστημα αφορά, όπως τονίσαμε, παγκόσμια κλίματα, τα δεδομένα αυτά είναι προσβάσιμα και από άλλες ερευνητικές ομάδες μέσω των τοπικών ψευδωνύμων τους.

Συνεχίζοντας οι ερευνητές μελετούν τους ασθενείς από δύο οπτικές, αυτή του ιατρού και αυτή των ιατρικών ερευνητών, όπου στη δεύτερη περίπτωση τα δεδομένα θα πρέπει να εμφανίζονται υπό τη μορφή ψευδωνύμων, αντί των προσωπικών πληροφοριών για τη διαφύλαξη της ιδιωτικότητας, εξετάζοντας, πώς η ψευδωνυμοποίηση μπορεί να εφαρμοστεί σε ένα μεγάλο εύρος δεδομένων υγείας, ώστε να μπορεί να προστατέψει τα υποκείμενα (ασθενείς) μειώνοντας το ρίσκο ταυτοποίησης κάποιου φυσικού προσώπου.

Περνώντας στο στάδιο της ψευδωνυμοποίησης, προκειμένου να μπορέσουν να διαμοιράσουν τα δεδομένα υγείας, οι ερευνητές χρησιμοποιούν ένα σύστημα πολυμορφικής κρυπτογράφησης και ψευδωνυμοποίησης (PEP system), μελετώντας τεχνικές ψευδωνυμοποίησης και προτείνοντας μία βελτιωμένη εκδοχή αυτών, βασισμένη σε αλγορίθμους ασύμμετρης κρυπτογράφησης. Το σύστημα αυτό δημιουργεί πολυμορφικά ψευδώνυμα για να μελετήσει τους συμμετέχοντες. Ο όρος «πολυμορφικά» χρησιμοποιείται για να δείξει ότι, με χρήση προηγμένων τεχνικών κρυπτογράφησης, μπορούν να παραχθούν πολλά διαφορετικά ψευδώνυμα ανά ασθενή, δηλαδή ξεχωριστό ψευδώνυμο ανά αποδέκτη δεδομένων (έτσι ώστε διαφορετικοί αποδέκτες να μην μπορούν να συνδυάσουν τα ψευδωνυμοποιημένα τους δεδομένα), και αυτό μπορεί να γίνει με βάση μία κρυπτογραφημένη μορφή ενός αρχικού ψευδωνύμου, χωρίς να χρειάζεται το κλειδί αποκρυπτογράφησής του. Τα ψευδώνυμα αυτά αποτελούνται από ένα μεγάλο αριθμό, 65 χαρακτήρων σε δεκαεξαδική μορφή. Οι αριθμοί αυτοί έχουν μία εσωτερική κρυπτογραφική δομή, έτσι ώστε να μπορούν να μετασχηματιστούν σε έναν άλλο αριθμό, ο οποίος θα αποτελεί το τοπικό ψευδώνυμο, που θα χρησιμοποιηθεί από κάποιον ερευνητή συγκεκριμένης ερευνητικής ομάδας τους συστήματος PEP. Το σύστημα λοιπόν είναι έτσι διαμορφωμένο, χρησιμοποιώντας διάφορα ανεξάρτητα μέρη για τη παραγωγή του τελικού ψευδωνύμου, όπου κάθε μέρος του συστήματος αυτού δεν γνωρίζει, κατά τη διαδικασία, τη πραγματική ταυτότητα του συμμετέχοντα για τον οποίο δημιουργεί το ψευδώνυμο, αξιοποιώντας ένα μηχανισμό «τυφλής» μετάφρασης. Με αυτό το τρόπο, κάθε ερευνητική ομάδα που συμμετέχει στην έρευνα αυτή (PPP) λαμβάνει το δικό της τοπικό ψευδώνυμο με αποτέλεσμα, ακόμα και να αποκαλυφθεί κάποιο από τα ψευδώνυμα αυτά, οι επιπτώσεις θα είναι σε τοπική κλίμακα, χωρίς να επηρεάσει ολόκληρο το σύστημα.

Στη συνέχεια οι ερευνητές περιγράφουν το σύστημα PEP μέσα από πέντε φάσεις όπως φαίνεται στην εικόνα 3.7.



Who	Assessor	Producer	Repository	Researcher
Identity information	Identifiable	Pseudonymised	Encrypted/ Pseudonymised	Pseudonymised
Data	Partly Clear text	Clear Text	Encrypted	Clear text
Linkability	Linkable to person	Unlinkable	Unlinkable	Linkable to pseudonym

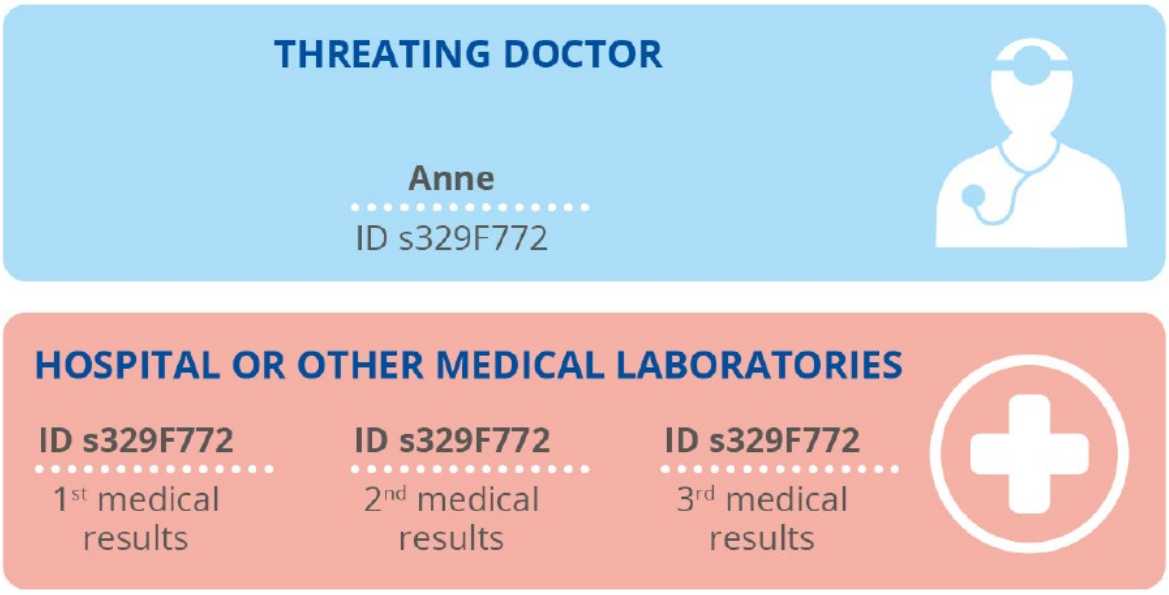
Εικόνα 3.7: Σύστημα κρυπτογράφησης PEP [33]

Η πρώτη φάση της συλλογής περιλαμβάνει επισκέψεις καθώς και συνεχή παρακολούθηση των ασθενών μέσω συσκευών παρακολούθησης της υγείας του ασθενούς. Κατά τη πρώτη επίσκεψη ό συμμετέχων λαμβάνει τη συσκευή αυτή η οποία είναι μόνιμα ενεργοποιημένη ώστε να συλλέγει τα δεδομένα που χρειάζονται για την έρευνα. Στη συνέχεια τα δεδομένα στέλνονται σε μια μορφή συνδυασμού του σειριακού αριθμού της συσκευής και του τοπικού ψευδωνύμου που δημιουργήθηκε. Έτσι οι ερευνητές δεν λαμβάνουν κάποια προσωπική πληροφορία του συμμετέχοντος. Συνεχίζοντας, ακολουθεί το στάδιο της παραγωγής, όπου τα δεδομένα διαμορφώνονται με τέτοιο τρόπο ώστε να μπορέσουν να «ανέβουν» στο σύστημα PEP, ωστόσο πριν σταλούν και αποθηκευτούν στο σύστημα θα πρέπει να καταργηθεί οποιοδήποτε αναγνωριστικό που μπορεί να συσχετίσει τα ψευδώνυμα με κάποιο πρόσωπο. Κατά το στάδιο της αποθήκευσης τα δεδομένα είναι ήδη κρυπτογραφημένα και διαμορφωμένα κατάλληλα, ενώ τα κλειδιά της κρυπτογράφησης κρατούνται εντός του συστήματος PEP. Από την άλλη κάθε ομάδα έχει τα τοπικά ψευδώνυμα της καθώς και το κλειδί της αποκρυπτογράφησης για αυτά, ώστε να μπορεί να τα επαναπροσδιορίσει όταν χρειαστεί, έχοντας πρόσβαση σε αυτά σε τοπική ψευδωνυμοποιημένη μορφή. Η τέταρτη φάση αφορά την επεξεργασία των ψευδωνυμοποιημένων δεδομένων όπου πραγματοποιείται σε ένα ασφαλές περιβάλλον. Τέλος υπάρχει και μία πρόσθετη φάση όπου μία ερευνητική ομάδα μπορεί να επιστρέψει κάποια αποτελέσματα στο σύστημα PEP, χρησιμοποιώντας τα δικά της τοπικά ψευδώνυμα για τη μεταφόρτωση. Έτσι, αφού τα αποτελέσματα μεταφορτωθούν στο σύστημα, δίνεται η

δυνατότητα σε άλλες συμμετέχουσες ερευνητικές ομάδες να έχουν πρόσβαση σε αυτά με τα δικά τους τοπικά ψευδώνυμα.

3.1.3 Άλλες Περιπτώσεις Εφαρμογής Της Ψευδωνυμοποίησης

Μελέτη που αναλύεται σε πρόσφατη αναφορά του ENISA [17] περιγράφει ένα σενάριο ανταλλαγής ιατρικών δεδομένων μεταξύ οργανισμών ή και τμημάτων εντός του Νοσοκομείου για διαγνωστικούς και θεραπευτικούς σκοπούς. Στη υπόθεση αυτή αναφέρεται σε έναν ασθενή ο οποίος εξετάστηκε σε εργαστήριο και θέλει να σταλούν τα αποτελέσματα στο ιατρό του. Κατά τη πρώτη επίσκεψη στα διαγνωστικά εργαστήρια, ζητήθηκαν στοιχεία του ασθενούς, της κατάστασής του, ιατρικό ιστορικό και δημογραφικά στοιχεία και του εκδόθηκε ένα μοναδικό ID ασθενούς. Με το ID αυτό το οποίο προκύπτει από έναν απλό μετρητή, μπορούν τα εργαστήρια να διαχωρίσουν τα προσωπικά στοιχεία του ασθενούς από τα αποτελέσματα των εξετάσεων χρησιμοποιώντας ένα απλό ψευδώνυμο.



Εικόνα 3.8: Ψευδωνυμοποίηση απλής αντικατάστασης

Έχοντας ολοκληρώσει τις εξετάσεις του ο ασθενής, τα εργαστήρια συσχετίζουν τα αποτελέσματα με το ID ασθενούς και όχι με τα προσωπικά του στοιχεία. Έτσι όταν ο ασθενής ζητήσει αντίγραφο των αποτελεσμάτων του, τα εργαστήρια θα τα αναζητήσουν με το ID που έχει αποδοθεί και στη συνέχεια θα το συσχετίσουν με τα προσωπικά στοιχεία. Στη περίπτωση που τα αποτελέσματα σταλούν απευθείας στον ιατρό του ασθενούς, τα εργαστήρια μπορούν να ενημερώσουν το ιατρό

με τα δεδομένα του συγκεκριμένου ID που ανήκει στον ασθενή τη πρώτη φορά, ενώ από πλευράς του ο ιατρός θα πρέπει να γνωρίζει από εκεί και ύστερα ότι το ID αυτό ανήκει στο συγκεκριμένο ασθενή του.

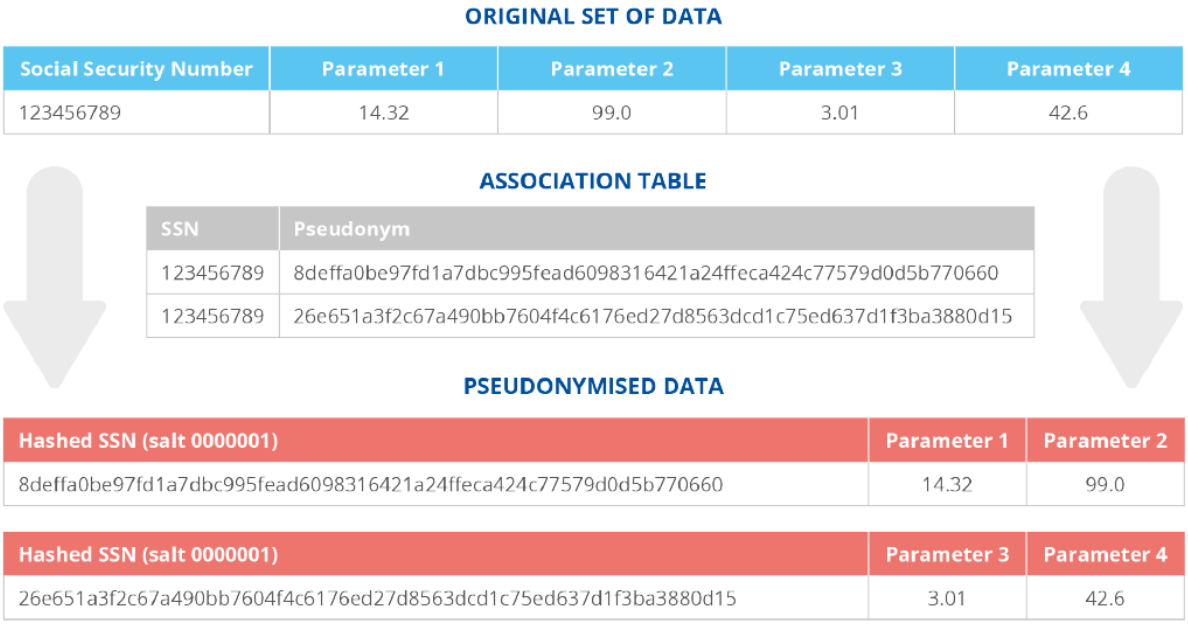
Μία άλλη περίπτωση [17] περιγράφει ένα σενάριο κλινικών δοκιμών που μελετούν νέες ιατρικές παρεμβάσεις και θεραπείες, αξιολογώντας τα αποτελέσματα τους καθώς και ενδεχόμενα παρενεργειών, ώστε να μπορέσουν να λάβουν τις απαραίτητες εγκρίσεις. Έτσι αποφασίζουν να ακολουθήσουν ένα τυπικό σενάριο της διπλής τυφλής μελέτης, χωρίζοντας άτομα με παρόμοια χαρακτηριστικά (όπως ασθένειες, φύλο, ηλικιακές ομάδες κλπ.) σε δύο υπο-ομάδες, όπου η μία ομάδα θα λάβει την υπό δοκιμή φαρμακευτική αγωγή, ενώ η άλλη ομάδα θα λάβει ένα εικονικό φάρμακο «placebo». Με το τρόπο αυτό, χωρίς να γνωρίζουν ούτε οι ερευνητές αλλά ούτε και οι ασθενείς σε ποια ομάδα ανήκουν, μπορούν να εξαχθούν πολύτιμα συμπεράσματα για την απόδοση της φαρμακευτικής αγωγής ή της ιατρικής παρέμβασης καθώς και αποτελέσματα που συσχετίζονται με ηλικιακές ομάδες, φύλο, επάγγελμα και άλλες μεταβλητές.

Σε τέτοιου είδους έρευνες η ταυτότητα των συμμετεχόντων δεν ωφελεί και οι ερευνητές δεν χρειάζεται να έχουν πρόσβαση σε αυτή. Ωστόσο, ο κίνδυνος ταυτοποίησης ενός ασθενούς με έμμεσο τρόπο μέσω άλλων πληροφοριών του δεν είναι μικρός, πληροφορίες που χρειάζονται για την μελέτη όπως το φύλο, η ηλικία, ο τόπος διαμονής κλπ. Στο πλαίσιο αυτό, η μελέτη εξετάζει την εφαρμογή ενός συστήματος ψευδωνυμοποίησης, που επιτρέπει την αξιοποίηση των συσχετίσεων μεταξύ ασθενών που μοιράζονται παρόμοια χαρακτηριστικά, ενώ παράλληλα δεν αποκαλύπτεται η πραγματική ταυτότητα του συμμετέχοντος. Έτσι, καθώς σε τέτοιες έρευνες όπως οι κλινικές δοκιμές συλλέγονται αρκετοί τύποι προσωπικών δεδομένων, η ψευδωνυμοποίηση θα πρέπει να προστατεύει τους συμμετέχοντες από μη εξουσιοδοτημένη αναγνώριση. Αυτό μπορεί να επιτευχθεί συνδυάζοντας δύο προσεγγίσεις,

1) Εφαρμογή ψευδωνυμοποίησης στα κύρια αναγνωριστικά δεδομένα του κάθε συμμετέχοντος

2) Χρήση περισσότερων του ενός ψευδωνύμων για κάθε αναγνωριστικό δεδομένο για διαφορετικές κλινικές παραμέτρους

Η προσέγγιση αυτή μπορεί να περιορίσει τα προσωπικά δεδομένα που σχετίζονται με κάθε ψευδώνυμο, αξιοποιώντας μία ισχυρή συνάρτηση κατακερματισμού όπως η SHA-2.

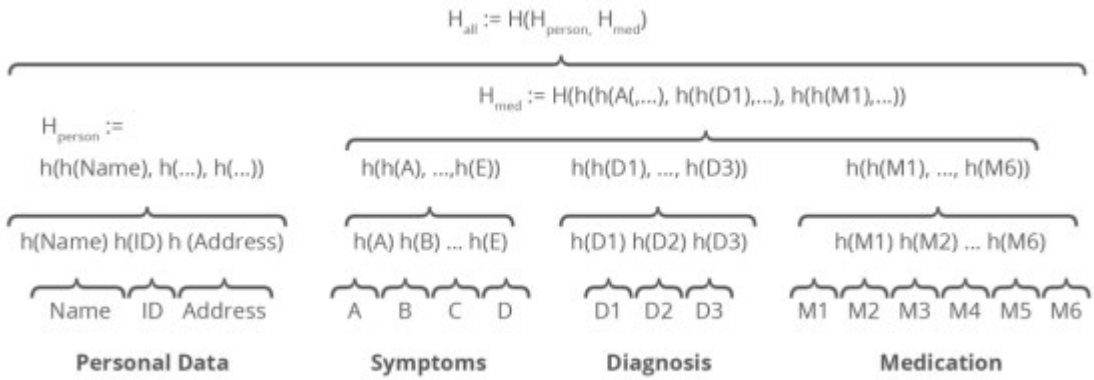


Εικόνα 3.9: Ψευδωνυμοποίηση με συνάρτηση κατακερματισμού

Όπως μπορούμε να δούμε και στην εικόνα 3.9, για συγκεκριμένο συμμετέχων έχουν δημιουργηθεί δύο ζεύγη παραμέτρων, κατά την υλοποίηση της ψευδωνυμοποίηση με συνάρτηση κατακερματισμού. Προκειμένου να καταστίσουν την ταυτοποίηση ακόμα πιο δύσκολη, στη διαδικασία εισήγαγαν και μία τυχαία τιμή 'salt', διαφορετική για κάθε ζεύγος κλινικών παραμέτρων, δημιουργώντας έτσι δύο διαφορετικά ψευδώνυμα για τον ίδιο συμμετέχων.

Προχωρώντας σε πιο προηγμένες τεχνικές ψευδωνυμοποίησης, μία άλλη μελέτη που αναλύεται από τον ENISA [28] περιγράφει, μέσω μιας προσπάθειας ασφαλέστερης υλοποίησής της, ένα σενάριο διαμοιρασμού εγγραφών ασθενών μεταξύ δύο νοσοκομείων. Στην υπόθεση αυτή τα δύο νοσοκομεία μοιράζονται την ίδια ενημερωμένη έκδοση ιατρικών φακέλων συγκεκριμένων ασθενών. Ωστόσο λόγω κάποιων καθυστερήσεων κατά τη ψηφιοποίηση μετά τη μεταφορά τους μεταξύ των δύο εμπλεκόμενων φορέων και σε μία προσπάθεια τα δεδομένα αυτά να παραμένουν σαφή και ενημερωμένα, καθίσταται αναγκαία η εκτέλεση πρωτοκόλλου ώστε να συγκρίνει τις εγγραφές αυτές και να τις κρατάει ενημερωμένες στις δύο πλευρές. Τα δεδομένα αυτά αποτελούν προσωπικά ευαίσθητα δεδομένα ασθενών καθώς πέρα από δημογραφικά στοιχεία περιέχουν και πληροφορίες υγείας όπως ασθένειες, συμπτώματα και ιατρικές διαγνώσεις. Στη συνέχεια περιγράφει τις ανάγκες ψευδωνυμοποίησης των δεδομένων αυτών καθώς χωρίς αυτή θα ήταν αδύνατο να μην αποκαλυφθούν τα προσωπικά δεδομένα των ασθενών.

Η προσέγγιση που ακολουθεί η συγκεκριμένη περίπτωση είναι με χρήση συναρτήσεων κατακερματισμού, ωστόσο πάει τη τεχνική αυτή ένα βήμα παραπέρα διασφαλίζοντας επιπλέον μέτρα και καθιστώντας σχεδόν αδύνατη την αποκάλυψη ευαίσθητων πληροφοριών σε τρίτες μη εξουσιοδοτημένες οντότητες. Ουσιαστικά ορίζει κάποια επίπεδα ψευδωνυμοποίησης με τη μορφή δένδρου Merkle κατακερματισμών [29].



Εικόνα 3.10: Ψευδωνυμοποίηση πολλαπλών αναγνωριστικών [28]

Όπως φαίνεται από την εικόνα 3.10 έχει χωρίσει τα προσωπικά δεδομένα σε τέσσερις κατηγορίες (δημογραφικά, συμπτώματα, διαγνώσεις, αγωγές). Στο πρώτο επίπεδο ακολουθεί το κατακερματισμό του κάθε γνωρίσματος ξεχωριστά. Ακολουθώντας ωστόσο τη ντετερμινιστική πολιτική ψευδωνυμοποίησης, μπορεί να αποκαλύψει αρκετές πληροφορίες σε περιπτώσεις κοινών ονομάτων, ή κοινών διαγνώσεων. Έτσι περνάει σε δεύτερο επίπεδο ψευδωνυμοποίησης όπου αντί για τον κατακερματισμό μεμονωμένων γνωρισμάτων, ο κατακερματισμός εφαρμόζεται ξεχωριστά σε ολόκληρες τις εκάστοτε κατηγορίες (δημογραφικά, διαγνώσεις κλπ.) ακολουθώντας μία μορφή κατακερματισμού πολλαπλών γνωρισμάτων, προσφέροντας ένα ακόμα επίπεδο ασφαλείας στην υλοποίηση της ψευδωνυμοποίηση των προσωπικών δεδομένων των ασθενών. Ανεβαίνοντας ένα ακόμα επίπεδο ακολουθώντας την ίδια διαδικασία χωρίζει τις κατηγορίες σε δημογραφικά δεδομένα και δεδομένα ιατρικού χαρακτήρα και δημιουργεί μία ακολουθία κατακερματισμού των δύο κατηγοριών για κάθε εγγραφής, περιορίζοντας έτσι, στο τρίτο επίπεδο, τη σύγκριση των δύο ακολουθιών μεταξύ των νοσοκομείων αυτών χωρίς το φόβο της αποκάλυψης κάποιας πληροφορίας. Τέλος μπορεί να κατακερματίσει το σύνολο όλων των δεδομένων της εγγραφής εκμηδενίζοντας έτσι την όποια ταυτοποίηση φυσικού προσώπου από τρίτους, ωστόσο στο σημείο αυτό τα δεδομένα είναι χρήσιμα μόνο για τη σύγκριση μεταξύ των δύο κατακερματισμένων τιμών από του δύο φορείς και δεν μπορούν να αποκαλύψουν καμία πληροφορία.

Μία άλλη μελέτη στην ίδια αναφορά του ENISA παρουσιάζει την ανάγκη αποθήκευσης ιατρικών δεδομένων σε περισσότερες από μία βάσεις διαφορετικών οργανισμών για λόγους ασφάλειας και διαθεσιμότητας [28]. Στη περίπτωσή μας τα ιατρικά δεδομένα των ασθενών αποθηκεύονται σε διαφορετικά νοσοκομεία, κάτι που θέτει επιτακτική την ψευδωνυμοποίηση των αναγνωριστικών, καθώς σε άλλη περίπτωση θα μπορούσε ο κάθε ένας να έχει πρόσβαση σε όλα τα αναγνωριστικά άλλων φορέων όπως ονόματα, διευθύνσεις κλπ. κάτι που θα αποτελούσε ρίσκο ως προς τη ιδιωτικότητα και τα δεδομένα των ασθενών. Για την προστασία λοιπόν των προσωπικών δεδομένων μπορεί να υλοποιηθεί κρυπτογράφηση των αναγνωριστικών ως μέσο ψευδωνυμοποίησης ώστε να μπορούν να διαμοιραστούν τα δεδομένα αυτά με ασφάλεια μεταξύ των εμπλεκόμενων φορέων. Έτσι τα δεδομένα υγείας παραμένουν διαθέσιμα στους εμπλεκόμενους φορείς, ωστόσο η πρόσβαση στα πραγματικά αναγνωριστικά περιορίζεται μόνο σε εξουσιοδοτημένους χρήστες που έχουν και πρόσβαση στο κλειδί της κρυπτογράφησης. Έχοντας λοιπόν το κλειδί αυτό ως μυστικό μέσο της ψευδωνυμοποίησης, προτείνεται μία προσέγγιση διαχωρισμού του μυστικού κλειδιού σε ξεχωριστά μέρη όπου κάθε νοσοκομείο θα έχει το δικό του μυστικό μέρος του κλειδιού. Ως εκ τούτου σε περίπτωση επαναπροσδιορισμού του αρχικού αναγνωριστικού θα πρέπει όλα τα εμπλεκόμενα νοσοκομεία να εκχωρήσουν το δικό τους μυστικό μέρος ώστε να αποκρυπτογραφηθεί το ψευδώνυμο και να ταυτοποιηθεί ο ασθενής. Αυτή η τεχνική μπορεί να εφαρμοστεί με συμμετρική καθώς και με ασύμμετρη κρυπτογράφηση.

Στη περίπτωση της συμμετρικής χρησιμοποιείται το ίδιο μυστικό κλειδί για την κρυπτογράφηση και την αποκρυπτογράφηση, έτσι κανείς δεν μπορεί να αποκαλύψει το ψευδώνυμο χωρίς να έχει πρόσβαση σε ολόκληρο το μυστικό της ψευδωνυμοποίησης.

Όσον αφορά την ασύμμετρη κρυπτογράφησης περιλαμβάνει δύο είδη κλειδιών, το δημόσιο και το ιδιωτικό. Σε αυτή τη περίπτωση το ιδιωτικό κλειδί αφορά το μυστικό που χρειάζεται για να αποκαλύψουμε τα ψευδώνυμα που δημιουργήθηκαν με το δημόσιο κλειδί.

3.2 Εφαρμογές Ανωνυμοποίησης

Στην ενότητα αυτή θα αναφερθούμε στην ανωνυμοποίηση δεδομένων δίνοντας ιδιαίτερη έμφαση στον τομέα τη υγειονομικής περίθαλψης και τα δεδομένα υγείας. Σκοπό έχουμε να εξετάσουμε ορισμένες έρευνες βασιζόμενες στην ανωνυμοποίηση δεδομένων υγείας, στοχεύοντας σε μία βαθύτερη κατανόηση της τρέχουσας κατάστασης και δημιουργώντας τις κατάλληλες κατευθύνσεις στα πλαίσια μίας διαδικασίας έρευνας. Έχουμε ήδη αναφερθεί στο γεγονός ότι η

ανωνυμοποίηση αποτελεί μία αρκετά απαιτητική διαδικασία, καθώς πρέπει πρώτα να καθοριστούν οι ανάγκες πραγματοποίησης της έρευνας. Επίσης υπάρχουν αρκετές τεχνικές εφαρμογής της μέσω κατάλληλα διαμορφωμένων εργαλείων, παρ' όλα αυτά στην παρούσα έρευνα στοχεύουμε να εστιάσουμε στις βασικότερες από τις τεχνικές αυτής.

3.2.1 Ανωνυμοποίησης Δεδομένων Υγείας

Έρευνα του 2017 μελετά την ανωνυμοποίηση σε ιατρικά δεδομένα ασθενών [34]. Παρουσιάζει την ανωνυμοποίηση των δεδομένων υγείας συνδυαστικά με την επιστήμη της Πληροφορικής ως ένα πολύ καίριο θέμα των σημερινών ημερών σε μία προσπάθεια εξαγωγής χρήσιμων συμπερασμάτων από τα δεδομένα υγείας, διατηρώντας πάντα ασφαλή την ιδιωτικότητα τους ασθενούς, βάση των κανονισμών που τη προστατεύουν.

Το σενάριο αυτό αποτελεί μία πραγματική περίπτωση νοσοκομείου, περιέχοντας ένα ικανοποιητικά μεγάλο σύνολο δεδομένων με αρκετά γνωρίσματα δημογραφικά, ιατρικά αλλά και ευαίσθητα δεδομένα ασθενειών. Το σύνολο αυτό θα πρέπει να ανωνυμοποιηθεί ώστε να αποσταλεί σε ερευνητικά κέντρα για ανάγκες έρευνας.

Στην έρευνα αυτή έχει χρησιμοποιηθεί το εργαλείο ARX το οποίο είναι ένα ανοιχτού κώδικα λογισμικό ανωνυμοποίησης δεδομένων [35]. Για να καλύψουν τις ανάγκες της έρευνας, το διαμοιρασμό δηλαδή του συνόλου δεδομένων, οι ερευνητές αποφάσισαν να ανωνυμοποιήσουν τα δεδομένα αυτά παρουσιάζοντας τη μέθοδο που ακολούθησαν και αξιολογώντας τα ανώνυμα δεδομένα ως προς τη διατήρηση της χρησιμότητάς τους αλλά και τους κινδύνους αποκάλυψης της ταυτότητας των υποκειμένων των δεδομένων αυτών. Κατά τη προσέγγισή τους χρησιμοποιούν την k-ανωνυμία για τιμή k = 3 λόγω του μεγάλου και πολυδιάστατου συνόλου δεδομένων, στοχεύοντας να κρατήσουν ένα υψηλό επίπεδο χρήσιμης πληροφορίας καθώς και λόγω του περιορισμού μεταξύ των φορέων που θα διαμοιραστούν τα δεδομένα αυτά. Όσον αφορά την αποκάλυψη γνωρισμάτων, οι ερευνητές αποφάσισαν να μη χρησιμοποιήσουν κάποιο από τα μοντέλα απορρήτου όπως η l-διαφορετικότητα καθώς το εμπιστευτικό γνώρισμα (ασθένεια) εμφανίζεται σε αρκετά χαμηλό ποσοστό.

Μελέτη που πραγματοποιήθηκε από το υπουργείο υγείας της Γαλλίας το 2013 για πειραματικούς σκοπούς εστιάζει στα δεδομένα υγείας και τον ασφαλή διαμοιρασμό τους καθώς και το ρίσκο επαναπροσδιορισμού της ταυτότητας του συνόλου δεδομένων μετά την ανωνυμοποίησή τους [36].

Το έργο τους εστιάζει στα δεδομένα υγείας, σε τεχνικές ανωνυμοποίησης και στο πως τα πραγματικά δεδομένα θα μπορούσαν να οργανωθούν ώστε να παρέχουν χρήσιμη πληροφορία και να διατηρούνται ως απόρρητα ώστε να μπορούν να τεθούν για δημόσια χρήση. Για τη συγκεκριμένη έρευνα χρησιμοποιούνται δύο εργαλεία ανωνυμοποίησης δεδομένων (το ARX και το μ-Argus). Πριν ξεκινήσουν την ανωνυμοποίηση βέβαια οι ερευνητές έδωσαν ιδιαίτερη προσοχή στο ίδιο το σύνολο δεδομένων αναλύοντας αρκετές πτυχές όπως τα χαρακτηριστικά που εμφανίζονται πολύ σπάνια στο σύνολο και άλλες πληροφορίες ώστε να κρατήσουν σε πολύ υψηλό επίπεδο τη χρησιμότητα των ανώνυμων δεδομένων. Κατά την προσέγγιση ανωνυμοποίησης των δεδομένων οι ερευνητές χρησιμοποίησαν τα μοντέλα απορρήτου k-anonymity και l-diversity για τιμές k = 10 και l = 3 στοχεύοντας να κρατήσουν τα δεδομένα ασφαλή από επικείμενες επιθέσεις αποκάλυψης της ταυτότητας ή αποκάλυψης των γνωρισμάτων.

Τέλος υλοποίησαν δύο προσεγγίσεις της ανωνυμοποίησης, μία με το εργαλείο μ-Argus και μία με το ARX και εξήγαγαν τα συμπεράσματά τους. Κατά τα αποτελέσματά τους φάνηκε η εξαιρετική δυσκολία μιας καλής αντιστάθμισης μεταξύ του ρίσκου αποκάλυψης και της διατήρησης χρήσιμης πληροφορίας μέσω της ανωνυμοποίησης.

Μία άλλη πιο πρόσφατη έρευνα του 2019 δοκιμάζει μία εντελώς διαφορετική προσέγγιση της ανωνυμοποίησης δεδομένων υγείας και υπόσχεται πολύ ικανοποιητικά αποτελέσματα [22]. Κατά τη προσέγγιση αυτή οι ερευνητές, για την ανωνυμοποίηση των δεδομένων υγείας, έχοντας ήδη ερευνήσει τις «κλασικές» τεχνικές ανωνυμοποίησης, παρουσιάζουν μία εναλλακτική ανωνυμοποίηση δεδομένων υγείας, με τη χρήση κρυπτογραφικών αλγορίθμων. Πιο συγκεκριμένα έχοντας τα δεδομένα υγείας από ένα νοσοκομείο, αρχικά έχουν κατηγοριοποιήσει τα γνωρίσματα σε αναγνωριστικά, ψευδο-αναγνωριστικά και εμπιστευτικά ώστε να ανωνυμοποιηθούν μέσω της κρυπτογράφησής τους. Πρακτικά ο στόχος τους (και γενικότερα ο στόχος της ανωνυμοποίησης) είναι η διατήρηση του απορρήτου και της χρησιμότητας των δεδομένων. Για την επίτευξη αυτή έχουν επιλεγεί τέσσερις αλγόριθμοι κρυπτογραφίας, δύο συμμετρικής (DES και AES) και δύο ασύμμετρης (RSA και ElGamal), δημιουργώντας έτσι ένα ικανοποιητικό επίπεδο ασφαλείας στα δεδομένα, καθώς ο επιτιθέμενος δεν μπορεί να αντιστοιχίσει τα πραγματικά δεδομένα χωρίς το κλειδί της κρυπτογράφησης. Στη δεύτερη φάση της ανωνυμοποίησης το σύνολο των δεδομένων ομαλοποιείται και συγκρίνονται τα δύο σύνολα (αρχικά και ανώνυμα δεδομένα).

Τέλος συγκρίνουν τα αποτελέσματά τους με τεχνικές όπως η k-ανωνυμία και αποδεικνύουν βασιζόμενοι στη μελέτη τους, ότι τα ανώνυμα δεδομένα διατηρούν αποτελεσματικότερα τη χρησιμότητά τους συγκριτικά με τη k-ανωνυμία.

3.3 Αξιοποίηση Και Προστασία Δεδομένων Υγείας

Στην ενότητα αυτή θα εξετάσουμε κάποιους κανονισμούς που αφορούν τα προσωπικά δεδομένα υγείας και την ανάγκη προστασίας τους αλλά και διάθεσής τους για ανάγκες έρευνας και αποτελεσματικότερης απόδοσης υγειονομικής περίθαλψης. Στα πλαίσια αυτά θα εξεταστεί και μία πρόταση της Ευρωπαϊκής Επιτροπής για την αξιοποίηση των δεδομένων υγείας, επωφελώντας τόσο τον ίδιο τον ασθενή, όσο και τις ανάγκες ερευνητικού ενδιαφέροντος.

3.3.1 Κανονισμός Ευρωπαϊκού Χώρου Δεδομένων Υγείας

Στο προηγούμενο κεφάλαιο κάναμε μία σύντομη αναφορά στη νέα πρόταση της Ευρωπαϊκής Ένωσης για τον Ευρωπαϊκό Χώρο Δεδομένων Υγείας (EHDS) [15]. Όπως αναφέραμε η πρόταση αυτή στοχεύει στην υποστήριξη των υποκειμένων των δεδομένων όσον αφορά τον έλεγχό τους σε αυτά, στη καλύτερη εξυπηρέτηση των δεδομένων υγεία για περιπτώσεις αποτελεσματικότερης περίθαλψης, ανάγκες έρευνας και καινοτομίας, καθώς και γενικότερα τη δυνατότητα της ΕΕ να μπορεί να αξιοποιήσει και να επαναχρησιμοποιήσει τα δεδομένα υγείας για το διαμοιρασμό τους πάντα με ασφάλεια. Πάνω στο πεδίο αυτό, με τον Κανονισμό EHDS η ΕΕ προσπαθεί να δημιουργήσει ένα οικοσύστημα για την υγεία το οποίο θα ακολουθεί κοινούς κανόνες και πρακτικές, όπως και υποδομές αποσκοπώντας στα παρακάτω.

1. Στην ενδυνάμωση των δυνατοτήτων των ατόμων μέσω της ψηφιακής πρόσβασης, να μπορούν να έχουν τον έλεγχο των ηλεκτρονικών δεδομένων της υγείας τους σε Εθνικό και Ευρωπαϊκό επίπεδο (κύρια χρήση των δεδομένων).

2. Στη παροχή των δεδομένων αυτών, ως αξιόπιστη και αποτελεσματική πηγή χρήσης τους σε έρευνες, δραστηριότητες καινοτομίας και χάραξης πολιτικής από τρίτους (δευτερογενής χρήση των δεδομένων).

Ως κανονισμός, ο EHDS προτάθηκε το 2022, και σημαντικό ρόλο σε αυτό έπαιξε η πανδημία COVID-19 η οποία ανέδειξε την ανάγκη ύπαρξης ενιαίων και ενημερωμένων δεδομένων υγείας

προκειμένου να λαμβάνονται πιο άμεσα αποφάσεις που σχετίζονται με τη δημόσια υγεία καθώς και τη διαχείριση κρίσεων στο τομέα της υγείας. Η πρόταση αυτή δείχνει να έχει μεγάλα πλεονεκτήματα τόσο αναφορικά με τη κύρια χρήση της όσο και τη δευτερογενή που αφορά τις ανάγκες επιστημονικής έρευνας, καινοτομίας και χάραξης πολιτικής [37] [2].

Καθώς τα δεδομένα αυτά, κατά τον GDPR αποτελούν πληροφορίες ευαίσθητου χαρακτήρα, από νομικής βάσης, η πρόταση αναφέρεται στον GDPR περί της νόμιμης επεξεργασίας των δεδομένων αυτών και επικεντρώνεται στα άρθρα 6 και 9, και πιο συγκεκριμένα ειδικά για τη δευτερογενή χρήση τους στη παράγραφο 2 του άρθρου 9, στοιχεία ζ) έως ι). Επίσης βάση των κανονισμών οι χρήστες των δεδομένων θα πρέπει να αποδείξουν κατά το άρθρο 6 παράγραφος 1 στοιχεία ε) ή στ) ότι συμμορφώνονται με αυτά και η πρόσβαση στα δεδομένα και η χρήση τους είναι απαραίτητη για την εκτέλεση εργασιών έχοντας έννομο συμφέρον. Επίσης για τους κατόχους των δεδομένων βάση του άρθρου 6 παράγραφος 1 στοιχείο γ) του GDPR θα πρέπει να αποκαλύπτουν τις πληροφορίες αυτές σε φορείς που αιτούνται πρόσβαση και διασφαλίζουν ότι η πρόσβαση σε αυτά παρέχεται βάσει των λόγων που την έχουν ζητήσει.

Όπως αναφέραμε και παραπάνω η πρόταση αυτή εστιάζει τόσο στην κύρια χρήση των δεδομένων υγείας, όσο και τη δευτερογενή τους χρήση. Παρακάτω θα εξετάσουμε τις δύο αυτές περιπτώσεις και τα πλεονεκτήματα που μπορούν να μας δώσουν.

Για τη κύρια χρήση των δεδομένων η πρόταση αναφέρεται στην επεξεργασία προσωπικών ηλεκτρονικών δεδομένων υγείας που αποσκοπεί στην άμεση υγειονομική περίθαλψη του ασθενούς. Η χρήση αυτή στοχεύει ώστε να κάνει ευκολότερη, για τους Ευρωπαίους πολίτες τη πρόσβαση και το διαμοιρασμό των δεδομένων υγείας τους εκτός συνόρων, αλλά εντός της Ευρωπαϊκής Ένωσης και συνδέεται με τα δικαιώματα πρόσβασης και λήψης πληροφοριών, διόρθωση και φορητότητα των δεδομένων τους βάσει του GDPR. Με το τρόπο αυτό οι ασθενείς θα έχουν τη δυνατότητα να προσθέτουν πληροφορίες στα ηλεκτρονικά δεδομένα υγείας τους, να ελέγχουν τη πρόσβαση σε αυτά και να μπορούν να ελέγξουν ποιος τα βλέπει. Για την αποτελεσματικότερη πρόσβαση στα δεδομένα αυτά από παρόχους υγειονομικής περίθαλψης σε ολόκληρη την Ευρωπαϊκή Ένωσή, κάθε κράτος μέλος θα πρέπει να ορίσει μία αρχή ψηφιακής υγείας ως υπεύθυνη για την εφαρμογή και τη διασφάλιση της συνεργασία με άλλα κράτη για την ανταλλαγή δεδομένων υγείας μεταξύ τους. Έτσι αντιλαμβανόμαστε ότι με αυτό το τρόπο μέσω του διαμοιρασμού των δεδομένων υγείας μεταξύ άλλων κρατών μελών της ΕΕ, ως πολίτες και επισκέπτες σε μία άλλη χώρα, θα μπορούμε κατά της επίσκεψή μας σε μονάδα υγείας αυτής να έχουμε πρόσβαση ηλεκτρονικά στα δεδομένα υγείας μας.

Από την άλλης η δευτερογενής χρήση των δεδομένων υγείας περιλαμβάνει την επεξεργασία τους για σκοπούς επιστημονικής έρευνας, καινοτομίας, χάραξης πολιτικής και άλλους παρόμοιους σκοπούς. Το άρθρο 33 της πρότασης του EHDS αναγράφει τις ελάχιστες κατηγορίες ηλεκτρονικών δεδομένων που μπορούν να αξιοποιηθούν για δευτερογενή χρήση. Μέσα σε αυτές περιλαμβάνονται τα ηλεκτρονικά δεδομένα υγείας (EHRs), γενετικά δεδομένα, δεδομένα κλινικών δοκιμών και άλλα. Συνεχίζει με το άρθρο 34 της πρόταση όπου αναφέρει τους σκοπούς που επιτρέπεται η χρήση των δεδομένων αυτών όπως δραστηριότητες δημοσίου συμφέροντος, επιστημονικές έρευνες, δραστηριότητες εκπαίδευσης στην υγεία κλπ., ενώ το άρθρο 35 περιλαμβάνει απαγορευμένους σκοπούς χρήσης τους. Κατά τη δευτερογενή χρήση των δεδομένων αυτών, ο χρήστης των δεδομένων υποβάλει αίτημα πρόσβασης σε αυτά, η οποία πρόσβαση αξιολογείται και παρέχεται από τον φορέα πρόσβασης στα δεδομένα υγείας, ο οποίος είναι και υπεύθυνος για την παροχή άδειας πρόσβασης ανάλογα με τους σκοπούς χρήσης των δεδομένων κατά την υποβολή [2].

Ωστόσο, παρά τα προτερήματα που δείχνει να φέρει ο κανονισμός αυτός ως πρόταση, είναι σημαντικό να αναφερθούν και κάποιες επιφυλάξεις που έχουν εκφράσει τόσο το Ευρωπαϊκό Συμβούλιο Προστασίας Δεδομένων (EDPB) όσο και ο Ευρωπαϊκός Επόπτης Προστασίας Δεδομένων (EDPS) αναφορικά με τους κινδύνους που μπορεί να δημιουργήσει ο κανονισμός αυτός, εξετάζοντας κάποιες από τις διατάξεις του. Πιο συγκεκριμένα, ενώ συμμερίζονται και κατανοούν τους σκοπούς της πρότασης τόσο για την κύρια αλλά και τη δευτερογενή χρήση των δεδομένων υγείας, θεωρούν ότι σε κάποιες περιπτώσεις μπορεί να αποδυναμώσει τη προστασία των δικαιωμάτων στην ιδιωτικότητα και τη προστασία των δεδομένων κυρίως για τους σκοπούς που σχετίζονται με τη δευτερογενή χρήση τους. Για την ακρίβεια επισημαίνουν κάποια σημεία της πρότασης, τα οποία δεν είναι εντελώς ξεκάθαρα σε σχέση με τους κανονισμούς του GDPR και απαιτούν περεταίρω σαφήνεια και σε κάποιες περιπτώσεις σχετικά με τα άρθρα της, κρίνουν ότι πρέπει να ενισχυθούν ή να τροποποιηθούν ώστε να συμβαδίζουν με τον GDPR, ενώ σε άλλες να καταργηθούν [38].

Συνοψίζοντας η πρόταση αυτή έχει ιδιαίτερο ενδιαφέρον για το τομέα της υγειονομικής περίθαλψης καθώς και τις ανάγκες επιστημονικής έρευνας. Βάσει των στόχων που έχει θέσει θα μπορούσε να φέρει επαναστατικά αποτελέσματα στον τομέα συνδυάζοντας την επιστήμη της πληροφορικής και να διευκολύνει σε μεγάλο βαθμό ολόκληρη την ιατρική κοινότητα και γενικότερα τη παροχή μίας αποτελεσματικότερης υγειονομικής περίθαλψης σε Ευρωπαϊκό επίπεδο.

3.3.2 Νόμος HIPAA

Ο Νόμος περί φορητότητας και λογοδοσίας για την ασφάλιση της υγείας (HIPAA) αποτελεί ομοσπονδιακό νόμο που θεσπίστηκε στις Ηνωμένες Πολιτείες το 1966 με σκοπό να προστατεύσει την ιδιωτικότητα και την ασφάλεια των πληροφοριών υγείας των ασθενών [39]. Ορίζει τη δημιουργία εθνικών προτύπων για τη προστασία ευαίσθητων πληροφοριών της υγείας των ασθενών από την αποκάλυψη χωρίς τη συγκατάθεσή τους ή εν' αγνοία τους. Ο Κανονισμός Απορρήτου HIPAA ισχύει για όλους τους φορείς υγείας, όπως νοσοκομεία, ασφαλιστικές και φαρμακευτικές εταιρείες και άλλους παρόχους υγειονομικής περίθαλψης.

Ο Κανονισμός αυτός αφορά τη χρήση καθώς και την αποκάλυψη πληροφοριών υγείας ατόμων που υπόκεινται στον Κανόνα Προστασίας Προσωπικών Δεδομένων, αναφέροντας τα άτομα αυτά ως καλυπτόμενες οντότητες. Ο κανόνας απορρήτου περιέχει επίσης πρότυπα για τα δικαιώματα των ατόμων να μπορούν να ελέγχουν το τρόπο χρήσης των πληροφοριών υγείας τους από τρίτους. Στόχος είναι να διασφαλιστεί ότι οι πληροφορίες υγείας των ατόμων προστατεύονται κατάλληλα, διατηρώντας τις ωστόσο έτσι ώστε να παρέχουν υψηλή ποιότητα υγειονομικής περίθαλψης.

Ο Κανονισμός αυτός καλύπτει ένα ευρύ φάσμα στο τομέα της υγείας συμπεριλαμβανομένου του απορρήτου, της ασφάλειας και της ειδοποίησης παραβίασης. Έτσι οι πάροχοι υγειονομικής περίθαλψης θα πρέπει να θέτουν σε εφαρμογή κατάλληλες τεχνικές διασφαλίζοντας τη προστασία των πληροφοριών των ασθενών από μη εξουσιοδοτημένες προσβάσεις σε αυτές. Οι πληροφορίες αυτές ονομάζονται προστατευμένες πληροφορίες υγείας (PHI) [40].

Ως καλυπτόμενες οντότητες ορίζουμε φυσικά πρόσωπα και οργανισμούς ως υποκείμενα του κανόνα απορρήτου. Σε αυτά συμπεριλαμβάνονται οι πάροχοι υγειονομικής περίθαλψης που εμπλέκονται και μεταφέρουν ηλεκτρονικά πληροφορίες υγείας, τα σχέδια υγείας και τα γραφεία συμψηφισμού υγειονομικής περίθαλψης. Στα υγειονομικά σχέδια ανήκουν οι ασφαλιστές υγείας, οργανισμοί συντήρησης υγείας, ασφαλιστές μακροχρόνιας περίθαλψης, διάφορα προγράμματα υγείας.

Στη συνέχεια περιγράφονται περιπτώσεις όπου επιτρέπεται από μία καλυπτόμενη οντότητα να χρησιμοποιεί και να αποκαλύπτει τις προστατευμένες πληροφορίες υγείας, χωρίς την εξουσιοδότηση κάποιου ατόμου περιγράφοντας τους ακόλουθους σκοπούς:

1. Αποκάλυψη των πληροφοριών στο ίδιο το άτομο

2. Ανάγκες θεραπείας και πράξεις υγειονομικής περίθαλψης

3. Περιορισμένο σύνολο των πληροφοριών για ανάγκες έρευνας ή δημόσιας υγείας

4. Δυνατότητα συμφωνίας ή αντίρρησης για γνωστοποίηση των πληροφοριών

5. Δραστηριότητες δημοσίου συμφέροντος όπως η δημόσια υγεία, όταν απαιτείται από το νόμο, επιβολή του νόμου, έρευνες υπό προϋποθέσεις, βασικές κυβερνητικές λειτουργείες, πρόληψη ή μείωση σοβαρής απειλής για την υγεία ή την ασφάλεια καθώς και άλλες νόμιμες περιπτώσεις.

Τέλος έχουμε τους κανόνες ασφάλειας οι οποίοι προστατεύουν ένα υποσύνολο πληροφοριών που καλύπτονται από τον κανόνα απορρήτου. Για τη συμμόρφωση με τους κανόνες ασφάλειας όλες οι καλυπτόμενες οντότητες θα πρέπει να διαβεβαιώσουν την εμπιστευτικότητα, την ακεραιότητα και τη διαθεσιμότητα όλων των προστατευόμενων πληροφοριών υγείας, να προστατεύουν τις πληροφορίες αυτές από επικείμενες μη επιτρεπόμενες χρήσεις τους ή αποκάλυψή τους, να εντοπίζουν και να προστατεύουν τις πληροφορίες έναντι αναμενόμενων απειλών για την ασφάλειά τους [40].

Κεφάλαιο 4
Συλλογή Δεδομένων

Προχωρώντας προς το πρακτικό σκέλος της παρούσας διατριβής, όπου θα εστιάσουμε σε μία μελέτη περίπτωσης ως ρεαλιστικό σενάριο, αρχικά θα πρέπει να συλλέξουμε τα δεδομένα υγείας που θα αξιοποιήσουμε. Η επιλογή της μεθόδου αποτελεί ένα σημαντικό κομμάτι καθώς εξετάζουμε τα πλεονεκτήματα και τα μειονεκτήματα της κάθε μεθόδου, αλλά και μία πρόκληση. Σημαντικό είναι να έχουμε προσδιορίσει έως τώρα τις ανάγκες της έρευνας καθώς είναι ένας παράγοντας που παίζει καθοριστικό ρόλο στην επιλογή μας. Πιο συγκεκριμένα θα πρέπει να εξετάσουμε τον στόχο της μελέτης και την ανάγκη αξιοποίησης πραγματικών δεδομένων υγείας ή άλλων ρεαλιστικών προσεγγίσεων.

4.1 Είδη Δεδομένων

Όσον αφορά τα δεδομένα υγείας που μπορούμε να χρησιμοποιήσουμε για την μελέτη μας υπάρχουν διαφορετικές προσεγγίσεις και τρόποι συλλογής τους [21]. Πέραν των πραγματικών δεδομένων υγείας από εγγραφές ιατρικών φακέλων, ασφαλιστικών ταμείων και γενικότερα νοσοκομείων ή κέντρων υγείας, υπάρχουν και δεδομένα διαδικτυακά διαθέσιμα που προσομοιάζουν τα πραγματικά και μπορούν να αξιοποιηθούν για ανάγκες έρευνας. Πρόκειται για δεδομένα διαθέσιμα που έχουν χρησιμοποιηθεί σε προηγούμενες μελέτες, ή έχουν δημιουργηθεί για παρόμοιες ανάγκες. Τέλος τα τελευταία χρόνια έχει αναπτυχθεί τεχνολογία παραγωγής τυχαίων ή ψευδο-τυχαίων δεδομένων, μία εναλλακτική προσέγγιση που υπόσχεται ικανοποιητικά αποτελέσματα και ξεπερνά πολλά εμπόδια των παραπάνω μεθόδων.

4.1.1 Πραγματικά Δεδομένα Υγείας

Τα πραγματικά δεδομένα υγείας (Real-World HealthCare Data) αφορούν δεδομένα που αποτελούν κομμάτι των μεγάλων δεδομένων (Big Data) στον τομέα της υγείας και πρόκειται για μεγάλες ποσότητες πληροφορίας από διάφορες πηγές, όπως ηλεκτρονικές εγγραφές υγείας (EHRs), συσκευές παρακολούθησης υγείας, δεδομένα υγείας από βάσεις δεδομένων διαφόρων

φορέων υγείας [41]. Αφορά δεδομένα που αντιστοιχούν σε πραγματικούς ασθενείς και δεν έχουν προκύψει από κάποια τυχαιοποιημένη επεξεργασία ή πειραματική μελέτη. Πρόκειται για ευαίσθητα και εμπιστευτικά δεδομένα, αφορούν υπαρκτά πρόσωπα και χρήζουν ειδικής μεταχείρισης λόγω της μεγάλης ανάγκης προστασίας τους όσον αφορά την ιδιωτικότητα. Έχουν οριστεί από τον GDPR ως ευαίσθητα προσωπικά δεδομένα, ο οποίος καθορίζει συγκεκριμένες μεθόδους και συνθήκες κάτω από τις οποίες μπορούμε να τα αξιοποιήσουμε. Μπορούν να συλλεχθούν από νοσοκομεία ή διάφορα ιδρύματα υγείας, μέσω εγγραφών ιατρικών φακέλων, ασφαλιστικών ταμείων, λογαριασμούς κλινικών ή ιατρικών συνεντεύξεων [42]. Πιο συγκεκριμένα οι πηγές προέλευσής τους είναι κυρίως μητρώα ασθενών, βάσεις δεδομένων υγειονομικής περίθαλψης, καθώς και εταιρείες ασφαλιστικής υγείας. Γενικότερα θεωρούνται ως μία πολύ καλά δομημένη μορφή πληροφορία και μπορεί να εντοπιστεί σε κατάλληλα σχεδιασμένα αποθετήρια. Η χρήση τους κρίνεται επιτακτική σε μελέτες όπως το να εξακριβωθεί η αποτελεσματικότητα κάποιας θεραπείας, την πρόληψη εμφάνισης κάποιας επιδημίας ή την εξέταση κάποιου φαρμάκου λόγω κυρίως της ακρίβειας που παρέχουν.

Αναφορικά με τα δεδομένα αυτά υπάρχουν αρκετά οφέλη χρήσης τους σε διάφορους τομείς, ωστόσο θα δούμε ότι προκύπτουν και αρκετές προκλήσεις. Προκλήσεις που σίγουρα θα πρέπει να ξεπεράσουμε για να μπορέσουμε να τα αξιοποιήσουμε.

1. Στα πλεονεκτήματα χρήσης πραγματικών δεδομένων υγείας αρχικά απευθυνόμαστε σε υπαρκτά δεδομένα [43]. Αναφερόμαστε σε μία ακριβή προσέγγιση δεδομένων για μια ρεαλιστική οπτική της έρευνας. Τα αποτελέσματά που θα παίρναμε θα ήταν πολύ ακριβή καλύπτοντας ένα πραγματικό σενάριο. Το γεγονός αυτό θα μας εξάλειφε το χρόνο και το κόστος που θα απαιτούσε η δημιουργία ενός θεωρητικά ρεαλιστικού σεναρίου και η παραγωγή νέων δεδομένων. Επίσης αναφορικά με τα δεδομένα αυτά, έχουμε υπό συνθήκες μία καλή ποιοτική μορφή δεδομένων υγείας που σίγουρα λειτουργεί θετικά στις ανάγκες της έρευνας.

2. Ωστόσο αρκετά σημαντικά ηθικά αλλά και νομικά ζητήματα προκύπτουν όταν επεξεργαζόμαστε πραγματικά δεδομένα και πιο συγκεκριμένα ειδικές περιπτώσεις όπως αυτά της υγείας. Οι προκλήσεις αυτές αφορούν την ευαισθησία των δεδομένων αυτών και το πώς μπορούν να εκθέσουν πραγματικά πρόσωπα [44]. Η ανάγκη προστασίας τους από πλευράς ιδιωτικότητας είναι πολύ μεγάλη και προστατεύεται, όπως προαναφέρθηκε, από τον GDPR. Ένα σημαντικό ακόμα μειονέκτημα των δεδομένων αυτών είναι οι ασφαλείς τρόποι απόκτησής τους. Το να εξασφαλίσουμε τέτοια δεδομένα είναι ένα αρκετά δύσκολο

έργο καθώς θα πρέπει να μας τα παραχωρήσουν οργανισμοί ή μονάδες υγείας, παρουσιάζοντας από πλευράς μας έναν ορθό και ασφαλή τρόπο αξιοποίησής τους καθώς και το λόγο χρήσης τους.

4.1.2 Υπάρχοντα Μη Πραγματικά Δεδομένα Υγείας

Τα δεδομένα αυτά ή αλλιώς συνθετικά δεδομένα αφορούν δεδομένα που δεν συνδέονται με πραγματικούς ασθενείς και κατά συνέπεια δεν θίγονται ανθρώπινα δικαιώματα όπως η ιδιωτικότητα φυσικών προσώπων [21]. Είναι μια εναλλακτική λύση συλλογής δεδομένων υγείας χωρίς να περιορίζουν τη διαδικασία της έρευνας από πλευράς καταπάτησης προσωπικών δεδομένων, περιέχοντας πληροφορίες υγείας όμοιες με αυτές των πραγματικών δεδομένων υγείας, αγγίζοντας μία πολύ καλή προσέγγιση σε κάποιες περιπτώσεις. Πρόκειται για εγγραφές δεδομένων υγείας που χρησιμοποιήθηκαν σε παλαιότερες έρευνες ή δημιουργήθηκαν για τέτοιους σκοπούς και είναι διαθέσιμα προς χρήση. Κάποια από τα δεδομένα αυτά διατίθενται ελεύθερα στο διαδίκτυο και άλλα προσφέρονται από τρίτους ανάλογα τις ανάγκες. Σημαντικό ρόλο παίζει η ποιότητα των δεδομένων αυτών καθώς και η ευκολία εύρεσής τους.

1. Τα δεδομένα αυτά παρουσιάζουν αρκετά οφέλη καθώς δεν εγείρουν ζητήματα προσβολής ιδιωτικότητας και προστασίας τους κατά την επεξεργασία, την αποθήκευση και την ασφαλή μεταφορά τους. Ως βάση, βρίσκοντας ποιοτικά δεδομένα υγείας τέτοιου είδους, μιλάμε για δεδομένα πολύ καλά δομημένα, ρεαλιστικά και μία πολύ καλή προσομοίωση των πραγματικών δεδομένων υγείας που θα συνδράμει θετικά στην έρευνά μας χωρίς την ανάγκη επιπλέον εργασίας και χρόνου που θα απαιτούσε η παραγωγή νέων δεδομένων [45].

2. Πέρα από τα οφέλη που μας δίνει η συγκεκριμένη μέθοδος, συναντάμε και μερικές προκλήσεις κατά τη χρήση της. Μία από αυτές είναι ο εντοπισμός τέτοιων δεδομένων και η δυνατότητα απόκτησής τους. Όπως αναφέραμε ήδη κάποια από τα δεδομένα αυτά διατίθενται ελεύθερα, ωστόσο κάποια άλλα θα πρέπει να απευθυνθούμε σε τρίτους ώστε να τα αποκτήσουμε οι οποίοι θα πρέπει να εγκρίνουν τους τρόπους που θα τα αξιοποιήσουμε και να μας τα παραχωρήσουν. Ένα άλλο σημαντικό ζήτημα που πρέπει να αναφερθεί είναι η ποιότητα των δεδομένων αυτών και κατά πόσο θα συνεισφέρει σε μία καλή έρευνα ή θα οδηγήσει σε δυσκολίες.

4.1.3 Δημιουργία Ρεαλιστικών Δεδομένων Υγείας

Μία ακόμα επιλογή είναι η δημιουργία ρεαλιστικών μη υπαρκτών δεδομένων υγείας. Πρόκειται για σχετικά νέα τεχνολογία παραγωγής ψευδο-τυχαίων δεδομένων που μπορεί να αξιοποιηθεί ξεπερνώντας δυσκολίες των προηγούμενων μεθόδων. Η λύση αυτή είναι σίγουρα πιο χρονοβόρα από αυτή των ήδη υπαρχόντων δεδομένων, καθώς πρέπει να δημιουργήσουμε εμείς τα δεδομένα αυτά. Ωστόσο υπάρχουν αρκετά εργαλεία διαθέσιμα στο διαδίκτυο παραγωγής δεδομένων τέτοιου τύπου που μπορούμε να προσαρμόσουμε τις ανάγκες μας κατάλληλα προσομοιάζοντας έτσι σε ικανοποιητικό βαθμό τα υπάρχοντα δεδομένα υγείας [46] [47].

1. Πλεονεκτήματα της δημιουργίας ρεαλιστικών δεδομένων υγείας είναι ότι σε αντίθεση με τα πραγματικά δεδομένα υγείας δεν εμπίπτουμε σε κανέναν περιορισμό καταπάτησης ανθρωπίνων δικαιωμάτων και ιδιωτικότητας καθώς μιλάμε για δεδομένα μη υπαρκτά τα οποία μάλιστα δημιουργούμε εμείς. Επίσης δεν χρειάζεται να απευθυνθούμε σε κανένα οργανισμό ή κοινότητα για να τα αποκτήσουμε και δεν επιφέρουν καμία ανάγκη προστασίας τους κατά την αποθήκευση και τη χρήση τους. Τέλος ένα σημαντικό κομμάτι χρήσης τους είναι το ότι μπορούμε να προσαρμόσουμε όπως εμείς επιθυμούμε τα δεδομένα αυτά, κατάλληλα διαμορφωμένα για τις ανάγκες της έρευνάς μας και γενικότερα υπάρχει μία μεγάλη ελευθερία επιλογών [21].

2. Στα μειονεκτήματα τώρα χρήσης τους έχουμε ότι δεν προσομοιάζουν σε θέματα ακρίβειας τα πραγματικά δεδομένα υγείας. Επίσης σε αντίθεση με τα ήδη υπάρχοντα δεδομένα υγείας, η παραγωγή τέτοιων δεδομένων αποτελεί πρόκληση. Αρχικά θα πρέπει να είμαστε σε θέση να παράγουμε τέτοια δεδομένα καθώς απαιτεί κάποια εκπαίδευση χρήσης των κατάλληλων εργαλείων καθώς και την ικανότητα να παράξουμε ρεαλιστικά και ποιοτικά δεδομένα υγείας που θα συνεισφέρουν θετικά στο έργο μας. Μία χαμηλής ποιότητας δημιουργία δεδομένων υγείας θα έχει και τα αντίστοιχα αποτελέσματα επηρεάζοντας αρνητικά την έρευνά μας. Η παραγωγή δεδομένων εξ' αρχής αποτελεί και μια χρονοβόρο διαδικασία σε σχέση με της άλλες μας επιλογές. Τέλος η επιλογή του κατάλληλου εργαλείου μπορεί και αυτή να αποτελέσει πρόκληση όσον αφορά την ευκολία χρήσης του, τα επιπρόσθετα χαρακτηριστικά που διαθέτει και τον παράγοντα ψευδό-τυχαιότητας που μας προσφέρει.

4.2 Συλλογή Δεδομένων

Στη παραπάνω ενότητα παρουσιάσαμε τρείς διαφορετικούς τρόπους επιλογής μεθόδου που μπορούμε να χρησιμοποιήσουμε για τη συλλογή των δεδομένων υγείας που εξυπηρετούν την έρευνά μας με όλα τα οφέλη καθώς και τις προκλήσεις που θα συναντήσουμε ανάλογα την επιλογή μας.

1. Τη χρήση πραγματικών δεδομένων υγείας που παρέχουν μία πολύ ακριβή παρουσίαση, ωστόσο έρχονται με αρκετούς περιορισμούς λόγω της ευαισθησίας τους, τους κανονισμούς που προκύπτουν κατά τη χρήση τους και τις δυσκολίες απόκτησής τους.

2. Τη χρήση υπαρκτών μη πραγματικών δεδομένων υγείας, που αποτελούν μία πολύ πειστική προσομοίωση των πραγματικών δεδομένων υγείας σε αρκετές περιπτώσεις, ωστόσο περιορίζονται σε ακρίβεια καθώς και τους τρόπους εύρεσής τους.

3. Τη δημιουργία ρεαλιστικών δεδομένων υγείας τα οποία λύνουν ένα μεγάλο πρόβλημα, αυτό του περιορισμού επεξεργασίας και αποθήκευσης πραγματικών δεδομένων υγείας καθώς και την ευκολία και την ελευθερία δημιουργίας δεδομένων ακριβώς στις ανάγκες μίας συγκεκριμένης έρευνας, παρ' όλα αυτά απαιτούν χρόνο και εκπαίδευση για τη δημιουργίας τους και αντιμετωπίζουν προβλήματα ακρίβειας ανάλογα πάντα τους σκοπούς.

4.2.1 Επιλογή Μεθόδου

Καταλήγοντας στη μέθοδο που θα εξυπηρετήσει τους σκοπούς και τις ανάγκες τις έρευνάς μας επιλέγουμε αυτή της δημιουργίας ρεαλιστικών δεδομένων υγείας. Ο μεγάλος χρόνος δημιουργίας τους υπερτερεί της ανάγκης συλλογής πραγματικών ή έγκυρων δεδομένων υγείας σε σχέση με τις άλλες μεθόδους. Αναλύοντας και τις τρεις επιλογές μας καταλήξαμε ότι η χρήση πραγματικών δεδομένων υγείας που θα πρόσφερε μεγάλη ακρίβεια θα απαιτούσε αρκετό χρόνο επίσης εύρεσης ενός ασφαλούς τρόπου συλλογή τους, επεξεργασίας τους και αποθήκευσή τους χωρίς να θίξουμε θέματα ιδιωτικότητας και προσωπικών δεδομένων, έχοντας πάντα τον κίνδυνο διαρροής τους από λαθεμένη χρήση που θα επέφερε σοβαρές συνέπειες. Η απόκτησή τους ακόμη θα ήταν μία πρόκληση διότι δεν είναι εύκολα προσβάσιμα. Επίσης η ακρίβεια δεν είναι κάτι που μας απασχολεί στη παρούσα έρευνα διότι ξεπερνά τα πλαίσια του έργου μας.

Η χρήση υπαρχόντων μη πραγματικών δεδομένων υγείας θα μπορούσε να είναι μία καλή επιλογή καλύπτοντας τους κινδύνους διαρροής πραγματικών δεδομένων και χωρίς να χρειάζεται να

δημιουργηθούν νέα δεδομένα υγείας, δεδομένα τα οποία θα μπορούσαν να αντιπροσωπεύσουν σε αρκετά κοντινό στάδιο τα πραγματικά και θα απαιτούσαν σημαντικά λιγότερο χρόνο επεξεργασίας, ωστόσο επιλέξαμε τη παραγωγή νέων δεδομένων υγείας για να τα προσαρμόσουμε στις ανάγκες της έρευνάς μας. Παρά την απαίτηση χρόνου επεξεργασίας και εύρεσης τρόπων δημιουργία τους μας δίνει τη δυνατότητα να δημιουργήσουμε ένα δικό μας σενάριο με κατάλληλα διαμορφωμένα γνωρίσματα, όλα αυτά προσαρμοσμένα στις δικές μας ανάγκες.

4.2.2 Ποιότητα Και Μορφή Δεδομένων

Επιλέγοντας τη δημιουργία νέων δεδομένων ως μέθοδο συλλογής η πρόκληση είναι μεγαλύτερη σε μία προσπάθεια παραγωγής ρεαλιστικών δεδομένων υγείας. Ένα από τα σημαντικά ζητήματα στο στάδιο αυτό είναι ότι τα δεδομένα που χρειάζεται να παράγουμε πρέπει να αντικατοπτρίζουν σε μεγάλο βαθμό τα πραγματικά δεδομένα υγείας περιέχοντας όλες της πληροφορίες που θα παρείχαν τα δεύτερα. Η ποιότητά τους λοιπόν παίζει καθοριστικό ρόλο και πρέπει να είναι κατάλληλα διαμορφωμένα, ώστε να εξυπηρετούν πλήρως τον σκοπό τους [48]. Ελλιπή δεδομένα υγείας ή δεδομένα τα οποία θα είχαν περιττές πληροφορίες θα χαρακτηρίζονταν ως χαμηλής ποιότητας και δεν θα αποτελούσαν καλή προσέγγιση του στόχου μας. Γίνεται σαφές πως τα δεδομένα που θα δημιουργηθούν λοιπόν πρέπει να προσομοιάζουν πλήρως αυτά που θα λαμβάναμε από ένα νοσοκομείο ή κάποιο ίδρυμα υγείας. Το επόμενο κομμάτι που θα πρέπει να εξετάσουμε είναι ο όγκος των δεδομένων αυτών. Σε μία μελέτη όπως αυτή, εξετάζοντας τεχνικές ψευδωνυμοποίησης και ανωνυμοποίησης, το πλήθος των εγγραφών θα πρέπει να είναι ικανοποιητικά μεγάλο ώστε να καλύπτει τις ανάγκες μας. Όπως έχουμε ήδη δει σε παραπάνω κεφάλαιο, οι διαφορετικές προσεγγίσεις ειδικά της ανωνυμοποίησης απαιτούν ένα μεγάλο ποσοστό δεδομένων διασφαλίζοντας έτσι ικανοποιητικά την ανάλυση τους καθώς τον παράγοντα της τυχαιοποίησης. Επίσης για να μπορέσουμε να εξετάσουμε τις προκλήσεις που συναντάμε αναλύοντας ευαίσθητα δεδομένα όπως αυτά της υγείας και ειδικά υπό τη σκοπιά της ανωνυμοποίησης τους όπως και μιας θεωρητικά ιατρικής μελέτης έχοντας ως γνώμονα τις δυσκολίες που παρουσιάζονται και τις απαιτήσεις που ορίζει ο GDPR, δεδομένα όπως αυτά των ηλικιακών ομάδων ή χαρακτηριστικά γνωρίσματα όπως τα ονόματα είναι σημαντικό να υπάρχουν.

Περνώντας τώρα στη μορφή των δεδομένων αυτών, εξετάσαμε την ανάγκη της έρευνάς μας που είναι η προσπάθεια ψευδωνυμοποίησης και ανωνυμοποίησής τους ώστε να περιοριστεί κατά το μέγιστο δυνατό η πιθανότητα αναγνώρισης κάποιου προσώπου μέσα σε αυτά. Όπως έχουμε ήδη

αναφέρει, η παραγωγή τους θα πρέπει να προσομοιάζει αυτή των πραγματικών δεδομένων υγείας με όλα τα χαρακτηριστικά που θα λαμβάναμε από κάποιον φορέα υγείας σε άλλη περίπτωση. Στη περίπτωση αυτή λοιπόν θα συμπεριληφθούν δημογραφικά στοιχεία όπως ονοματεπώνυμα ασθενών, ηλικία (η οποία προκύπτει αυτόματα από την ημερομηνία γέννησής του ασθενούς), ο αριθμός μητρώου κοινωνικής ασφάλισής (ΑΜΚΑ), το φύλο καθώς και ο ταχυδρομικός κώδικας κατοικίας τους όπως αυτά συλλέγονται κατά τη προσκόμισή τους σε ένα νοσοκομείο. Στη συνέχεια ο κάθε ασθενής κατά την πρώτη επίσκεψή του σε κάποιο νοσοκομείο λαμβάνει και έναν μοναδικό αριθμό ασθενούς που τον ακολουθεί για πάντα κατά τις επισκέψεις του στο νοσοκομείο ως χαρακτηριστικό γνώρισμα. Συνεχίζοντας, μιας και αναφερόμαστε σε νοσοκομειακούς ασθενείς αποφασίσαμε να συμπεριλάβουμε μία σειρά από πιθανές ασθένειες χωρισμένες σε δύο κατηγορίες, τις χρόνιες και τις προσωρινές, την πιθανότητα εγκυμοσύνης, κάποιου χειρουργείου ή συμπεριφοράς τα οποία, υπό συνθήκες, αποτελούν ευαίσθητες πληροφορίες που πρέπει να προστατευτούν. Τέλος σε περίπτωση εισαγωγής ενός ασθενούς στο νοσοκομείο θα πρέπει να καταγράφεται η ημερομηνία εισαγωγής του, η ημερομηνία εξιτηρίου καθώς και οι μέρες νοσηλείας του.

Έχοντας ως στόχο τη ρεαλιστικότητα των δεδομένων που θέλουμε να επιτύχουμε, υπάρχει ακόμη ένας παράγοντας που κρίνεται καθοριστικός, αυτός της τυχαιότητας. Θα πρέπει να δοθεί μεγάλο βάρος ώστε τα δεδομένα μας να μη παρουσιάζονται εντελώς τυχαία κάτι που θα υστερούσε σε ρεαλισμό συγκριτικά με τα πραγματικά δεδομένα υγείας αλλά στη πραγματικότητα να πετύχουμε κάποια ψευδο-τυχαιότητα δίδοντας έτσι διαφορετικό βάρος στα γνωρίσματά μας. Διαφορετικοί ασθενείς αντιδρούν διαφορετικά σε κάποια ασθένεια, όπως επίσης η ηλικιακή ομάδα στην οποία ανήκουν αλληλοεπιδρά διαφορετικά. Τέλος δεν πάσχουν όλοι οι ασθενείς που επισκέπτονται ένα νοσοκομείο ή μία μονάδα υγείας από κάποια ασθένεια κάτι που θα πρέπει να επισημανθεί στη προσπάθεια ενός πολύ ρεαλιστικού αποτελέσματος.

4.2.3 Διαθέσιμα Εργαλεία Παραγωγής Δεδομένων

Επιλέγοντας τη δημιουργία νέων δεδομένων υγείας ως μέθοδο συλλογής θα πρέπει να εξετάσουμε και εργαλεία με τα οποία μπορούμε να παράγουμε τα δεδομένα αυτά και εξυπηρετούν τους σκοπούς μας. Υπάρχουν δύο διαθέσιμοι τρόποι υλοποίησης, αυτός της χειροκίνητης διαδικασίας και η αξιοποίηση εργαλείων παραγωγής τυχαίων δεδομένων [49].

Ο πρώτος τρόπος μπορεί να υλοποιηθεί με ένα εργαλείο όπως το Excel της Microsoft και τη χρήση πληθώρας συναρτήσεων που προσφέρει, ωστόσο θα καθιστούσε πάρα πολύ χρονοβόρα τη

διαδικασία αυτή δημιουργώντας χειροκίνητα όλα τα πεδία και τα γνωρίσματα που απαιτούνται, καθώς επίσης θα επηρέαζε σε μεγάλο βαθμό τον παράγοντα τυχαιοποίησης σε μία προσπάθεια να αποδώσουμε ρεαλιστικά δεδομένα.

Για τους παραπάνω λόγους οδηγούμαστε στη δεύτερη επιλογή μέσα από δωρεάν εργαλεία, διαδικτυακά κυρίως που προσφέρουν παραγωγή τυχαίων δεδομένων ορίζοντας συγκεκριμένα γνωρίσματα και προσαρμόζοντάς τα στις ανάγκες του έργου που θέλουμε να εξετάσουμε. Παρακάτω παρουσιάζουμε κάποια από τα εργαλεία αυτά εξετάζοντας τις λειτουργίες του κάθε ενός σε μία προσπάθεια να επιλέξουμε το καταλληλότερο για τη διαδικασία που επιθυμούμε.

1. **GenerateData** [50]: Πρόκειται για μία πολύ απλή πλατφόρμα παραγωγής τυχαίων δεδομένων προσφέροντας αντίστοιχα αρκετά βασικούς τύπους δεδομένων. Παρέχει ένα αρκετά εύκολο προς χρήση περιβάλλον δίνοντας την επιλογή προεπισκόπησης των δεδομένων που έχουμε παράγει. Τέλος μας δίνει τη δυνατότητα εξαγωγής των δεδομένων αυτών σε μία πληθώρα τύπων αρχείου. Παρά την ευκολία του και τη φιλικότητα προς τον χρήστη, το εργαλείο αυτό δεν καλύπτει τις ανάγκες μας λόγω των περιορισμένων χαρακτηριστικών που προσφέρει.

2. **OnlineDataGenerator** [51]: Αποτελεί ένα αρκετά πιο σύνθετο εργαλείο προσφέροντας αρκετούς συνδυασμούς τύπων δεδομένων και πληροφοριών που μπορεί να εξάγει. Αντίστοιχα με το προηγούμενο εργαλείο που εξετάσαμε προσφέρει αρκετούς τύπους αρχείων κατά την εξαγωγή. Επίσης δίνει τη δυνατότητα εξαγωγής ενός μεγάλου όγκου εγγραφών και αποτελεί μία καλή επιλογή για τις ανάγκες μας, όντας ένα δωρεάν διαθέσιμο εργαλείο. Εξετάζοντάς το αναλυτικότερα μπορούμε να δούμε ότι μας δίνει αρκετές επιλογές γνωρισμάτων, ωστόσο θα πρέπει να το ελέγξουμε αρκετά πιο ενδελεχώς για να δούμε αν καλύπτει πλήρως τις ανάγκες μας.

3. **Mockaroo** [52]: Ένα ακόμα αρκετά εξειδικευμένο εργαλείο παραγωγής τυχαίων δεδομένων το οποίο δείχνει να προσφέρει αρκετούς συνδυασμούς επιλογών και μεγάλη πληθώρα γνωρισμάτων, υπόσχοντας πολύ ρεαλιστικά αποτελέσματα. Παρέχοντας αρκετούς τύπους δεδομένων και πιο συγκεκριμένα μία ειδική κατηγορία δεδομένων υγείας το καθιστά αρκετά χρήσιμο για τους σκοπούς μας. Δίνει επίσης τη δυνατότητα εξαγωγής διαφόρων τύπων αρχείων. Το συγκεκριμένο εργαλείο είναι αρκετά πιο περίπλοκο από τα υπόλοιπα ως προς την εκμάθησή του, ωστόσο αυτό οφείλεται στη πληθώρα επιλογών που προσφέρει και μπορεί να παράξει δεδομένα πλήρως

προσαρμοσμένα στις ανάγκες μας. Στην δωρεάν έκδοσή του παρέχει περιορισμένο όγκο δεδομένων προς εξαγωγή.

4.2.3 Επιλογή Εργαλείου

Εξετάζοντας ξεχωριστά τα διαθέσιμα εργαλεία παραγωγής τυχαίων δεδομένων και τις λειτουργίες που μας προσφέρουν καθώς και την ευκολία χρήσης τους καταλήξαμε στην επιλογή του Mockaroo το οποίο μας δίνει μία πληθώρα επιλογών και προσαρμογής των δεδομένων ακριβώς στις ανάγκες μας. Για τις ανάγκες αυτές αρκούμαστε στην δωρεάν έκδοσή του, χωρίς να χρειαστεί να απευθυνθούμε στις επί πληρωμή. Όπως αναφέραμε πρόκειται για ένα αρκετά περίπλοκο εργαλεία συγκριτικά με τις υπόλοιπες επιλογές μας, ωστόσο αυτό αποτελεί και προτέρημά του δίνοντας μας την ευχέρεια μίας πολύ πιο ρεαλιστικής αναπαράστασης. Το εργαλείο αυτό έχει αναπτυχθεί σε γλώσσα προγραμματισμού ruby, μία γλώσσα αρκετά απλή στην εκμάθηση για τους σκοπούς μας παρέχοντας αρκετές επιλογές παραμετροποίησης του κάθε πεδίου σε αυτή. Επίσης συνοδεύεται από έναν αρκετά πλήρη οδηγό για τις ενέργειες και τα πρόσθετα που προσφέρει. Αξιοσημείωτο είναι και το γεγονός ότι προσφέρει ειδική κατηγορία που αφορά δεδομένα υγείας, ωστόσο επιλέξαμε να δημιουργήσουμε δικά μας γνωρίσματα, κατάλληλα τροποποιημένα προσαρμόζοντάς τα στο σενάριο μας.

Κεφάλαιο 5
Μελέτη Περίπτωσης

Το παρόν κεφάλαιο αποτελεί τη μελέτη περίπτωσης που θα εξετάσουμε κατά την διεξαγωγή της έρευνάς μας. Αφορά μία "θεωρητική" επιδημία που έχει εμφανιστεί και την ανάγκη που έχει προκύψει, συλλέγοντας δεδομένα από κλινικές σε μία παγκόσμια κλίμακα για τη πλήρη κατανόησή της κατάστασης και την αντιμετώπισής της. Τα δεδομένα αυτά συλλέγονται για να αποσταλούν σε Ερευνητικό Κέντρο προς μελέτη, διασφαλίζοντας τα προσωπικά δεδομένα και την ιδιωτικότητα των φυσικών προσώπων ως υποκείμενα της έρευνας και ακολουθώντας πιστά τους νόμους και τους κανονισμούς που ορίζονται από τον GDPR.

5.1 Μελέτη

Στα πλαίσια της επιδημίας HRS που δείχνει να μαστίζει και την Ελλάδα τους τελευταίους 14 μήνες και μετά τη ξαφνική εμφάνιση του Covid-19 που δοκίμασε σε μεγάλο βαθμό την ετοιμότητα των νοσοκομείων, παρουσιάζεται η επιτακτική ανάγκη να ληφθούν τα απαραίτητα μέτρα για τη προστασία των πολιτών.

Οι μέχρι τώρα ενδείξεις με βάση τα δεδομένα του Παγκόσμιου Οργανισμού Υγείας δείχνουν πως η νόσος HRS δεν επηρεάζει άμεσα τη υγεία των φορέων της νόσου καθώς έχει ελεγχθεί ένα αρκετά μεγάλο ποσοστό ασθενών παγκοσμίως, ασθενών διαφόρων παθήσεων και ηλικιακών ομάδων. Έχει καταστεί σαφές λοιπόν ότι δεν επηρεάζει τη πορεία της κατάστασης της σωματικής υγείας τους αλλά ούτε και τις ημέρες νοσηλείας τους κάτι που σαφέστατα δείχνει καθησυχαστικό. Ωστόσο από την επιδημία αυτή, με τα δεδομένα που έχουμε έως τώρα δείχνει να επηρεάζεται η ψυχική υγεία θετικών ασθενών στην HRS νόσο υπό συγκεκριμένες συνθήκες και καταστάσεις της ζωής τους. Για παράδειγμα, γυναίκες σε εγκυμοσύνη, που πάσχουν από σακχαρώδη διαβήτη και βρέθηκαν θετικές στην HRS δείχνουν να παρουσιάζουν κατάθλιψη, ασθενείς θετικοί στην HRS συγκεκριμένης ηλικιακής ομάδας δείχνουν να παρουσιάζουν επιθετική συμπεριφορά.

5.1.1 Ανάγκη Αντιμετώπιση Της

Με βάση τα παραπάνω δείγματα αιτήθηκε από τον Π.Ο.Υ προς όλες τις Χώρες να καταγραφούν για το έτος 2022 τα δεδομένα ασθενών των νοσοκομείων οι οποίοι έχουν εξεταστεί στη συγκεκριμένη νόσο κατά την επίσκεψη ή τη νοσηλεία τους. Επίσης η εγκύκλιος αναφέρει ρητά πως πέραν τη καταγραφής των πιθανών ασθενειών των εξεταζόμενων θα πρέπει να συμπεριληφθεί και η κοινωνική συμπεριφορά αυτών για της ανάγκες της έρευνας. Τα δεδομένα αυτά αφού συλλεχθούν από το εκάστοτε νοσοκομείο θα αποσταλούν σε ερευνητικά κέντρα για να προχωρήσει η ανάλυση των έως τώρα ευρημάτων, καταλήγοντας σε μία πλήρη εικόνα της κατάστασης.

Με τη σειρά της η διοίκηση του νοσοκομείου μας, ζήτησε από τις πέντε κλινικές του(Χειρουργική, Γυναικολογική, Παθολογική, Καρδιολογική, Πνευμονολογική) να αποστείλουν δεδομένα σχετικά με τους ασθενείς που εξετάστηκαν στη νόσο HRS για το συγκεκριμένο έτος και πέραν των δημογραφικών τους στοιχείων να συμπεριληφθούν και πιθανές ασθένειες χρόνιες ή προσωρινές, πιθανές εγκυμοσύνες καθώς και ενδείξεις αλλαγής της συμπεριφοράς τους. Τα δεδομένα αυτά θα συγκεντρωθούν σε έναν ενιαίο πίνακα και θα αποσταλούν στο ερευνητικό κέντρο του Πανεπιστημίου Αθηνών.

5.1.2 Αναγκαιότητα Προστασίας Των Δεδομένων

Τα δεδομένα αυτά αποτελούν ευαίσθητα προσωπικά δεδομένα υγείας ασθενών και απαιτούν ιδιαίτερη μεταχείριση. Για την ανάγκη αυτή ζητήθηκε από τη διοίκηση η βοήθεια της Πληροφορικής καθώς και του Υπεύθυνου Προστασίας Προσωπικών Δεδομένων(DPO) του νοσοκομείου για την ασφαλή διανομή των δεδομένων αυτών από τις κλινικές προς το τμήμα ερευνών του νοσοκομείο και στην συνέχεια στο Ερευνητικό Κέντρο του Πανεπιστημίου από το οποίο θα αναλυθούν τα δεδομένα αυτά.

Προκειμένου να αποφευχθεί η διαρροή ευαίσθητων πληροφοριών υγείας που μπορεί να οδηγήσει σε σοβαρές συνέπειες και να προστατευτεί η ιδιωτικότητα και η ασφάλεια του απορρήτου των ασθενών, η πρόταση της Διεύθυνσης Πληροφορικής ήταν να ψευδωνυμοποιηθούν αρχικά τα δεδομένα και πιο συγκεκριμένα τα αναγνωριστικά τους πεδία από κάθε κλινική πριν αποσταλούν στο τμήμα ερευνών. Λαμβάνοντας υπόψιν ότι βάσει δικαιωμάτων στην εφαρμογή που κρατάει τα δεδομένα των ασθενών, οι χρήστες έχουν πρόσβαση μόνο στα δεδομένα που αφορούν τη κλινική τους και δεν μπορούν να δουν τους ασθενείς των άλλων κλινικών, ο προϊστάμενος κάθε

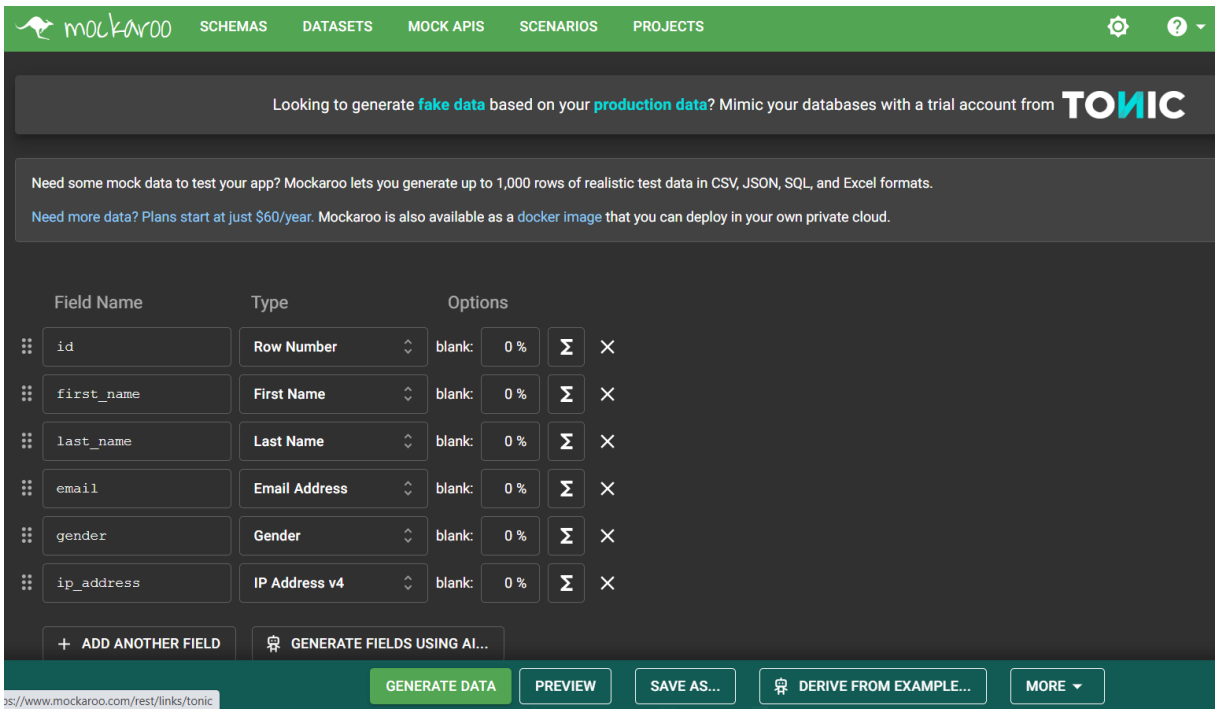
κλινικής φέρει ευθύνη να συγκεντρώσει τα δεδομένα των ασθενών που ζητήθηκαν για τη δική του κλινική σε ένα αρχείο Excel χωρίς το φόβο ότι θα δουν άλλες κλινικές τα δεδομένα της δικής του και να τα αποστείλει στο τμήμα ερευνών. Δικαίωμα πρόσβασης σε όλες τις κλινικές έχουν μόνο οι διαχειριστές της εφαρμογής. Τέλος το τμήμα ερευνών του νοσοκομείου φέρει την ευθύνη της δημιουργίας του τελικού πίνακα με τα ψευδωνυμοποιημένα πλέον δεδομένα και να τα χειριστεί αναλόγως πριν τα στείλει το Ερευνητικό Κέντρο.

5.2 Παραγωγή Των Δεδομένων

Στη παρούσα ενότητα θα εξετάσουμε αναλυτικά τα δεδομένα που δημιουργήσαμε για τις ανάγκες της έρευνάς μας. Από το προηγούμενο κεφάλαιο έχουμε επιλέξει τη παραγωγή νέων δεδομένων υγείας ως μέθοδο συλλογής. Στο σημείο αυτό θα πρέπει να αναλύσουμε και να χωρίσουμε τα δεδομένα μας ανάλογα τα χαρακτηριστικά τους, τις πληροφορίες που μας δίνουν, καθώς και την ευαισθησία τους. Η προσπάθεια προσέγγισής τους έγινε με μία απόπειρα ρεαλιστικής απεικόνισης μίας καρτέλας ασθενούς σε ένα Ελληνικό νοσοκομείο.

5.2.1 Διαμόρφωση Των Δεδομένων

Ως εργαλείο παραγωγής των δεδομένων μας επιλέξαμε το Mockaroo, όπως αναφέραμε και στο προηγούμενο κεφάλαιο κρίνοντας το έτσι, μέσα από μία σειρά επιλογών, ως το αποτελεσματικότερο για την έρευνά μας.



Εικόνα 5.1: Mockaroo – Random Data Generator Tool [52]

Στη συνέχεια έπρεπε να αποφασίσουμε για τις πληροφορίες που θα συμπεριληφθούν στα δεδομένα αυτά δημιουργώντας ένα πίνακα με τον οποίο θα ξεκινήσουμε το έργο μας. Για τις ανάγκες της έρευνας αποφασίσαμε να δημιουργήσουμε 100.000 μοναδικές εγγραφές οι οποίες αποτελούν ξεχωριστούς ασθενείς που εξετάστηκαν στην HRS νόσο στο νοσοκομείο μας κατά το παραπάνω αναγραφόμενο έτος. Στους ασθενείς αυτούς έχουν συμπεριληφθεί διάφορα γνωρίσματα ως πληροφορίες που αφορούν την επίσκεψή τους στο νοσοκομείο και την ενδεχόμενη νοσηλεία τους. Τα γνωρίσματα αυτά χωρίζονται σε αναγνωριστικά, ψευδο-αναγνωριστικά και άλλα αποτελούν ευαίσθητα χαρακτηριστικά. Αφού δημιουργήθηκε ο πίνακας των δεδομένων μας με όλα τα πιθανά σενάρια, στη συνέχεια χωρίστηκε σε πέντε μικρότερους που αφορούν τις κλινικές του νοσοκομείου και ανάλογα τις διαγνώσεις των ασθενών, τη πιθανότητα εγκυμοσύνης ή κάποιου χειρουργείου, μετατέθηκαν οι ασθενείς σε αυτές. Τέλος τα δεδομένα αυτά ακολουθώντας τη διαδικασία του νοσοκομείου για την απόκρυψη των προσωπικών δεδομένων των ασθενών, συλλέχθηκαν σε ένα πίνακα ψευδωνυμοποιημένα πλέον και στάλθηκαν στο τμήμα ερευνών του νοσοκομείου ώστε να τα προετοιμάσει για να τα αποστείλει με τη σειρά του στο Ερευνητικό Κέντρο προς ανάλυση.

5.2.2 Δημογραφικά Στοιχεία

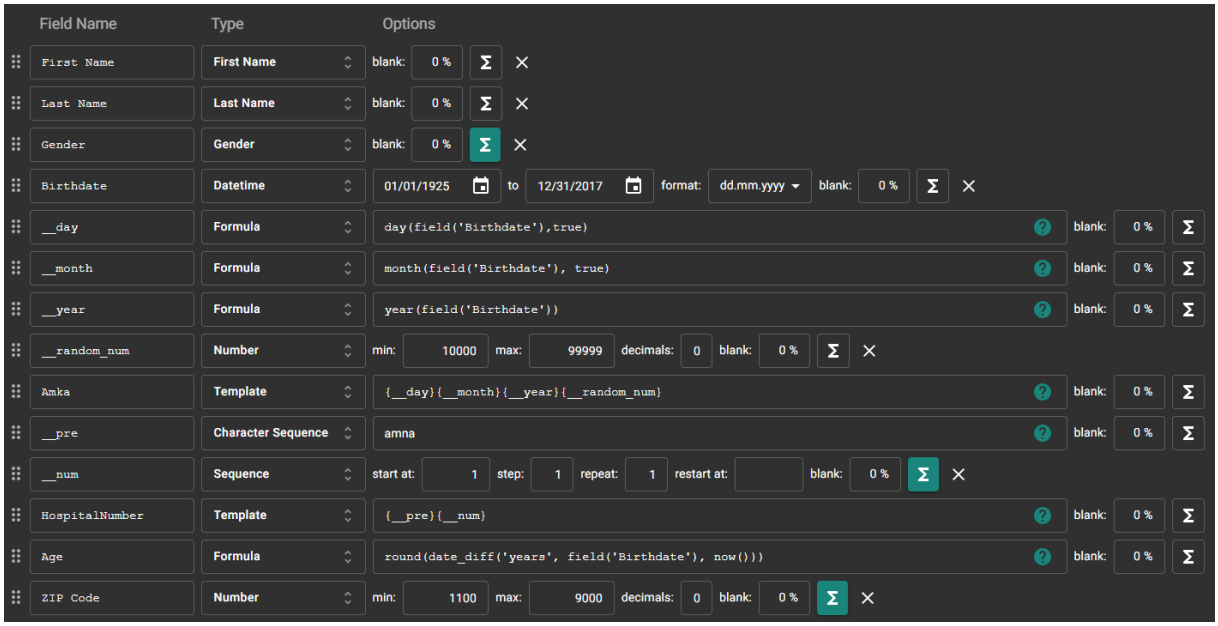
Αναφερόμενοι σε ασθενείς ενός Νοσοκομείου αρχικά τα πρώτα δεδομένα που θα πρέπει να καταγραφούν είναι τα δημογραφικά τους στοιχεία. Σε αυτά έχουν συμπεριληφθεί ως ξεχωριστά γνωρίσματα το Όνομα, το Επώνυμο καθώς και το Φύλο του ασθενούς. Μέσω του εργαλείου Mockaroo δίνεται η δυνατότητα παραγωγής τυχαίων ονομάτων και επωνύμων για κάθε εγγραφή και για τη περίπτωση του φύλου είναι διαμορφωμένο έτσι ώστε ανάλογα αν το όνομα είναι αρσενικού ή θηλυκού γένους, να αποδίδεται αντίστοιχα και το φύλο.

Στη συνέχεια σε μία προσπάθεια να προσδιορίσουμε την περιοχή στην οποία διαμένουν οι ασθενείς, κάτι που θα μπορούσε να δώσει πληροφορίες για τη γεωγραφική εξάπλωση της νόσου, αποφασίσαμε να συμπεριλάβουμε τον ταχυδρομικό κώδικα της κατοικίας τους. Ο Τ.Κ. αποδίδεται μέσα από μία γεννήτρια τυχαίων αριθμών σε ένα εύρος {1000-9999}. Στη συνάρτηση αυτή, καθώς οι Τ.Κ της Αθήνας αποτελούν 5ψήφιους αριθμούς και ξεκινούν πάντα από 1, προστέθηκε και ο αριθμός 1 μπροστά από τη 4ψήφια ακολουθία, καθιστώντας την παραγωγή του Τ.Κ. ως ψευδο-τυχαία.

Συνεχίζοντας, στα δημογραφικά δεδομένα του ασθενούς συμπεριλαμβάνεται η ηλικία και η ημερομηνία γέννησής τους. Για τη ημερομηνία γέννησης, το εργαλείο προσφέρει ειδικά διαμορφωμένη συνάρτηση τυχαίων ημερομηνιών, συμπεριλαμβανομένης της ημέρας, του μήνα και του έτους. Επίσης προσφέρει τη δυνατότητα να ορίσουμε εμείς το εύρος των χρονολογιών που επιθυμούμε. Για τη περίπτωσή μας επιλέξαμε το εύρος {1925-2017}. Η ηλικία του ασθενούς προκύπτει από την διαφορά της ημερομηνίας γέννησης με τη σημερινή ημερομηνία. Αυτό επιτυγχάνεται επίσης μέσω συνάρτησης που προσφέρει το Mockaroo για τον υπολογισμό της διαφοράς μεταξύ δύο αριθμών.

Το επόμενο γνώρισμα που έχουμε είναι ο Αριθμός Μητρώου Κοινωνικής Ασφάλισης(ΑΜΚΑ) του ασθενούς. Πρόκειται για μοναδικό αριθμό που υιοθετεί πλέον από τη γέννησή του πολίτης της Ελλάδας και αποτελείται από την ημερομηνία γέννησής του σε μορφή ακολουθίας αριθμών, έχοντας ως διψήφιο το έτος γέννησης, ακολουθούμενο από πέντε τυχαίους αριθμούς. Για τη παραγωγή του ΑΜΚΑ με το εργαλείο μας, αρχικά δημιουργήσαμε μέσω συνάρτησης του, ακολουθία αριθμών όμοια με την ημερομηνία γέννησης του ασθενούς της μορφής ddmmyyyy. Στη συνέχεια δημιουργήσαμε ακόμα μία γεννήτρια τυχαίων αριθμών σε ένα εύρος {10000-99999} ώστε να αποδώσουμε τους 5 τελευταίους αριθμούς. Τέλος συνενώσαμε τα δύο παραπάνω στοιχεία για να δημιουργήσουμε το πεδίο ΑΜΚΑ. Για την αποφυγή πιθανής συσχέτισης του πεδίου αυτού με κάποιο πραγματικό ΑΜΚΑ κρατήσαμε ως 4ψήφιο το έτος γέννησης εν αντιθέσει των δύο ψηφίων που έχει το πραγματικό.

Τέλος έχουμε τον Μοναδικό Αριθμό Ασθενή. Πρόκειται για έναν επίσης μοναδικό αριθμό που λαμβάνει ο ασθενής κατά τη πρώτη του επίσκεψή στο νοσοκομείο και τον ακολουθεί για πάντα ως αναγνωριστικό στοιχείο και στις επόμενες πιθανές επισκέψεις του. Το γνώρισμα αυτό αποτελείται από τέσσερεις χαρακτήρες και ακολουθείται από οκτώ αριθμούς ξεκινώντας από το 10000001 καθώς ανεβαίνει κατά ένα ως μετρητής σε κάθε επόμενο ασθενή. Στη περίπτωσή μας για να δημιουργήσουμε τον Μοναδικό Αριθμό Ασθενή, αρχικά δημιουργήσαμε μία ακολουθία τεσσάρων χαρακτήρων με το όνομα 'AMNA' σε λατινικούς χαρακτήρες (Αριθμός Μητρώου Νοσηλευόμενου Ασθενή) καθώς και μία ακολουθία αριθμών που αυξάνεται διαδοχικά κατά 1, ξεκινώντας από το 10000001. Τέλος με άλλη συνάρτηση συνδέσαμε τα δύο αυτά στοιχεία και παρουσιάζεται σε μία μορφή 'amna10000001'.



Εικόνα 5.2: Mockaroo – Δημιουργία γνωρισμάτων Δεδομένων Υγείας

5.2.2 Ευαίσθητα Πεδία

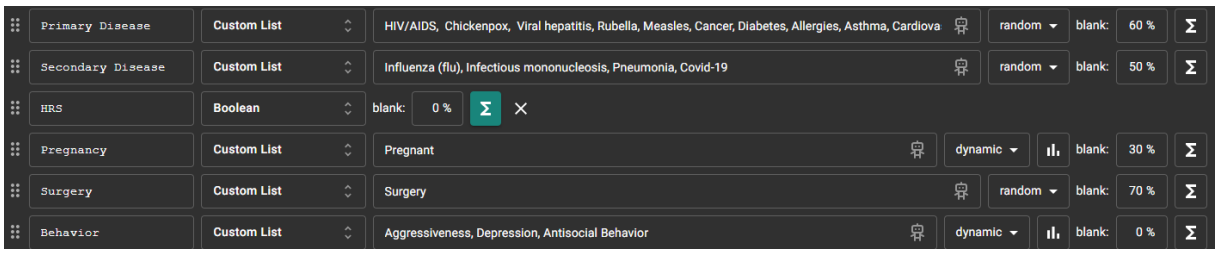
Μετά τα δημογραφικά στοιχεία του ασθενούς έχουμε τα ευαίσθητα γνωρίσματα τα οποία και χρήζουν ιδιαίτερης προσοχής κατά την επεξεργασία, όντας ο κύριος λόγος διασφάλισης του απορρήτου και της ιδιωτικότητας του ασθενούς. Όπως αναφέραμε και σε προηγούμενο κεφάλαιο πρόκειται για άκρως ευαίσθητα προσωπικά δεδομένα ή ειδική κατηγορία προσωπικών δεδομένων όπως έχουν οριστεί και από το Γενικό Κανονισμό Προστασίας Δεδομένων (GDPR).

Αρχικά έχουμε ως Boolean γνώρισμα(αληθές ή ψευδές) την εξέταση του ασθενούς στην νόσο HRS. Το πεδίο αυτό είναι είτε θετικό(HRS+), είτε αρνητικό(HRS-) χαρακτηρίζοντας έτσι το αποτέλεσμα της εξέτασης του ασθενούς.

Στη συνέχεια έχουμε δύο ακόμα γνωρίσματα, ως πιθανές ασθένειες, που αφορούν, η πρώτη κάποια χρόνια ασθένεια από την οποία μπορεί να πάσχει ο ασθενής και η δεύτερη μία προσωρινή νόσο, η οποία πιθανό να είναι και ο λόγος επίσκεψής τους στο Νοσοκομείο. Για τη πρώτη ασθένεια δημιουργήσαμε ένα γνώρισμα ως Primary Disease το οποίο αποτελεί μία λίστα διάφορων ονομάτων που δίνουμε εμείς και αποδίδει τυχαία σε κάθε εγγραφή ένα από τα ονόματα της λίστας. Ως ονόματα ορίσαμε 10 κύριες ασθένειες (HIV/AIDS, Chickenpox, Viral hepatitis, Rubella, Measles, Cancer, Diabetes, Allergies, Asthma, Cardiovascular Disease) και αποδίδεται μία από αυτές σε κάθε ασθενή. Μέσα από το εργαλείο προσφέρεται και δυνατότητα κατά ποσοστό % εμφάνισης κενών πεδίων σε κάποιο γνώρισα. Δεδομένου ότι δεν πάσχουν όλοι οι ασθενείς που επισκέπτονται το Νοσοκομείο από κάποια από τις παραπάνω ασθένειες, αποφασίσαμε στον πίνακά μας το γνώρισμα αυτό να περιέχει κατά 60% κενές τιμές. Για τις προσωρινές νόσους ακολουθήσαμε την ίδια διαδικασία περιέχοντας στη λίστα τέσσερις προσωρινές νόσους(Influenza (flu), Infectious mononucleosis, Pneumonia, Covid-19). Το ποσοστό εμφάνισης κενών πεδίων στο γνώρισμα αυτό βρίσκεται στο 50% κρίνοντας ότι αρκετά περιστατικά προσέρχονται στο Νοσοκομείο για άλλους λόγους πέραν κάποιου από τις παραπάνω προσωρινές νόσους.

Στη συνέχεια έχουμε τη πιθανότητα χειρουργείου από ασθενείς που επισκέφτηκαν τη χειρουργική κλινική. Το ποσοστό κενών πεδίων, κρίνοντας ότι το ένα τρίτο περίπου των ασθενών υπεβλήθη σε χειρουργείο, είναι στο 70%. Το χειρουργείο επίσης επηρεάζει τις ημέρες νοσηλείας του ασθενούς αυξάνοντάς τις από 1 έως 10 ημέρες με συνθήκη που δημιουργήθηκε στο εργαλείο.

Κλείνοντας με τα ευαίσθητα γνωρίσματα έχουμε τη συμπεριφορά τους ασθενούς. Πρόκειται για επικείμενες αλλαγές της συμπεριφοράς θετικών ασθενών στη νόσο HRS υπό συγκεκριμένες συνθήκες ή καταστάσεις της ζωής τους. Αρχικά έχουμε ότι γυναίκες σε εγκυμοσύνη, που πάσχουν από σακχαρώδη διαβήτη καθώς και ασθενείς με καρδιοπάθειες που διαγνώσθηκαν θετικοί στην HRS παρουσιάζουν κατάθλιψη, ασθενείς θετικοί στην HRS ηλικίας άνω των 70 ετών παρουσιάζουν επιθετική συμπεριφορά. Τέλος θετικοί ασθενείς στην HRS αρσενικού γένους συγκεκριμένης ηλικιακής ομάδα {18-40} καρκινοπαθείς καθώς και θετικοί ασθενείς, που πάσχουν από HIV παρουσιάζουν αντικοινωνική συμπεριφορά.



Εικόνα 5.3: Mockaroo – Ευαίσθητα Γνωρίσματα

5.2.3 Μη Εμπιστευτικά Πεδία

Ένα γνώρισμα κατηγοριοποιείται ως μη εμπιστευτικό ή ευαίσθητο στη παρούσα έρευνα το οποίο αφορά της γυναίκες και είναι η πιθανότητα εγκυμοσύνης. Πρόκειται για γυναίκες που εξετάστηκαν για τη πιθανότητα αυτή στη γυναικολογική κλινική. Δημιουργήθηκε χρησιμοποιώντας συνάρτηση λίστας, περιέχοντας ένα μόνο στοιχείο, αυτό της εγκυμοσύνης. Το συγκεκριμένο γνώρισμα δημιουργήθηκε βάσει κανόνων που περιλαμβάνουν το φύλο και την ηλικία. Για να υπάρχει η πιθανότητα εγκυμοσύνης αρχικά θα πρέπει να το φύλο να είναι θηλυκού γένους και η ηλικία να περιορίζεται στο εύρος {18-37}. Επίσης αποδώσαμε και ένα ποσοστό 30% κενών πεδίων κρίνοντας ότι δεν είναι σε εγκυμοσύνη όλες η γυναίκες που επισκέφτηκαν το γυναικολόγο και βρίσκονται σε αυτό το εύρος ηλικιών.

5.2.4 Πρόσθετα πεδία αναφορικά με τη νοσηλεία

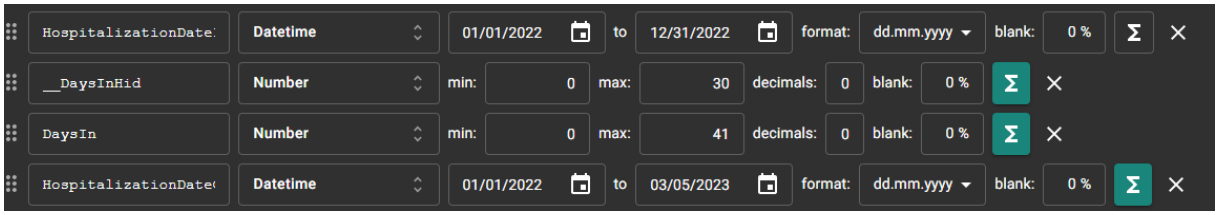
Στη συνέχεια σε μία προσπάθεια να αποδώσουμε ρεαλιστικότητα στα δεδομένα μας, αποφασίσαμε να συμπεριλάβουμε τρία ακόμα γνωρίσματα που αφορούν την ενδεχόμενη νοσηλεία του ασθενούς στο Νοσοκομείο. Τα γνωρίσματα αυτά κατηγοριοποιούνται ως εμπιστευτικά, καθώς μπορούν συνδυαστικά με κάποιο άλλο ευαίσθητο πεδίο να ταυτοποιήσουν κάποιο φυσικό πρόσωπο και αποτελούν πληροφορία που δεν είναι δημόσια διαθέσιμη.

Αρχικά ορίσαμε το πρώτο γνώρισμα ως ημέρα εισαγωγής ή επίσκεψης του ασθενούς στο Νοσοκομείο. Αναφερόμενοι στο έτος 2022, χρησιμοποιήσαμε τη συνάρτηση τυχαίων ημερομηνιών στο εύρος του συγκεκριμένου έτους.

Το επόμενο πεδίο αφορά τις πιθανές ημέρες νοσηλείας του ασθενούς ξεκινώντας από την ημέρα που επισκέφτηκε το Νοσοκομείο και εξαρτάται από κάποιους παράγοντες που αφορούν την ασθένεια ή τα πιθανά χειρουργεία. Όπως αναφέραμε και παραπάνω οι ασθενείς που έχουν

χειρουργηθεί νοσηλεύονται στο Νοσοκομείο σε ένα τυχαίο εύρος {1-10} περισσότερες μέρες από το αναμενόμενο μέσω ειδικής συνάρτησης τυχαίων αριθμών που προσφέρει το εργαλείο. Επίσης οι ημέρες νοσηλείας επηρεάζονται και από τις προσωρινές καθώς και τις χρόνιες ασθένειες. Έχουν διαμορφωθεί μέσω του εργαλείου ειδικές συνθήκες ως κανόνες στο γνώρισμα της ημέρας επίσκεψης, όπου ανάλογα την ασθένεια και την ηλικιακή ομάδα που ανήκει ο ασθενής, νοσηλεύεται αντίστοιχες ημέρες μέσω συνάρτησης τυχαίων αριθμών σε κάποιο προκαθορισμένο εύρος. Ένα χαρακτηριστικό παράδειγμα είναι ότι ασθενείς με ένα απλό κρύωμα το πιθανότερο θα νοσηλευτούν λιγότερες ημέρες από άλλους που έχουν πνευμονία. Επίσης ασθενείς μεγαλύτερων ηλικιών, είναι πιο ευάλωτες ομάδες και πιθανό να έχουν περισσότερες ημέρες νοσηλείας από άλλους νεότερους.

Τέλος δημιουργήσαμε την ημέρα εξόδου του ασθενούς η οποία προκύπτει από την ημερομηνία εισόδου συν τη διαφορά που προσθέτουν οι ημέρες νοσηλείας του.



Εικόνα 5.4: Mockaroo – Ημέρες Νοσηλείας

5.2.5 Πίνακας Γνωρισμάτων

Στο σημείο αυτό παρουσιάζεται ο πίνακας γνωρισμάτων έτσι όπως έχουν περιγράφει από τη παραπάνω υποενότητα, χωρίζοντάς τα σε αναγνωριστικά, ψευδο-αναγνωριστικά και ευαίσθητα σε μία προσπάθεια να γίνει πιο κατανοητή η διαδικασία δημιουργίας των δεδομένων.

Attribute	**Μορφή**	**Category**
First Name	String	Identifier
Lat Name	String	Identifier
AMKA	Number Sequence	Identifier

Hospital Number	Alphanumeric	Identifier
Gender	String	Quasi-Identifier
DOB	Date	Quasi-Identifier
Age	Integer	Quasi-Identifier
ZIP Code	Random Number Sequence	Quasi-Identifier
Hospitalization Date In / Out	Date	Sensitive
Days In	Integer	Sensitive
HRS	Boolean	Sensitive
Primary Disease	Custom List	Sensitive
Secondary Disease	Custom List	Sensitive
Pregnancy	Custom List	Insensitive
Surgery	Custom List	Sensitive
Patient Behavior	Custom List	Sensitive

Πίνακας 5.1: Γνωρίσματα Δεδομένων Υγείας

Κεφάλαιο 6
Ψευδωνυμοποίηση - Προσέγγιση

Το κεφάλαιο αυτό αφορά τη ψευδωνυμοποίηση των δεδομένων για τη μελέτη περίπτωσης που εξετάζουμε. Έχοντας πλέον συλλέξει τα δεδομένα υγείας, και αφού τα έχουμε ταξινομήσει σε πέντε κλινικές ανάλογα το περιστατικό, καλούμαστε για την ασφαλή διανομή τους προς έρευνα να τα ενοποιήσουμε σε έναν ενιαίο πίνακα. Για την επίτευξη αυτή αρχικά αποφασίσαμε να αντικαταστήσουμε τα αναγνωριστικά πεδία τους με ψευδώνυμα από κάθε κλινική, πριν τα στείλουμε στο τμήμα Ερευνών του Νοσοκομείου, σε μία προσπάθεια να προστατέψουμε την ιδιωτικότητα των ασθενών αποκρύπτοντας έτσι τη πραγματική τους ταυτότητα. Η ενέργεια αυτή αποσκοπεί στο να αποτρέψει, από μη εξουσιοδοτημένα άτομα εντός του Νοσοκομείου που θα αποκτήσουν πρόσβαση στα δεδομένα αυτά, να μπορέσουν να αναγνωρίσουν κάποιο φυσικό πρόσωπο μέσα σε αυτά.

6.1 Επιλογή Μεθόδου Ψευδωνυμοποίησης

Έχοντας μελετήσει αρκετές τεχνικές ψευδωνυμοποίησης από τα προηγούμενα κεφάλαια, καθώς και άλλες πιο προηγμένες προσεγγίσεις από έρευνες που διερευνήθηκαν, αναλύσαμε τα οφέλη, τις προκλήσεις αλλά και τα μειονεκτήματα της κάθε τεχνικής, δεδομένης της εφαρμογής τους σε ένα μεγάλο όγκο ευαίσθητων δεδομένων υγείας. Στόχος μας είναι να εξασφαλίσουμε τη προστασία της ιδιωτικότητας των ασθενών από επικείμενες επιθέσεις άμεσης αναγνώρισης εντός του περιβάλλοντος του Νοσοκομείου.

6.1.1 Διερεύνηση Μεθόδων

Από τα προηγούμενα κεφάλαια έχουμε εξετάσει διάφορες μεθόδους και τεχνικές ψευδωνυμοποίησης, άλλες απλούστερες ως προς την υλοποίηση και άλλες πιο περίπλοκες. Επίσης σε ορισμένες από τις τεχνικές αυτές κατά την υλοποίηση, το ψευδώνυμο εξαρτάται από το αναγνωριστικό κατά τη διαδικασία καθώς παράγεται από αυτό ενώ σε άλλες είναι εντελώς ανεξάρτητο. Διαχωρίζοντάς τις έτσι με βάση την εξάρτησή τους έχουμε δύο περιπτώσεις.

Στη περίπτωση των μη εξαρτώμενων τεχνικών όπως είδαμε έχουμε τον κοινό μετρητή και τη γεννήτρια (RNG) παραγωγής τυχαίων αριθμών ή χαρακτήρων. Οι μέθοδοι αυτοί, ούσες πλήρως ανεξάρτητες από το αναγνωριστικό που θέλουμε να ψευδωνυμοποιήσουμε, σε αρκετές περιπτώσεις θα μπορούσαν να προσφέρουν μεγαλύτερη ασφάλεια απόκρυψης του πραγματικού αναγνωριστικού γνωρίσματος και κατά συνέπεια προστασίας του υποκειμένου. Ωστόσο η εφαρμογή τους μπορεί να αξιοποιηθεί κυρίως σε μικρά μεγέθη συνόλων δεδομένων καθώς σε μεγαλύτερες βάσεις θα είχαμε προβλήματα αρίθμησης ή πιθανότητες «συγκρούσεων» μεταξύ διαφορετικών ψευδωνύμων παράγοντας το ίδιο ψευδώνυμο για διαφορετικές εισόδους. Επιτακτικό επίσης είναι να αποθηκεύεται κάπου με ασφάλεια και σε ξεχωριστό χώρο ένας πίνακας αντιστοίχισης του αναγνωριστικού με το ψευδώνυμο ώστε να μπορούμε ανά πάσα στιγμή να ανακτήσουμε τα πραγματικά δεδομένα. Στη περίπτωσή μας, καθώς αναφερόμαστε σε μεγάλο όγκο δεδομένων οι τεχνικές αυτές δεν θα μπορούσαν να εφαρμοστούν με αποδοτικό τρόπο, καθώς θα δημιουργούσαν αρκετές δυσκολίες ως προς την υλοποίηση. Η ακολουθία του μετρητή μάλιστα θα μπορούσε να προδώσει και κάποιο πρόσωπο στη περίπτωση που τα αναγνωριστικά είναι ταξινομημένα κατά αλφαβητική σειρά ή έστω κατά χρονολογική σειρά με βάση την ημερομηνία εισαγωγής.

Στη δεύτερη περίπτωση έχουμε τεχνικές στις οποίες τα ψευδώνυμα εξαρτώνται και δημιουργούνται με βάση το αρχικό αναγνωριστικό. Είναι αρκετά λειτουργικές σε μεγάλα σύνολα δεδομένων, ξεπερνώντας αρκετές δυσκολίες όπως την αρίθμηση σε μεγάλα σύνολα ή τις

πιθανότητες συγκρούσεων της τεχνικής RNG. Στις τεχνικές αυτές συγκαταλέγονται οι παρακάτω. α) Η κρυπτογράφηση του αναγνωριστικού ως μέσο ψευδωνυμοποίησης με κάποιο κλειδί κρυπτογράφησης. Είναι πολύ σημαντικό τόσο το κλειδί, όσο και ο αλγόριθμος κρυπτογράφησης να είναι αρκετά ισχυρός, για την αποφυγή επαναπροσδιορισμού του αρχικού αναγνωριστικού. Η περίπτωση αυτή δεν απαιτεί πίνακα συσχέτισης, καθώς με το κλειδί τη κρυπτογράφησης μπορούμε και να αποκρυπτογραφήσουμε το ψευδώνυμο για ανάκτηση του αρχικού γνωρίσματος. Οπότε κρίνεται απαραίτητη και η ασφαλής φύλαξη του κλειδιού καθώς με πιθανή απώλειά του δεν είμαστε σε θέση να ανακτήσουμε το αρχικό αναγνωριστικό.

β) Συνεχίζοντας έχουμε τη περίπτωση συνάρτησης κατακερματισμού (hash function) [53] η οποία αποτελεί μία πολύ ισχυρή τεχνική ψευδωνυμοποίησης αν υλοποιηθεί σωστά. Στη περίπτωση αυτή το αναγνωριστικό μέσω της συνάρτησης αυτής μετατρέπεται σε μία ακολουθία σταθερού πάντα μεγέθους ανεξάρτητα του μεγέθους της εισόδου, βασιζόμενο στον αλγόριθμο που χρησιμοποιήθηκε και πρόκειται για αλφαριθμητική ακολουθία. Η ακολουθία αυτή είναι πάντα η ίδια, όσο η είσοδος στη συνάρτηση παραμένει αναλλοίωτη και χρησιμοποιείται ο ίδιος αλγόριθμος κατακερματισμού. Ένα ακόμα πλεονέκτημα της μεθόδου αυτής είναι ότι οι συναρτήσεις κατακερματισμού αποτελούν μία διαδικασία μιας κατεύθυνσης (one-way), με την έννοια ότι έχοντας μία τιμή κατακερματισμού είναι υπολογιστικά αδύνατο να επαναπροσδιορίσουμε το αρχικό μήνυμα, ωστόσο υπάρχουν έμμεσες τεχνικές επιβεβαίωσης του αρχικού μηνύματος. Τέλος ένα σημαντικό κομμάτι της τεχνικής αυτής είναι ότι χρησιμοποιώντας μια καλή συνάρτηση κατακερματισμού όπως η sha-256 μπορούμε να αποφύγουμε τις πιθανότητες «σύγκρουσης» ψευδωνύμων, καθώς οι συναρτήσεις αυτές είναι ιδιαίτερα ευαίσθητες σε αλλαγές της εισόδου όπου ακόμα και μία ελάχιστη αλλαγή στην είσοδο της συνάρτησης θα φέρει ως αποτέλεσμα μία εντελώς διαφορετική έξοδο. Ωστόσο ακόμα και οι συναρτήσεις αυτές, αν και δείχνουν αρκετά ασφαλείς, έχουν και ένα μειονέκτημα, το οποίο οφείλεται στην ίδια την ιδιότητα τους. Γνωρίζοντας ότι κατακερματίζοντας το ίδιο ακριβώς μήνυμα με συγκεκριμένη συνάρτηση, δίνει πάντα την ίδια ακολουθία, μπορούμε έχοντας την ακολουθία αυτή στα χέρια μας να αναπαράξουμε τη διαδικασία και με έμμεσο τρόπο να εντοπίσουμε το αρχικό μήνυμα, επαληθεύοντας τις δύο αυτές hash τιμές. Επίσης υπάρχουν και αρκετές κατηγορίες επιθέσεων (rainbow tables [54], dictionary attacks, brute force attacks) [55] οι οποίες κάνουν σύγκριση μεταξύ δυο κατακερματισμένων τιμών επαληθεύοντας το αρχικό μήνυμα μέσα από διαδικασίες ελέγχου πιθανών κατακερματισμένων τιμών και της πραγματικής τιμής hash που βρίσκεται αποθηκευμένη.

Μία άλλη περίπτωση ψευδωνυμοποίησης είναι η χρήση κώδικα αυθεντικοποίησης μηνύματος (MAC) [56]. Οι συναρτήσεις αυτές υλοποιούνται με παρόμοιο τρόπο όπως και αυτές του κατακερματισμού, με τη διαφορά ότι κατά τη διαδικασία υπεισέρχεται και ένα κλειδί που έχει ορίσει ο αποστολέας, στη περίπτωσή μας ο υπεύθυνος για την επεξεργασία των δεδομένων και υλοποίησης της ψευδωνυμοποίησης. Η προσθήκη του κλειδιού, κάνει τις συναρτήσεις MAC ασφαλέστερες έναντι αυτών του κατακερματισμού ξεπερνώντας το ζήτημα του έμμεσου επαναπροσδιορισμού της εισόδου, όπως είδαμε παραπάνω, ωστόσο έχουν και αυτές άλλα μειονεκτήματα, όχι τόσο ως προς την υλοποίηση τους αλλά ως προς το χρήστη. Έτσι το κλειδί αυτό καθίσταται απολύτως απαραίτητο ώστε να μπορέσουμε να επαναπροσδιορίσουμε τα αρχικά γνωρίσματα καθώς σε περίπτωση απώλειάς του δεν έχουμε τη δυνατότητα ανάκτησή τους. Επίσης το κλειδί αυτό πρέπει να φυλάσσεται κάπου με ασφάλεια καθώς στη περίπτωση διαρροής του εκθέτουμε όλα τα δεδομένα του αρχικού πίνακα. Τόσο στη περίπτωση των συναρτήσεων κατακερματισμού όσο και στις MAC συναρτήσεις δεν μπορούμε με άμεσο τρόπο να επαναπροσδιορίσουμε το αρχικό μήνυμα, για αυτό και απαιτείται ένας πίνακας αντιστοίχισης του αναγνωριστικού και του ψευδωνύμου σε ασφαλή εξωτερικό χώρο ώστε να μπορούμε να ανακτήσουμε τα αρχικά δεδομένα σε περίπτωση ανάγκης.

6.1.2 Επιλογή Μεθόδου

Αφού εξετάσαμε τα προτερήματα καθώς και τις δυσκολίες διαφορετικών τεχνικών ψευδωνυμοποίησης, επιλέγουμε τη συνάρτηση κατακερματισμού ως την αποτελεσματικότερη για τη υλοποίησή μας. Ως πολιτική ψευδωνυμοποίησης θα ακολουθούσαμε τη ντετερμινιστική εκδοχή, όπου σε περίπτωση περισσότερων του ενός όμοιων γνωρισμάτων θα είχαμε και τα ίδια ψευδώνυμα – αυτό κρίνεται απαραίτητο, προκειμένου να μπορούν οι ερευνητές να εξάγουν ακριβή συμπεράσματα, αφού το ίδιο πρόσωπο θα εμφανίζεται πάντα με το ίδιο ψευδώνυμο (διαφορετικά, αν είχε διαφορετικό ψευδώνυμο, δεν θα προέκυπτε ότι, π.χ., διαφορετικές εγγραφές για το ίδιο πρόσωπο αφορούν πράγματι το ίδιο πρόσωπο). Ωστόσο θα ακολουθήσουμε μια προηγμένη τεχνική ψευδωνυμοποίησης με βάση τις συναρτήσεις κατακερματισμού, ως μία αρκετά απλούστερη παραλλαγή της ψευδωνυμοποίησης πολλαπλών αναγνωριστικών σε ένα ψευδώνυμο όπως είδαμε σε προηγούμενο κεφάλαιο, δημιουργώντας ακολουθίες γνωρισμάτων μοναδικές για κάθε εγγραφή.

Έχοντας απορρίψει τεχνικές όπως το μετρητή και την μέθοδο RNG που θα δημιουργούσαν προβλήματα λόγο του μεγάλου όγκου εγγραφών όπως αναφέραμε και σε προηγούμενη ενότητα,

κληθήκαμε να επιλέξουμε την τεχνική μας μεταξύ της συνάρτησης κατακερματισμού, κρυπτογράφησης και κώδικα αυθεντικοποίησης μηνύματος. Και οι τρείς αυτές τεχνικές μπορούν να προσφέρουν μεγάλη ασφάλεια αξιοποιώντας τες σωστά και να ψευδωνυμοποιήσουν τα αναγνωριστικά των ασθενών με αποτελεσματικό τρόπο.

Η τεχνική της συμμετρικής κρυπτογράφησης θα απαιτούσε ένα ισχυρό κλειδί καθώς σε περίπτωση αποκάλυψης του θα μπορούσαν να αποκρυπτογραφηθούν άμεσα όλα τα αναγνωριστικά των φυσικών προσώπων. Η ασύμμετρη κρυπτογράφηση από την άλλη θα απαιτούσε και δεύτερη οντότητα κάνοντας χρήση δύο κλειδιών, του δημόσιου για τη δημιουργία των ψευδωνύμων και του ιδιωτικού για τον επαναπροσδιορισμό του αρχικού γνωρίσματος. Επίσης κατά πιθανή απώλεια του κλειδιού και στις δύο περιπτώσεις θα ήταν αδύνατο να ανακτήσουμε τα αρχικά γνωρίσματα σε περίπτωση ανάγκης. Η αποκάλυψη ή η απώλεια του κλειδιού θα προκαλούσε τα ίδια προβλήματα και στη χρήση MAC.

Καταλήγοντας και αναλύοντας τη χρήση των συναρτήσεων κατακερματισμού, επιλέγοντάς αυτή ως μέθοδο ψευδωνυμοποίησης εντοπίζουμε κάποιες δυσκολίες και απειλές τις οποίες πρέπει πρώτα να ξεπεράσουμε. Όπως έχουμε αναφέρει οι συναρτήσεις κατακερματισμού μετατρέπουν ένα μήνυμα ανεξαρτήτου μεγέθους σε μία σταθερού μεγέθους αλφαριθμητική ακολουθία. Επίσης αναφέραμε ότι η αντίστροφη διαδικασία, δηλαδή ο επαναπροσδιορισμός του αρχικού μηνύματος από την ακολουθία αυτή είναι υπολογιστικά αδύνατος με οποιοδήποτε τρόπο. Ωστόσο, όπως αναφέραμε και παραπάνω, οι κίνδυνοι της έμμεσης αναγνώρισης μίας κατακερματισμένης τιμής αποτελούν ένα σημαντικό ζήτημα που πρέπει να αντιμετωπίσουμε.

Σε μία προσπάθεια να ξεπεράσουμε τους παραπάνω κινδύνους, οι συναρτήσεις κατακερματισμού μπορούν να εισάγουν μία επιπρόσθετη πληροφορία τυχαίας τιμής κατά τη διαδικασία του κατακερματισμού γνωστή και ως “salt”. Πρόκειται για μία ακολουθία αλφαριθμητικών ή και συμβόλων που κατά τη διαδικασία του κατακερματισμού προστίθεται στην αρχή ή στο τέλος της εγγραφής που θέλουμε να κατακερματίσουμε. Έτσι ο επιτιθέμενος καθώς δεν γνωρίζει τη τιμή αυτή, δεν μπορεί να αναπαράξει την διαδικασία και να εντοπίσει το μήνυμα που κατακερματίστηκε λόγω της ευαισθησίας που όπως αναφέραμε έχουν οι συναρτήσεις αυτές στις αλλαγές της εισόδου [57]. Κατά μία έννοια, το salt έχει το ρόλο μυστικού κλειδιού.

6.2 Εφαρμογή Της Ψευδωνυμοποίησης

Στο σημείο αυτό έχοντας καταλήξει στη χρήση συνάρτησης κατακερματισμού ως την αποτελεσματικότερη μέθοδο για το έργο μας, τα αναγνωριστικά των ασθενών θα αντικατασταθούν ως ψευδώνυμα από σταθερού μεγέθους μη αντιστρέψιμες ακολουθίες, έτσι ώστε να μη μπορεί κάποιος να ταυτοποιήσει άμεσα κάποιο φυσικό πρόσωπο μέσα στον πίνακα.

6.2.1 Ανάλυση Υλοποίησης

Η αρχική πρόταση έχοντας τέσσερα αναγνωριστικά πεδία στους πίνακες μας (Όνομα, Επίθετο, ΑΜΚΑ, Μοναδικός Αριθμός Ασθενούς) ήταν να αφαιρεθούν εντελώς τα ονοματεπώνυμα των ασθενών και να εφαρμοστεί ο κατακερματισμός στον Αριθμό Μητρώου Κοινωνικής Ασφάλισης ή τον Μοναδικό Αριθμό Ασθενούς και να καταργήσουμε το άλλο.

Ένα σημαντικό μειονέκτημα στο να κατακερματίσουμε ένα μοναδικό αριθμό ο οποίος είναι ευρέως γνωστός (ΑΜΚΑ) ή μπορεί εύκολα να εντοπιστεί (Μοναδικός Αριθμός Ασθενή) είναι η ίδια η διαδικασία που κάνουν οι συναρτήσεις κατακερματισμού όπως ήδη διαπιστώσαμε, δηλαδή να δημιουργούν πάντα την ίδια ακολουθία στην έξοδο [55]. Ο ΑΜΚΑ του ασθενούς αποτελεί χαρακτηριστικό παράδειγμα. Όπως αναφέραμε παραπάνω η διαδικασία του hashing είναι μη αναστρέψιμη, οπότε κατακερματίζοντας τον ΑΜΚΑ δεν θα μπορούσαμε να τον εντοπίσουμε από την ακολουθία που δημιουργήθηκε κατά τη διαδικασία.

Ωστόσο τί συμβαίνει στη περίπτωση που γνωρίζουμε κάποιο φυσικό πρόσωπο που νοσηλεύτηκε στο νοσοκομείο εκείνο το διάστημα. Αρκεί μόνο να βρούμε τον ΑΜΚΑ του προσώπου αυτού και τη συνάρτηση κατακερματισμού που χρησιμοποιήθηκε. Γνωρίζοντας λοιπόν τον ΑΜΚΑ, μπορούμε να τον κατακερματίσουμε με τη ίδια συνάρτηση γνωρίζοντας ότι η ακολουθία που θα μας δώσει είναι πάντα η ίδια και μοναδική για συγκεκριμένη είσοδο. Στη συνέχεια μπορούμε να συγκρίνουμε την ακολουθία που βρήκαμε με αυτές που βρίσκονται στο ψευδωνυμοποιημένο πίνακα και μιας και ο ΑΜΚΑ αποτελεί μοναδικό αριθμό για κάθε ασθενή μπορούμε να αναγνωρίσουμε έμμεσα ένα πρόσωπο. Το ίδιο ισχύει και για τον Μοναδικό Αριθμό ασθενή και είναι ένα μεγάλο μειονέκτημα στη προσπάθεια να ακολουθήσουμε αυτή τη πρακτική, διακινδυνεύοντας την διαρροή προσωπικών δεδομένων και την ιδιωτικότητα των ασθενών [55].

Μελετώντας λοιπόν τις εναλλακτικές μας, καταλήξαμε στον κατακερματισμό του συνδυασμού πολλαπλών αναγνωριστικών γνωρισμάτων χωρίς να αφαιρέσουμε κάποιο πεδίο. Η παραλλαγή αυτή βασίζεται στη τεχνική ψευδωνυμοποίησης πολλαπλών γνωρισμάτων σε ένα ψευδώνυμο, ωστόσο αποτελεί μία απλούστερη παραλλαγή της χωρίς να χρειάζονται επίπεδα

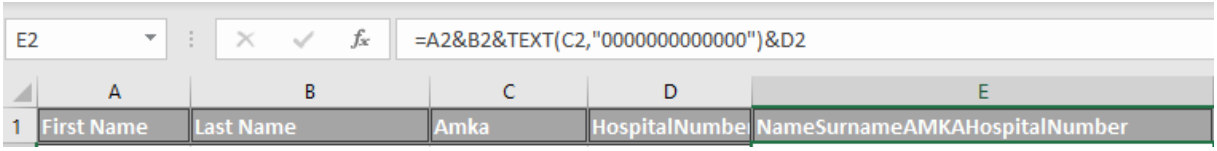
ψευδωνυμοποίησης όπως το δέντρο Merkle. Συνδυάζοντας όλα τα αναγνωριστικά πεδία (Όνομα, Επίθετο, ΑΜΚΑ, Hospital Number) και κατακερματίζοντάς τα θα δημιουργηθεί μία μοναδική ακολουθία που πολύ δύσκολα θα μπορέσει να συνδυάσει κανείς και να την αναπαράξει ώστε να ταυτοποιήσει κάποιον ασθενή [28]. Το γεγονός ότι οι χρήστες της εφαρμογής που αρχικά περιέχει τους ασθενείς, δεν έχουν πρόσβαση σε ασθενείς άλλων κλινικών διασφαλίζει ότι ακόμα και αν κάποιος μπορεί να εντοπίσει την κατακερματισμένη πλέον τιμή, μπορεί με εξαιρετικά μεγάλη δυσκολία να αναγνωρίσει κάποιο φυσικό πρόσωπο μέσα στο πίνακα αναπαράγοντας την ακολουθία.

Τέλος στη προσπάθεια να εξαλείψουμε ακόμα και την ελάχιστη πιθανότητα αναπαραγωγής κάποιας ακολουθίας και κατά συνέπεια ταυτοποίησης κάποιου φυσικού προσώπου μέσω του ψευδωνυμοποιημένου πίνακα αποφασίσαμε να συμπεριλάβουμε και ένα πρόθεμα ως 'salt' στη διαδικασία του hashing το οποίο θα προστατεύεται από κάθε μη εξουσιοδοτημένη πρόσβαση (κατάλληλοι μηχανισμοί, υπό την εποπτεία του υπευθύνου προστασίας προσωπικών δεδομένων (DPO) του Νοσοκομείου, θα αναπτυχθούν για το σκοπό αυτό). Η μορφή της εισόδου στην συνάρτηση κατακερματισμού πλέον θα είναι Salt_Name_LastName_AMKA_HospitalNumber.

6.2.2 Υλοποίηση Της Ψευδωνυμοποίησης

Αρχικά ακολουθώντας τη παραπάνω διαδικασία θα πρέπει να συγχωνεύσουμε τα τέσσερα αυτά αναγνωριστικά σε ένα ως συνεχόμενη ακολουθία (Name_LastName_AMKA_HospitalNumber). Τα δεδομένα μας για κάθε κλινική είναι αποθηκευμένα σε μορφή πινάκων όπως δημιουργήθηκαν από προηγούμενο κεφάλαιο σε μορφή Excel.

Αρχικά εξάγουμε τα τέσσερα αυτά αναγνωριστικά για όλες τις εγγραφές σε ξεχωριστό πίνακα ο οποίος θα αποτελέσει τον πίνακα αντιστοίχισης μεταξύ των αναγνωριστικών και των ψευδωνύμων τους. Αυτό μπορεί να υλοποιηθεί με συνάρτηση του εργαλείου Excel η οποία μπορεί να επιλέξει τα κελία που επιθυμούμε και να τα αποδώσει σε κάποιο νέο.



Εικόνα 6.1: Συνάρτηση Excel - Συγχώνευση αναγνωριστικών

Η συνάρτηση (**=A2&B2&TEXT(C2,"0000000000000")&D2**) εκχωρεί στο κελί Ε μία ακολουθία αποτελούμενη από τα κελία Α, Β, C, D. Στο κελί C όπου περιλαμβάνεται ο ΑΜΚΑ του ασθενούς δηλώνουμε με μηδενικά, όσα και το μήκος του ΑΜΚΑ, τη μορφή που θα εμφανίζεται καθώς μη δηλώνοντάς την, αλλοιώνονται οι ΑΜΚΑ που ξεκινούν με 0. Η διαδικασία αυτή ακολουθείται σε όλες τις εγγραφές του εκάστοτε πίνακα δημιουργώντας έτσι το πρώτο μέρος του πίνακα αντιστοίχισης.

Στη συνέχεια έχοντας τον πίνακα αντιστοίχισης που περιέχει τα αναγνωριστικά ασθενών όλων των εγγραφών του πίνακα, θα πρέπει να κατακερματίσουμε τις τιμές αυτές ώστε να παράξουμε το ψευδώνυμο για τον εκάστοτε ασθενή. Για την υλοποίηση του κατακερματισμού διαδοχικά σε όλες τις εγγραφές του πίνακα δημιουργήσαμε ένα script που κατακερματίζει την είσοδο που λαμβάνει χρησιμοποιώντας τον αλγόριθμο sha-256.

Αναλυτικότερα, το script αυτό αρχικά ανοίγει το αρχείο excel και το φύλλο που περιέχεται η ακολουθία των αναγνωριστικών των ασθενών που δημιουργήσαμε παραπάνω.

```
# Load the Excel COM object
$excel = New-Object -ComObject Excel.Application

# Open the Excel workbook
$workbook = $excel.Workbooks.Open("C:\Users\Doulge\Desktop\AYD_Thesis\Script\Pseudonymized_Data.xlsx")

# Select the worksheet and range of cells to read
$worksheet = $workbook.Worksheets.Item("Sheet2")
```

Εικόνα 6.2: Script - Άνοιγμα αρχείου Excel

Εισάγουμε ένα εύρος ίσο με τις εγγραφές των ασθενών στο κελί Ε, από όπου «διαβάζει» την ακολουθία της κάθε εγγραφής.

```
$range = $worksheet.Range("E2:E100001")
```

Εικόνα 6.3: Script – Εύρος εγγραφών

Στη συνέχεια δημιουργούμε μία μεταβλητή 'salt' αποδίδοντάς της μία προκαθορισμένη τιμή ως πρόθεμα της εισόδου στη διαδικασία του hashing.

```
# Create a salt to use for hashing
$salt = "wT8s45_"
```

Εικόνα 6.4: Script – Δημιουργία Salt

Συνεχίζοντας, δημιουργούμε ένα νέο φύλλο όπου θα κρατούνται οι τιμές hash ως ψευδώνυμα.

```
# Create a new worksheet to hold the hashed values
$newWorksheet = $workbook.Worksheets.Add()
$newWorksheet.Name = "Hashes"
```

Εικόνα 6.5: Script – Δημιουργία νέου φύλλου hash τιμών

Τέλος δημιουργούμε τη συνάρτηση που θα κάνει τον κατακερματισμό όλων των αναγνωριστικών δημιουργώντας τα ψευδώνυμα των ασθενών.

```
$sha256 = New-Object -TypeName System.Security.Cryptography.SHA256CryptoServiceProvider
$row = 1
foreach ($cell in $range) {
    $value = $cell.Value2
    $valueWithSalt = $salt + $value
    $hash = [System.BitConverter]::ToString($sha256.ComputeHash([System.Text.Encoding]::UTF8.GetBytes($valueWithSalt)))
    $hash = $hash.Replace("-", "").ToLower()
    $newWorksheet.Cells.Item($row, 1) = $hash
    $row++
}
```

Εικόνα 6.6: Script – Υλοποίηση συνάρτησης κατακερματισμού

Αρχικά δημιουργεί μία κλάση για την υλοποίηση συνάρτηση κατακερματισμού με τον αλγόριθμο sha-256 αποδίδοντάς τη σε μία μεταβλητή $sha256, αρχικοποιώντας μία μεταβλητή $row ως μετρητή. Στη συνέχεια παίρνει το κάθε κελί διαδοχικά στο εύρος που του δηλώσαμε προηγουμένως και κάνει τις εξής διαδικασίες. Πρώτα κρατάει σε μία μεταβλητή $value την τιμή του κελιού που βρίσκεται, συνεχίζει προσθέτοντας το salt στη αρχή της τιμής αυτής ως πρόθεμα δημιουργώντας τη ακολουθία που θέλουμε να κατακερματίσουμε και τη κρατάει σε μία νέα μεταβλητή $valueWithSalt. Δημιουργεί μία νέα μεταβλητή $hash όπου αποδίδει τη τιμή κατακερματισμού για είσοδο ίση με το περιεχόμενο της $valueWithSalt, αφού πρώτα υπολογίσει τη τιμή αυτή με τον αλγόριθμο sha-256 χρησιμοποιώντας τη μέθοδο ComputeHash και στη συνέχεια τη μετατρέψει σε string. Η επόμενη εντολή αφαιρεί τυχόν παύλες '-' που μπορεί να υπάρχουν στη κατακερματισμένη τιμή και τη μορφοποιεί σε πεζούς χαρακτήρες. Προχωρώντας εκχωρεί τη κατακερματισμένη τιμή που κρατάει στη $hash σε συγκεκριμένο κελί στο φύλλο εργασίας το οποίο κελί προσδιορίζεται από τη μεταβλητή $row υποδηλώνοντας και τη πρώτη

στήλη του φύλλου. Τέλος αυξάνει το μετρητή $row κατά ένα και επαναλαμβάνει τη διαδικασία για όλο το εύρος μέχρι να κατακερματίσει όλες τις τιμές του αρχικού πίνακα, δημιουργώντας έτσι όλα τα ψευδώνυμα των ασθενών και κατά την ολοκλήρωση της διαδικασία αποθηκεύει το excel αρχείο που πλέον περιέχει όλες τις ψευδωνυμοποιημένες τιμές των αναγνωριστικών των ασθενών.

Οι τιμές αυτές πλέον θα εκχωρηθούν επίσης στον πίνακα αντιστοίχισης ολοκληρώνοντας τη διαδικασία της ψευδωνυμοποίησης των αναγνωριστικών. Τα ψευδώνυμα πλέον μπορούν με ασφάλεια να αντικαταστήσουν τα τέσσερα αυτά αναγνωριστικά στους πίνακες των κλινικών και να αποσταλούν στο τμήμα Ερευνών του Νοσοκομείου ώστε να δημιουργηθεί ο ενιαίος πίνακας.

6.3 Σύνοψη

Στο σημείο αυτό έχουμε εξασφαλίσει ότι τα αρχεία που θα αποσταλούν από τις κλινικές προς το τμήμα Ερευνών του Νοσοκομείου δεν περιέχουν πλέον κάποιο από τα αναγνωριστικά που θα μπορούσαν να ταυτοποιήσουν άμεσα κάποιον ασθενή και αντί αυτού περιέχουν μόνο μια κατακερματισμένη ακολουθία που μπορεί να αναπαράξει μόνο ο υπεύθυνος ασφαλείας των προσωπικών δεδομένων του νοσοκομείου, αν και εφόσον λάβει έγκριση από το DPO του Νοσοκομείου. Επίσης με το τρόπο αυτό τα δεδομένα μας παραμένουν πλήρως λειτουργικά. Σημαντικό στο σημείο αυτό είναι να αναφερθεί ότι τα δεδομένα ακόμα και τώρα αποτελούν προσωπικά δεδομένα.

Ωστόσο υπάρχουν και άλλες προκλήσεις καθώς και κίνδυνοι που αφορούν την προστασία των προσωπικών δεδομένων των ασθενών και πρέπει να αντιμετωπιστούν. Οι προκλήσεις αυτές αφορούν τα ψευδο-αναγνωριστικά πεδία των εγγραφών που θα μπορούσαν έμμεσα να ταυτοποιήσουν κάποιο φυσικό πρόσωπο μέσα στον πίνακα. Ακόμα και αν τα ψευδώνυμα είναι μη αναστρέψιμα εφόσον προστατεύεται το κλειδί, δεν μπορεί να αποκλειστεί η δυνατότητα να αναγνωριστούν ένας ή περισσότεροι ασθενείς από τα ψευδο-αναγνωριστικά τους– και αυτό καταδεικνύει ότι μία αποτελεσματική ψευδωνυμοποίηση δεν διασφαλίζεται πάντα με τη δημιουργία ισχυρών, μη αναστρέψιμων, ψευδωνύμων. Προκειμένου να γίνει σαφής ο κίνδυνος αναγνώρισης ασθενών, αξίζει να ανακαλέσουμε τις σχετικές προβλέψεις του GDPR ακόμα και για τα ίδια τα ανώνυμα δεδομένα (άρα, πολύ περισσότερο δε για τα ψευδωνυμοποιημένα). Όπως αναφέρεται ρητά μέσα στον GDPR, για να μπορέσουμε να χαρακτηρίσουμε τα δεδομένα ως ανώνυμα, θα πρέπει να εξασφαλίσουμε ότι οι πληροφορίες τους δεν μπορούν να ταυτοποιήσουν

το υποκείμενο με οποιοδήποτε τρόπο είτε άμεσο, είτε έμμεσο. Ένα χαρακτηριστικό παράδειγμα είναι η πιθανότητα να υπάρχει ένα και μόνο φυσικό πρόσωπο μέσα στον πίνακα που πάσχει από μία σπάνια ασθένεια ή η γνώση μας πως επισκέφτηκε το Νοσοκομείο άτομο του κύκλου μας από συγκεκριμένη περιοχή, σε συγκεκριμένη χρονική περίοδο και συγκεκριμένης ηλικίας και είναι ο μοναδικός μέσα στο πίνακα που πληροί όλες τις προϋποθέσεις. Διατρέχοντας τον πίνακα ταυτοποιήσαμε έμμεσα το πρόσωπο αυτά και μάλιστα μάθαμε ότι πάσχει από καρκίνο. Άρα σίγουρα δεν μπορούμε να διατυπώσουμε ότι τα δεδομένα είναι ανώνυμα: ωστόσο, ακριβώς για τους ίδιους λόγους, είναι αμφίβολο και αν μπορούμε να διατυπώσουμε ότι τα δεδομένα είναι επαρκώς ψευδωνυμοποιημένα, αφού – ακριβώς για τους ίδιους λόγους – η μη αντιστρεψιμότητα του ψευδωνύμου δεν επαρκεί για να αποτρέψει την πιθανότητα αναγνώρισης κάποιου προσώπου.

Μπορούμε λοιπόν να καταλήξουμε σε ένα πρώτο συμπέρασμα αναλύοντας τα παραπάνω δεδομένα και να κατανοήσουμε ότι η ψευδωνυμοποίηση από μόνη της, όσο καλά σχεδιασμένη και αν είναι, δεν αρκεί εάν περιορίζεται απλά στην αντικατάσταση των αναγνωριστικών με ένα μη αναστρέψιμο ψευδώνυμο.

Κεφάλαιο 7
Εφαρμογή τεχνικών ανωνυμοποίησης - Προσέγγιση

Σε αυτό το κεφάλαιο θα μελετήσουμε μερικές από τις προσεγγίσεις που μπορούμε να ακολουθήσουμε κατά την ανωνυμοποίηση των δεδομένων μας τις οποίες και θα εφαρμόσουμε στη συνέχεια της έρευνάς μας. Έως τώρα έχουμε αναφερθεί σε ορισμένα από τα μοντέλα απορρήτου όπως και μεθόδους ανωνυμοποίησης που μπορούμε να εξετάσουμε με κατάλληλα εργαλεία και θα δούμε αναλυτικότερα σε επόμενο κεφάλαιο.

7.1 Μοντέλα Απορρήτου Και Μέθοδοι Ανωνυμοποίησης

Κατά τη εφαρμογή της ανωνυμοποίησης θα συνδυάσουμε κάποια από τα βασικότερα μοντέλα απορρήτου καθώς και μεθόδους ανωνυμοποίησης ώστε να ελαττώσουμε κατά το δυνατόν τη δυνατότητα αναγνώρισης προσώπου στα δεδομένα μας. Πρόκειται για μια αρκετά απαιτητική διαδικασία που χρήζει ιδιαίτερης προσοχής καθώς αφορά ιδιαίτερα ευαίσθητα δεδομένα.

7.1.1 Μοντέλα Απορρήτου

Όπως αναφερθήκαμε και παραπάνω κάποια από τα βασικά μοντέλα απορρήτου που έχουμε μελετήσει έως τώρα και θα περιορίσουμε την έρευνά μας πάνω σε αυτά είναι τα εξής:

1. k-ανωνυμία: όπως έχουμε αναφέρει προσφέρει ένα ελάχιστο επίπεδο ανωνυμίας στα ψευδο-αναγνωριστικά του συνόλου δεδομένων δημιουργώντας κλάσεις ισοδυναμίας οι οποίες περιέχουν κατ' ελάχιστο k εγγραφές που μοιράζονται τα ίδια ψευδο-αναγνωριστικά [27]. Το μοντέλο αυτό στοχεύει στη προστασία από την αποκάλυψη ταυτότητας.

2. l-διαφορετικότητα: το μοντέλο αυτό αντίθετα με τη k-ανωνυμία στοχεύει στη προστασία από την αποκάλυψη γνωρισμάτων. Εξασφαλίζει για τα εμπιστευτικά γνωρίσματα ότι κάθε κλάση k ισοδυναμίας θα περιέχει τουλάχιστον l διαφορετικές τιμές του εμπιστευτικού/ευαίσθητου πεδίου [27].

3. t-εγγύτητα: όπως και η l-διαφορετικότητα το μοντέλο αυτό επικεντρώνεται στη προστασία από την αποκάλυψη γνωρισμάτων. Επίσης εξασφαλίζει ότι η απόσταση μεταξύ της κατανομής του ευαίσθητου γνωρίσματος σε μία κλάση ισοδυναμίας συγκριτικά με τα υπόλοιπες κλάσεις του πίνακα είναι μικρότερη ή ίση με t [27].

Τέλος υπάρχουν αρκετά μοντέλα ακόμα (δ-disclosure, β-likeness, k-map, δ-presence, διαφορική ιδιωτικότητα) τα οποία δεν θα εξεταστούν περισσότερο στη παρούσα μεταπτυχιακή διατριβή καθώς σκοπεύουμε να επικεντρωθούμε στα βασικά μοντέλα απορρήτου.

7.1.2 Μέθοδοι Ανωνυμοποίησης

Όσον αφορά τις μεθόδους ανωνυμοποίησης είδαμε ότι υπάρχει μία πληθώρα περιπτώσεων (όπως η γενίκευση, η κάλυψη, η προσθήκη θορύβου, τα συνθετικά δεδομένα κλπ.), ωστόσο κατά την έρευνά μας, βάσει των ψευδο-γνωρισμάτων όπως έχουν διαμορφωθεί από τη δημιουργία των δεδομένων μας, θα περιοριστούμε στην εφαρμογή συγκεκριμένων μεθόδων όπως αυτή της γενίκευσης (generalization) και της κάλυψης (masking). Η γενίκευση αντικαθιστά μία τιμή με ένα εύρος τιμών όπως για παράδειγμα την ηλικία (έστω 25) σε ένα εύρος ηλικιών(20-30). Από την άλλη η κάλυψη στοχεύει στην απόκρυψη του τελευταίου στοιχείου (συνήθως αριθμητικών γνωρισμάτων) σε κάθε επίπεδο. Παράδειγμα της κάλυψης είναι ο ταχυδρομικός κώδικας όπου

εφαρμόζοντάς την έχοντας έστω έναν ΤΚ 1234 και μετατρέπεται σε 123* ενώ σε κάθε επόμενο επίπεδο καλύπτεται ένας ακόμα αριθμός από το τέλος [25].

7.2 Προσέγγιση

Κατά την έρευνά μας όσον αφορά το κομμάτι της ανωνυμοποίησης, όπως έχουμε αναφέρει και παραπάνω, στόχο έχουμε να εξετάσουμε διαφορετικές προσεγγίσεις της και να αναλύσουμε τα αποτελέσματά τους. Θέλουμε να εξετάσουμε την αποτελεσματικότητα της μίας προσέγγισης έναντι κάποιας άλλης ως προς το κατά πόσον μειώνεται ο κίνδυνος αποκάλυψης της ταυτότητας κάποιας οντότητας και της αποκάλυψης γνωρισμάτων, όσο και τη διατήρηση κατά το μέγιστο της χρήσιμης πληροφορίας μετά τη διαδικασία. Με το τρόπο αυτό αναμένουμε να κρατήσουμε τα δεδομένα μας όσο το δυνατό αναλλοίωτα για της ανάγκες μίας αποτελεσματικής έρευνας, προστατεύοντας παράλληλα θεμελιώδη δικαιώματα των προσώπων, βάση τους περιορισμούς που έχει ορίσει ο GDPR [58].

7.2.1 Κατηγοριοποίηση Γνωρισμάτων

Θα ξεκινήσουμε τη διαδικασία, ορίζοντας μέσα στο σύνολο των δεδομένων υγείας, τα αναγνωριστικά, τα ψευδο-αναγνωριστικά καθώς και ένα από τα ευαίσθητα γνωρίσματα ως εμπιστευτικό (τη νόσο HRS συγκεκριμένα). Θα αφήσουμε αρχικά τα υπόλοιπα ευαίσθητα γνωρίσματα ως μη εμπιστευτικά, κρατώντας τη πρώτη μας προσέγγισή ανεπηρέαστη από αυτά και θα εξετάσουμε συνδυασμούς των μοντέλων απορρήτου που μελετήσαμε (k-ανωνυμία και l-διαφορετικότητα κυρίως, καθώς η t-εγγύτητα δημιουργεί προβληματισμούς σχετικά με τα εξαχθέντα δεδομένα λόγω τη πολυπλοκότητας των γνωρισμάτων του αρχικού μας πίνακα) παρατηρώντας κατά πόσο επηρεάζει η κάθε προσέγγιση τα εξαχθέντα δεδομένα ως προς την απώλεια πληροφορίας αλλά και το ποσοστό του ρίσκου πιθανής αποκάλυψης κάποιας εγγραφής. Επίσης θα εξετάσουμε και το βάθος του επιπέδου γενίκευσης του κάθε ψευδο-αναγνωριστικού που απαιτείται ώστε να λάβουμε τον ιδανικότερο συνδυασμό χρησιμότητας/ρίσκου αποκάλυψης στο σύνολο των δεδομένων μας. Έχοντας ορίσει ένα γνώρισμα ως εμπιστευτικό κάνει τη διαδικασία ευκολότερη (ειδικά στη περίπτωση της νόσου η οποία παίρνει δύο τιμές, θετική ή αρνητική) συγκριτικά με το να έχουμε περισσότερα του ενός εμπιστευτικά γνωρίσματα.

Στη πορεία της διαδικασίας, σε επόμενες προσεγγίσεις θα συμπεριληφθούν και άλλα από τα ευαίσθητα γνωρίσματα ως εμπιστευτικά καθώς και διαφορετικές παράμετροι για τα μοντέλα

απορρήτου, με σκοπό να εξετάσουμε πως θα διαμορφωθούν τα δεδομένα μας καθώς και τα αποτελέσματά μας. Αναμένουμε αρκετές προκλήσεις και δυσκολίες όσο βλέπουμε να αυξάνονται τα εμπιστευτικά γνωρίσματα του συνόλου δεδομένων κατά τη διαδικασία ανωνυμοποίησής τους, κάτι που επαληθεύει το γεγονός ότι η ανωνυμοποίηση αποτελεί μία αρκετά απαιτητική διαδικασία που χρειάζεται ιδιαίτερη προσοχή, βασιζόμενοι πάντα στις ανάγκες που απαιτεί η έρευνα που επιθυμούμε να πραγματοποιήσουμε. Τέλος δημιουργείται και ένας προβληματισμός σχετικά με ένα γνώρισμα, αυτό της εγκυμοσύνης, το οποίο αν και δεν μπορεί να οριστεί ως ψευδο-αναγνωριστικό αλλά ούτε και ως εμπιστευτικό στη παρούσα μελέτη μας (καθώς δεν είναι, αλλά και πάλι θα δημιουργούσε σημαντικό ποσοστό απώλειας πληροφορίας καθιστώντας την ανωνυμοποίηση ανέφικτη), δημιουργεί διλλήματα ως προς την αξιοποίησή του. Αξιοσημείωτο είναι το γεγονός ότι ακόμη και ως μη εμπιστευτικό, το γνώρισμα αυτό δημιουργεί σημαντικούς κινδύνους αναγνώρισης κάποιου φυσικού προσώπου αν συνδυαστεί με κάποιο ψευδο-αναγνωριστικό ή εμπιστευτικό γνώρισμα, καταδεικνύοντας τη δυσκολία της ανωνυμοποίησης και τις προκλήσεις που δημιουργεί ως προς την αντιμετώπιση κάποιου γνωρίσματος, που υπό προϋποθέσεις μπορεί λειτουργήσεις ως ψευδο-αναγνωριστικό.

7.2.2 Διαδικασία Ανωνυμοποίησης

Έχοντας κατηγοριοποιήσει τα δεδομένα μας θα χρειαστεί να εφαρμόσουμε τη κατάλληλη μέθοδο ανωνυμοποίησης ανάλογα τα ψευδο-αναγνωριστικά, επιλέγοντας μεταξύ της γενίκευσης ή της κάλυψης βάση της δομής τους, όπως δημιουργήθηκαν στη διαδικασία της συλλογής τους σε προηγούμενο κεφάλαιο. Οι μέθοδοι αυτοί στοχεύουν ώστε συγκεκριμένες τιμές γνωρισμάτων που θα μπορούσαν έμμεσα να ταυτοποιήσουν κάποιο φυσικό πρόσωπο μέσα στο σύνολο δεδομένων, να αντικατασταθούν από πιο γενικευμένες εκδοχές τους ή να καλυφθούν ανάλογα, με στόχο τη προστασία κατά ένα βαθμό των υποκειμένων της έρευνας. Συνεχίζοντας θα δημιουργήσουμε επίπεδα βάθους της γενίκευσης ή της κάλυψης που θα εφαρμόσουμε σε κάθε ψευδο-αναγνωριστικό προετοιμάζοντάς τα για την ανωνυμοποίησή τους [59]. Στο σημείο αυτό απαιτείται να είμαστε ιδιαίτερα προσεκτικοί στις επιλογές μας, καθώς όσο αυξάνεται σε βάθος το επίπεδο γενίκευσης/κάλυψης, τόσο μειώνεται το ρίσκο αποκάλυψης, ωστόσο αυξάνεται η απώλεια χρήσιμης πληροφορίας για τις ανάγκες μιας αποτελεσματικής έρευνας.

Στη συνέχεια θα πρέπει να επιλέξουμε τα μοντέλα απορρήτου που θα εφαρμόσουμε στα δεδομένα μας και να προχωρήσουμε στην διαδικασία ανωνυμοποίησής τους. Τα τρία βασικά αυτά μοντέλα που θα εξετάσουμε κατά την έρευνά μας, όπως έχουμε ήδη δει έως τώρα δουλεύουν συνδυαστικά

ώστε να μπορούν να προστατέψουν τα δεδομένα μας από διαφορετικούς τύπους επιθέσεων [23], όπως και μοντέλα επιτιθέμενων βάση των γνώσεων που έχουν μέσα στο σύνολο δεδομένων και τους στόχου τους [20]. Κατά τη διαδικασία της ανωνυμοποίησης θα εξεταστούν διαφορετικές τιμές για τα k, l και t αντίστοιχα για το κάθε μοντέλο απορρήτου.

7.2.3 Ανάλυση Αποτελεσμάτων

Ολοκληρώνοντας τη διαδικασία της ανωνυμοποίησης σε κάθε προσέγγιση είναι σημαντικό να καταγράφουμε τα αποτελέσματα μας τόσο αναφορικά με τη χρησιμότητα των δεδομένων η οποία προκύπτει από την απώλεια της χρήσιμης πληροφορίας ή και ολόκληρων εγγραφών που θα προκύπτει από τη διαδικασία, όσο και του ρίσκου αποκάλυψης την ταυτότητας ή των γνωρισμάτων κάποιας εγγραφής μέσα στο σύνολο δεδομένων. Εξετάζοντας την ανάλυση του ρίσκου αλλά και της χρησιμότητας των δεδομένων μπορούμε να καταλήξουμε στη πιθανή ασφαλέστερη προσέγγιση ανωνυμοποίησης των δεδομένων μας, αλλά και να έρθουμε αντιμέτωποι με τις προκλήσεις που επιφέρει μία τέτοια διαδικασία.

Κεφάλαιο 8
Επιλογή Εργαλείου

Στο κεφάλαιο αυτό θα εξετάσουμε κάποια από τα εργαλεία που θα εξυπηρετήσουν την υλοποίηση των τεχνικών ανωνυμοποίησης καταλήγοντας στο καταλληλότερο για την έρευνά μας. Στόχος είναι να επιλέξουμε κάποιο εργαλείο το οποίο θα προσφέρει αρκετές λειτουργίες προσαρμόζοντας την ανωνυμοποίηση των δεδομένων κατάλληλα, προσφέροντας διαφορετικές προσεγγίσεις μεθόδων και εξέτασης διάφορων μοντέλων απορρήτου. Τα αποτελέσματα της διαδικασίας που θα αναλυθούν στη συνέχεια παίζουν και αυτά καθοριστικό ρόλο στην επιλογή μας. Τέλος θα θέλαμε ακόμη το εργαλεία αυτό να είναι σχετικά εύκολο στη χρήση και αρκετά φιλικό προς το χρήστη.

8.1 ARX

Το ARX αποτελεί εργαλείο ανοιχτού κώδικα και χρησιμοποιείται για ανωνυμοποίηση δεδομένων, παρέχοντας προηγμένες δυνατότητες για τη προστασία της ιδιωτικότητας στα δεδομένα αυτά. Περιλαμβάνει αρκετές προσεγγίσεις ανωνυμοποίησης, μοντέλα απορρήτου και ρίσκου καθώς και

αρκετές μεθόδους μετριασμού του κινδύνου των δεδομένων αυτών προστατεύοντάς τα από την αποκάλυψη εμπιστευτικών πληροφοριών μέσα σε αυτά [35]. Επίσης προσφέρει αρκετές μεθόδους ανάλυσης και εξέτασης των αποτελεσμάτων της ανωνυμοποίησης, του ρίσκου και της λειτουργικότητας των ανωνυμοποιημένων δεδομένων συγκριτικά με την απώλεια χρήσιμης πληροφορίας.

8.1.1 Προσέγγιση

Η προσέγγιση της ανωνυμοποίησης βασίζεται σε ιεραρχίες των γνωρισμάτων, ανάλογα με το είδος τους (ημερομηνίες, ακέραιους αριθμούς, συμβολοσειρές κλπ.). Κατά τη δημιουργία των ιεραρχιών μπορούμε να επιλέξουμε αν θα ακολουθήσουμε τη μέθοδος της γενίκευσης ή τη κάλυψη μέρους των γνωρισμάτων. Επίσης δίνεται η δυνατότητα απόδοσης βάρους σε συγκεκριμένα γνωρίσματα αξιοποιώντας αποτελεσματικότερα την ανωνυμοποίηση χωρίς να χαθεί μεγάλο μέρος χρήσιμης πληροφορίας. Τέλος κατά την υλοποίηση της διαδικασίας μπορούμε να επιλέξουμε μεταξύ διάφορων μετασχηματισμών στα γνωρίσματά μας, επιλέγοντας το επίπεδο της ανωνυμοποίησης σε κάθε γνώρισμα.

8.1.2 Μοντέλα Απορρήτου

Το εργαλείο αυτό προσφέρει μία πληθώρα μοντέλων απορρήτου που μπορούν να εξυπηρετήσουν το σκοπό μας ανάλογα τη προσέγγισή μας στην ανωνυμοποίηση [20]. Τα βασικότερα και πιο διαδεδομένα μοντέλα παρουσιάζονται παρακάτω.

1. k-anonymity
2. l-diversity
3. t-closeness
4. Differential privacy
5. k-map
6. δ-disclosure

7. β-likeness

Κάποια από τα μοντέλα αυτά λειτουργούν συνδυαστικά ενώ κάποια άλλα μπορούν αυτούσια να προσφέρουν ισχυρή ανωνυμοποίηση των δεδομένων. Επίσης κάποια εστιάζουν στην αποκάλυψη της ταυτότητας (identity disclosure) όπως το k-anonymity και το k-map, ενώ κάποια άλλα όπως το l-diversity, t-closeness, δ-disclosure εστιάζουν στην αποκάλυψη γνωρισμάτων(attribute disclosure).

8.1.3 Μέθοδοι Ανωνυμοποίησης

Το εργαλείο επίσης προσφέρει αρκετές μεθόδους ανωνυμοποίησης και μετριασμού των δεδομένων όπως η γενίκευση των πληροφοριών δημιουργώντας ιεραρχικές φόρμες προσαρμοσμένες στον τύπο δεδομένων, τυχαία δειγματοληψίας των εισαχθέντων δεδομένων, κατάργηση ολόκληρων εγγραφών ή συγκεκριμένων γνωρισμάτων, μικροσυσσωμάτωση αριθμητικών δεδομένων μετασχηματίζοντάς τα σε κοινές τιμές με συναρτήσεις συνάθροισης που μπορεί να ορίσει ο χρήστης, κατηγοριοποίηση των γνωρισμάτων δημιουργώντας κανόνες μετασχηματισμού [60]. Κατά τη διαδικασία της έρευνάς μας θα χρησιμοποιηθούν οι μέθοδοι της γενίκευσης των γνωρισμάτων και η κατάργηση δεδομένων.

8.1.4 Μοντέλα Ποιότητας/Χρησιμότητας

Το εργαλείο ARX επίσης παρέχει αρκετά μοντέλα που μπορούν να χρησιμοποιηθούν για τη βελτιστοποίηση των δεδομένων εξόδου κατά τη διαδικασία της ανωνυμοποίησης. Τα μοντέλα αυτά εξετάζουν κατά πόσο επηρεάζεται η ποιότητα των δεδομένων κατά την απώλεια πληροφορίας. Το εργαλείο βελτιστοποιεί την ποιότητα αυτή βασισμένο στις οντότητες ως υποκείμενα της ανωνυμοποίησης, στα γνωρίσματα τους ή σε ολόκληρες τις εγγραφές.

8.1.5 Ανάλυση Χρησιμότητας Και Ρίσκου

Το ARX προσφέρει έναν αρκετά αναλυτικό μηχανισμό εκτίμησης της χρησιμότητας και της ποιότητας των ανωνυμοποιημένων δεδομένων συγκρίνοντας τα δεδομένα εισόδου με αυτά της εξόδου μετά τη διαδικασία της ανωνυμοποίησης [61]. Μας δίνει πρόσβαση σε στατιστικά σχετικά με τα δεδομένα αυτά κατά πόσο εφαρμόστηκε κατάργηση ολόκληρων εγγραφών με βάση το μοντέλο που χρησιμοποιήθηκε. Στα πλαίσια αυτά παρουσιάζονται αρκετές πληροφορίες υπό τη

μορφή διαγραμμάτων προς ανάλυση καθώς και η ποσότητα της απολεσθείσας χρήσιμης πληροφορίας για την ανάγκη της ανωνυμοποίησης.

Συνεχίζοντας έχουμε τη ανάλυση του ρίσκου των ανωνυμοποιημένων δεδομένων. Αφορά το ρίσκο αναγνώρισης κάποιας εγγραφής μέσα από τον ανώνυμο πίνακα με βάση τα τρία μοντέλα επιθέσεων (prosecutor, journalist και marketer attacks) [62]. Μπορούμε να αναλύσουμε το ρίσκο που θα επιφέρει η προσέγγιση της ανωνυμοποίησης μέσω διαγραμμάτων όπως οι εγγραφές που έχουν το μεγαλύτερο ρίσκο αναγνώρισης καθώς και το συνολικό ρίσκο σε κάθε μοντέλο. Επίσης παρουσιάζει αναλυτικά το ελάχιστο ή το μέγιστο ρίσκο όπως και το ποσοστό των εγγραφών που επηρεάστηκαν από αυτό βασισμένα σε συγκεκριμένο μοντέλο επίθεσης.

8.2 Amnesia

Το Amnesia είναι ακόμα ένα εργαλείο ανωνυμοποίησης των δεδομένων το οποίο αποτελεί μια απλούστερη λύση συγκριτικά με το ARX που είδαμε παραπάνω [63]. Παρέχει μία πολύ πιο απλή υλοποίηση της ανωνυμοποίησης καθώς προσφέρει συγκεκριμένα μοντέλα απορρήτου καθώς και μεθόδους ανωνυμοποίησης. Είναι ιδιαίτερα εύκολο στη χρήση, προσφέροντας τη δημιουργία ιεραρχιών των γνωρισμάτων και ανωνυμοποίηση αυτών με βάση τα συγκεκριμένα μοντέλα.

8.2.1 Μέθοδοι Ανωνυμοποίησης

Όπως αναφέραμε το εργαλείο αυτό προσφέρει πολύ περιορισμένες μεθόδους ανωνυμοποίησης και αυτές είναι η γενίκευση, κάλυψη και η τυχαιοποίηση. Αφού τα δεδομένα «φορτωθούν» στο εργαλείο μας δίνεται η δυνατότητα να δημιουργήσουμε ιεραρχίες στα γνωρίσματα του συνόλου δεδομένων και από εκεί και πέρα να επιλέξουμε τη μέθοδο της ανωνυμοποίησης που θα χρησιμοποιήσουμε [64].

8.2.2 Μοντέλα Απορρήτου

Όσον αφορά τα μοντέλα απορρήτου, προσφέρει μοντέλα βασισμένα στην αποκάλυψη της ταυτότητας μόνο όπως η κ-ανωνυμία καθώς και μία πιο αδύναμη μορφή της, τη km-ανωνυμία, εφαρμόζοντας μετασχηματισμούς βασισμένους στις ιεραρχίες που δημιουργήθηκαν [64].

8.2.3 Ανάλυση

Τέλος και αφού η ανωνυμοποίηση έχει εφαρμοστεί στα δεδομένα μας, τα αποτελέσματα παρουσιάζονται στατιστικά μέσω γραφημάτων από όπου μπορούν να αναλυθούν αργότερα.

8.3 PyCanon

Το PyCanon αποτελεί μια βιβλιοθήκη που έχει αναπτυχθεί σε γλώσσα Python στοχεύοντας στον έλεγχο κα την αξιολόγηση του επιπέδου ανωνυμίας σε ένα σύνολο δεδομένων [65].

8.3.1 Προσέγγιση

Κυριότερος στόχος του εργαλείου αυτού είναι να ελέγξει και να επιβεβαιώσει την απόδοση της ανωνυμοποίησης. Για παράδειγμα μπορούμε να ανωνυμοποιήσουμε τα δεδομένα μας με το ARX και να εισάγουμε στο PyCanon το νέο σύνολο δεδομένων επιβεβαιώνοντας αν η ανωνυμοποίηση επιτεύχθηκε βάση των αποτελεσμάτων του ARX. Πρακτικά εισάγουμε το σύνολο δεδομένων που λάβαμε ως αποτέλεσμα της ανωνυμοποίησης στο εργαλείο, στη συνέχεια χωρίζουμε τα γνωρίσματα μας σε αναγνωριστικά, ψευδο-αναγνωριστικά και εμπιστευτικά. Συνεχίζοντας ορίζουμε τα μοντέλα απορρήτου καθώς και τις τιμές που αποδώσαμε στο κάθε μοντέλο. Τέλος παράγουμε μία αναφορά που περιγράφει τα αποτελέσματά μας [66].

8.3.2 Μοντέλα Απορρήτου

Τα μοντέλα απορρήτου που υποστηρίζει το PyCanon είναι τα k-anonymity, (α, k)-anonymity, l-diversity, entropy και recursive (c, l)-diversity, t-closeness καθώς και basic β-likeness και δ-disclosure για την αποδοτικότερη προσέγγιση ανωνυμοποίησης του συνόλου δεδομένων, ορίζοντας τα γνωρίσματα του συνόλου αυτού ανάλογα σε αναγνωριστικά, ψευδο-αναγνωριστικά και ευαίσθητα.

8.3.3 Ανάλυση

Κατά την ολοκλήρωση της ανωνυμοποίησης το εργαλείο παράγει μία αναφορά σε μορφή pdf, αρχείο JSON ή εμφάνιση στην οθόνη μέσω του command line, παρουσιάζοντας το επίπεδο ανωνυμίας που επιτεύχθηκε με τις παραπάνω μεθόδους, τις κλάσεις ισοδυναμίας καθώς και τα

αποτελέσματα την ανωνυμοποίησης βάση της πιθανότητας αναγνώρισης κάποιας εγγραφής μέσα στο σύνολο.

8.4 Σύνοψη

Εξετάζοντας τα παραπάνω εργαλεία επιλέγουμε το ARX για την ανωνυμοποίηση των δεδομένων μας. Γενικά το εργαλείο αυτό προσφέρει μια μεγάλη γκάμα επιλογών τόσο ως προς τα μοντέλα απορρήτου και τις μεθόδους ανωνυμοποίησης όσο και στην ανάλυση των αποτελεσμάτων σχετικά με την χρησιμότητα των δεδομένων αλλά και το ρίσκο αποκάλυψης κατά τα διάφορα μοντέλα επιθέσεων. Επίσης παρέχει αρκετούς τύπους ιεράρχησης των γνωρισμάτων όπως και μετασχηματισμούς τους, προσαρμόζοντας την ανωνυμοποίηση των δεδομένων στις ανάγκες τις έρευνάς μας. Τέλος είναι αρκετά φιλικό προς το χρήστη και αποτελεί μία ιδανική επιλογή για τις ανάγκες του έργου μας.

Κεφάλαιο 9
Ανωνυμοποίηση - Εφαρμογή

Στο σημείο αυτό και αφού έχουμε επιλέξει το εργαλείο μας, έχοντας εξετάσει τις διαφορετικές προσεγγίσεις της ανωνυμοποίησης που μπορούμε να ακολουθήσουμε, θα εφαρμόσουμε τεχνικές ανωνυμοποίησης στα δεδομένα υγείας. Όπως είδαμε και σε προηγούμενα κεφάλαια, κατά τη διαδικασία μπορούν να αξιοποιηθούν διαφορετικά μοντέλα απορρήτου, μετασχηματισμοί, όπως και μέθοδοι ανωνυμοποίησης, προσφέροντας η κάθε περίπτωση διαφορετικά αποτελέσματα. Τα αποτελέσματα αυτά αφορούν την απώλεια της χρήσιμης πληροφορίας, όσο και το ρίσκο αποκάλυψης προσώπων ή γνωρισμάτων και αυτό είναι κάτι που θα εξετάσουμε στο παρόν κεφάλαιο προσεγγίζοντας με διαφορετικούς τρόπους τη διαδικασία της ανωνυμοποίησης.

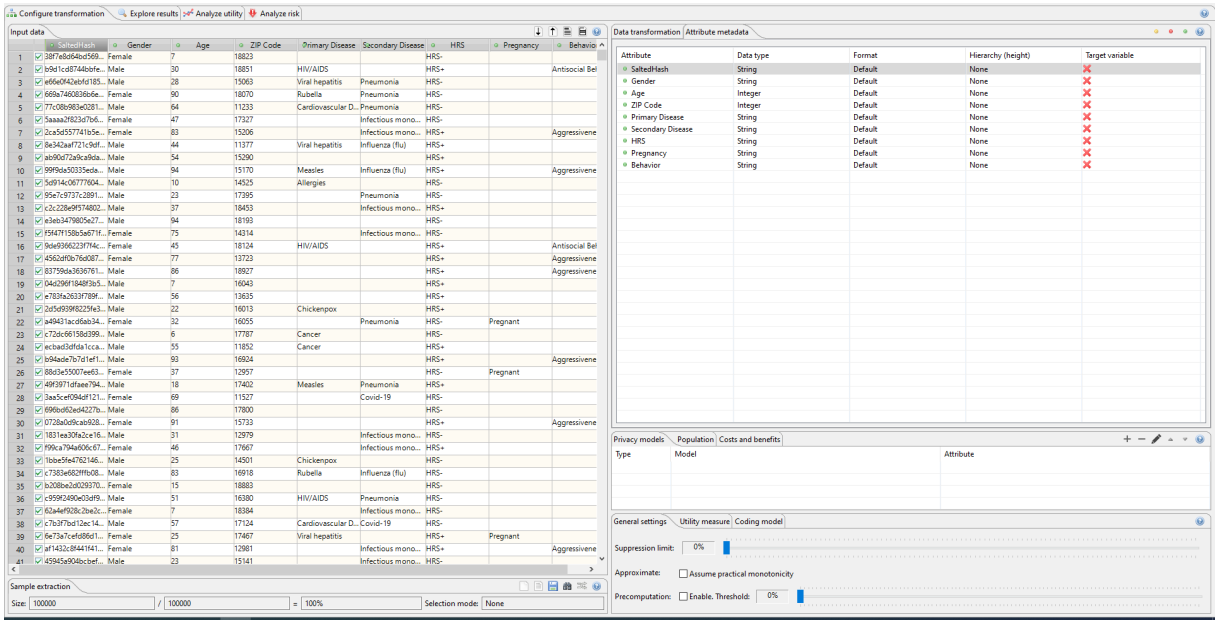
9.1 Προετοιμασία Δεδομένων

Αρχικά θα πρέπει να προετοιμάσουμε τα δεδομένα που δημιουργήσαμε σε προηγούμενο κεφάλαιο εισάγοντάς τα στο εργαλείο ARX. Συνεχίζοντας, ακολουθούμε μία σειρά ενεργειών για τη προετοιμασία τους όπως η κατηγοριοποίηση των γνωρισμάτων ως αναγνωριστικά, ψευδο-

αναγνωριστικά, εμπιστευτικά και μη εμπιστευτικά, η δημιουργία ιεραρχιών των ψευδο-αναγνωριστικών ανάλογα το είδος τους, η απόδοση βαρύτητας τους καθορίζοντας το ποσοστό της απώλειας που επιθυμούμε σε κάθε ψευδο-αναγνωριστικό καθώς και η μέγιστη απώλεια χρήσιμης πληροφορίας που επιθυμούμε να ορίσουμε.

9.1.1 Εισαγωγή Δεδομένων

Στο πρώτο στάδιο όπως αναφέραμε το σύνολο των δεδομένων υγείας που δημιουργήσαμε και φυλάσσονται σε μορφή excel θα πρέπει να εισαχθεί στο εργαλείο ARX όπως φαίνεται στη εικόνα 9.1.

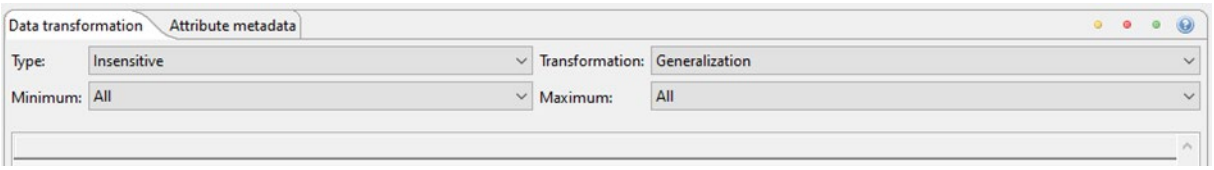


Εικόνα 9.1: ARX - Εισαγωγή δεδομένων υγείας

Όπως μπορούμε να δούμε η διεπαφή του ARX είναι χωρισμένη σε διαφορετικά παράθυρα όπου το κάθε ένα αποτελείται από συγκεκριμένες λειτουργίες. Επίσης βλέπουμε ότι στο πάνω μέρος της έχουμε τέσσερις καρτέλες που είναι η διαμόρφωση μετασχηματισμού, η εξερεύνηση αποτελεσμάτων, η ανάλυση χρησιμότητας και η ανάλυση ρίσκου. Ξεκινώντας με τη πρώτη καρτέλα (διαμόρφωση μετασχηματισμού) στην αριστερή στήλη του εργαλείου έχουμε τα δεδομένα μας όπως διαμορφώθηκαν κατά τη διαδικασία της ψευδωνυμοποίησης, περιέχοντας το ψευδώνυμο καθώς και τα υπόλοιπα γνωρίσματα της κάθε εγγραφής.

Η δεξιά στήλη στο πάνω μέρος περιλαμβάνει τα γνωρίσματα, ορίζοντας τον τύπο του κάθε ενός καθώς και το μετασχηματισμό των δεδομένων από όπου μπορούμε να κατηγοριοποιήσουμε τα

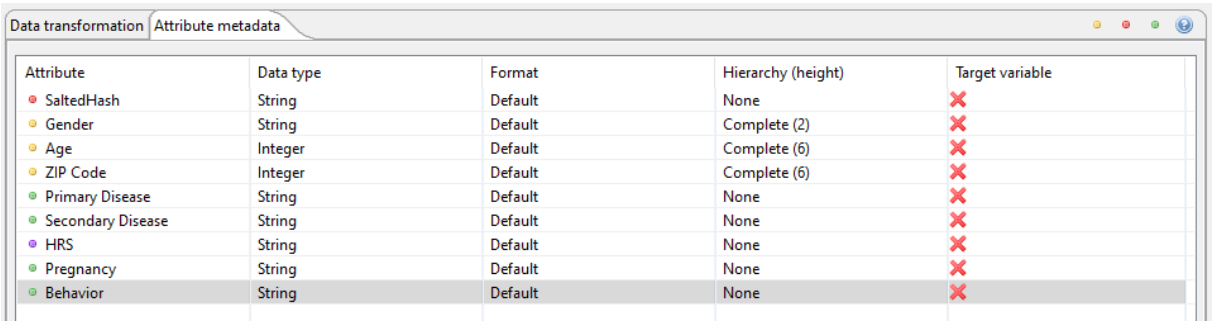
δεδομένα μας σε αναγνωριστικά, ψευδο-αναγνωριστικά, εμπιστευτικά ή μη-εμπιστευτικά και να επιλέξουμε τη μέθοδο μετασχηματισμού τους όπως βλέπουμε και στην εικόνα 9.2 παρακάτω. Όλοι οι μετασχηματισμοί μας θα ακολουθήσουν τη μέθοδο της γενίκευσης.



Εικόνα 9.2: ARX - Τύποι Και Μετασχηματισμοί Γνωρισμάτων

Όπως αναφέραμε και σε προηγούμενο κεφάλαιο σχετικά με τη προσέγγιση της ανωνυμοποίησης, αρχικά έχουμε τρία ψευδο-αναγνωριστικά που θα συμπεριληφθούν στη διαδικασία (ηλικία, φύλο και ταχυδρομικός κώδικας). Επίσης προκειμένου να εξετάσουμε διαφορετικές προσεγγίσεις της διαδικασίας της ανωνυμοποίησης και κατά πόσο η κάθε μία από αυτές προσφέρει καλύτερα αποτελέσματα έναντι κάποιας άλλης σχετικά με την αποκάλυψη ταυτότητας ή γνωρισμάτων καθώς το ποσοστό απώλειας χρήσιμης πληροφορίας, αποφασίσαμε αρχικά να ορίσουμε ένα μόνο από τα ευαίσθητα γνωρίσματα ως εμπιστευτικό και τα υπόλοιπα ως μη εμπιστευτικά δοκιμάζοντας διάφορες προσεγγίσεις και εξετάζοντας σε κάθε μία από αυτές τα αποτελέσματα που λαμβάνουμε. Συνεχίζοντας θα κατηγοριοποιηθούν και άλλα από τα ευαίσθητα γνωρίσματα ως εμπιστευτικά και αναλύοντάς τα σκοπεύουμε να καταλήξουμε στο καλύτερα συνδυασμό απώλειας πληροφορίας/ρίσκου αποκάλυψης.

Όπως βλέπουμε και στην εικόνα 9.3 μετά τη κατηγοριοποίηση των γνωρισμάτων, με κόκκινο χρώμα έχουμε τα αναγνωριστικά, με κίτρινο τα ψευδο-αναγνωριστικά, με μωβ τα εμπιστευτικά και με πράσινο τα μη εμπιστευτικά γνωρίσματα. Επίσης θα δούμε ότι κάποια από τα γνωρίσματα του αρχικού πίνακα δεν συμπεριλήφθηκαν στην ανωνυμοποίηση καθώς δεν επηρεάζουν ή επηρεάζονται από τη νόσο που εξετάζουμε δημιουργώντας περισσότερη σύγχυση στη διαδικασία της ανωνυμοποίησης χωρίς να προσφέρουν όφελος σε αυτή.

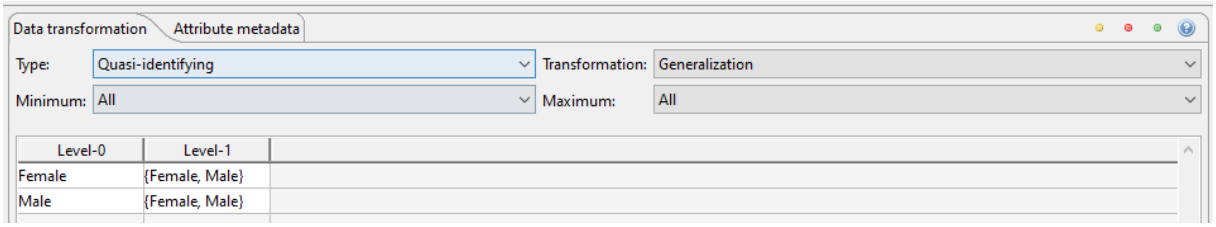


Εικόνα 9.3: ARX - Κατηγοριοποίηση γνωρισμάτων

9.1.2 Δημιουργία Ιεραρχιών

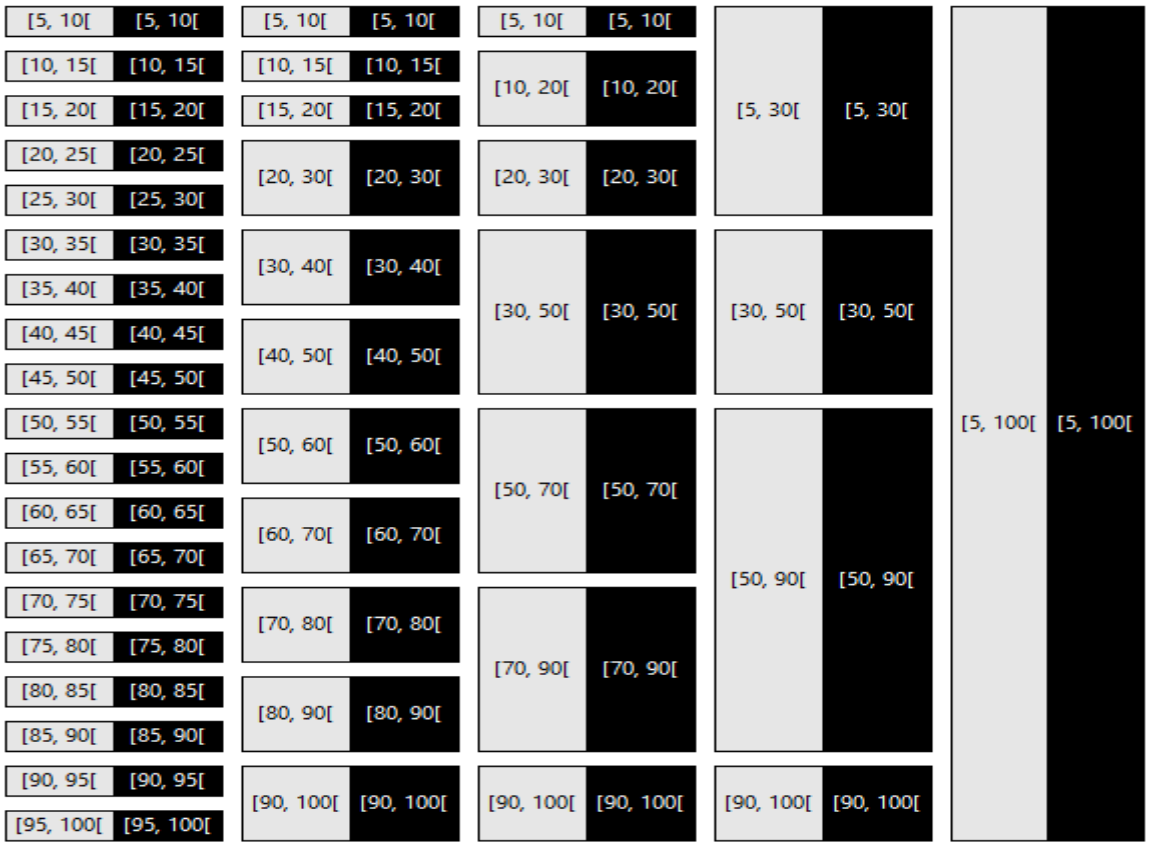
Συνεχίζοντας, μίας και έχουμε ολοκληρώσει την εισαγωγή των δεδομένων μας και έχοντας τα κατηγοριοποιήσει βάσει των τεσσάρων τύπων όπως είδαμε παραπάνω, θα πρέπει να δημιουργήσουμε ιεραρχίες στα ψευδο-αναγνωριστικά καθορίζοντας έτσι τα επίπεδα βάθους της γενίκευσης που θα εφαρμοστεί στο καθένα. Η ιεράρχηση των ψευδο-αναγνωριστικών θα βοηθήσει ώστε να πετύχουμε έναν ιδανικό συνδυασμό στον κάθε μετασχηματισμό που θα επιλέξουμε εφαρμόζοντας διαφορετικό επίπεδο γενίκευσης (generalization) ή κάλυψης (masking) σε κάθε γνώρισμα πετυχαίνοντας ένα υψηλό επίπεδο προστασίας των δεδομένων και ελαχιστοποιώντας όσο το δυνατό την απώλεια πληροφορίας.

Το πρώτο ψευδο-αναγνωριστικό είναι το φύλο του ασθενούς, που περιλαμβάνει δύο επιλογές (αρσενικού ή θηλυκού). Έχοντας δύο πιθανές επιλογές το μέγιστο επίπεδο γενίκευσης που μπορούμε να επιτύχουμε είναι ένα (Level 1) όπως μπορούμε να δούμε στην εικόνα 9.4.



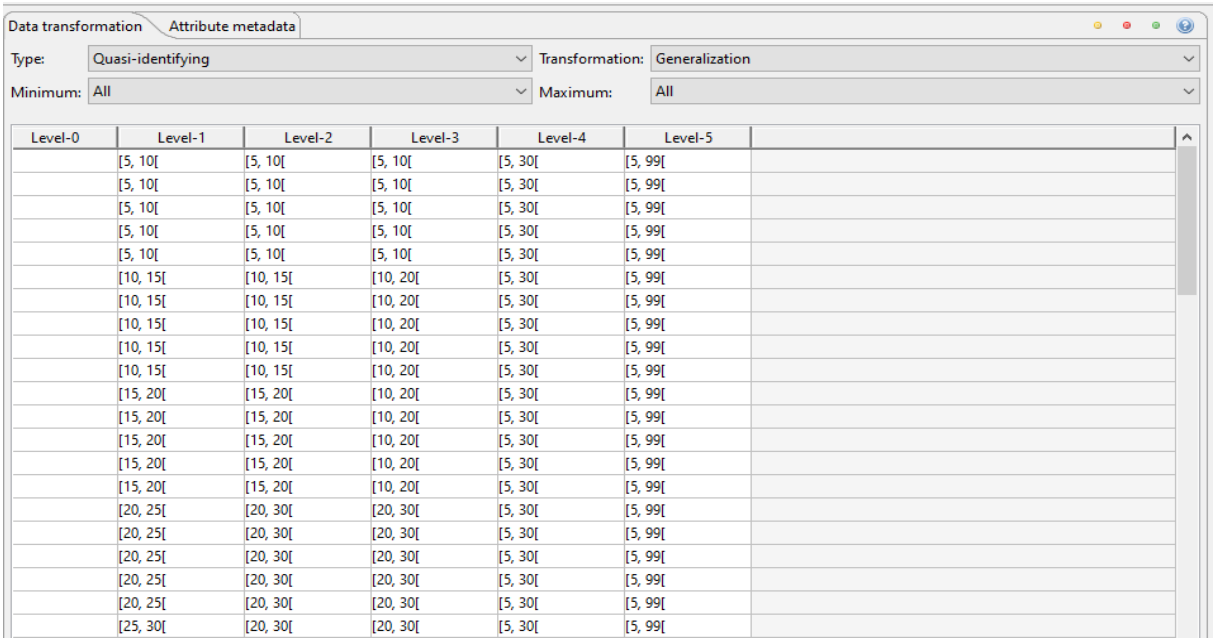
Εικόνα 9.4: ARX - Γενίκευση Φύλου

Το επόμενο ψευδο-αναγνωριστικό είναι η ηλικία του ασθενούς. Θα επιλέξουμε πέντε επίπεδα γενίκευσης όπως μπορούμε να δούμε και από την εικόνα 9.5 ξεκινώντας από ένα εύρος πενταετίας, γενικεύοντας ακόμα περισσότερο σε κάθε επόμενο επίπεδο.



Εικόνα 9.5: ARX - Γενίκευση Ηλικίας

Στην εικόνα 9.6 βλέπουμε πώς διαμορφώνονται τα πέντε επίπεδα γενίκευσης της ηλικίας.


Data transformation | Attribute metadata

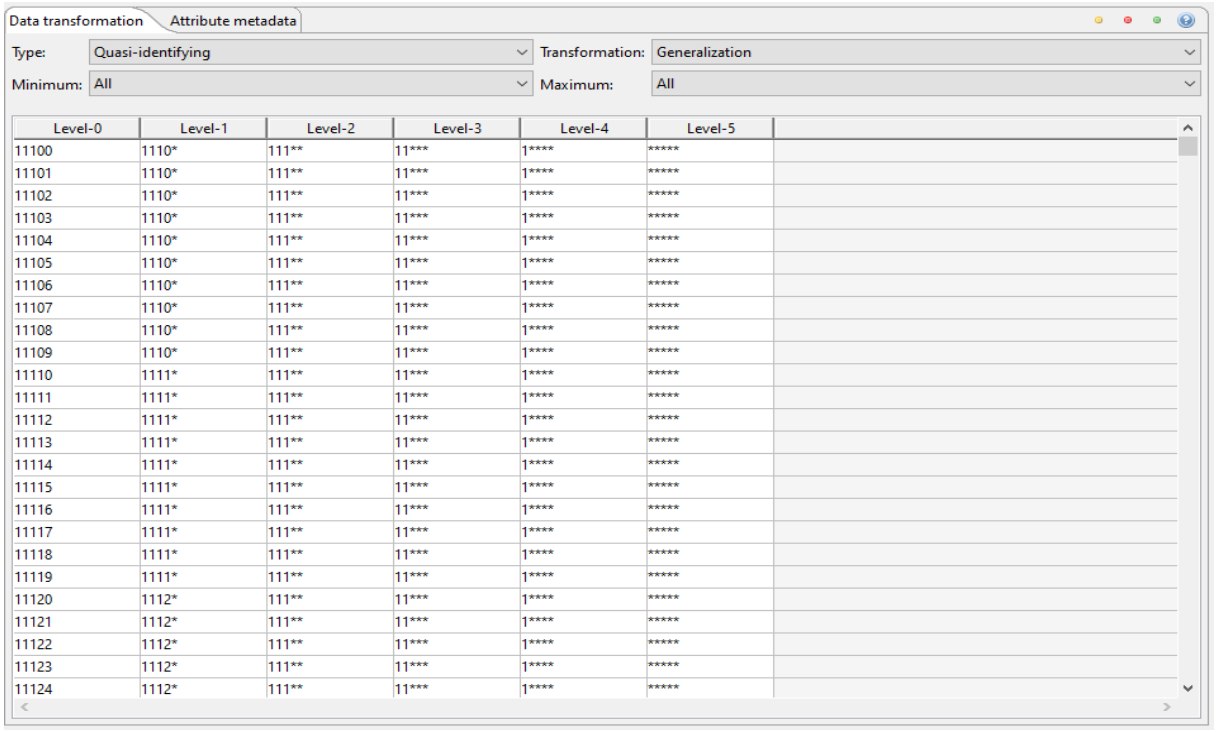
Type: Quasi-identifying | Transformation: Generalization

Minimum: All | Maximum: All

Level-0	Level-1	Level-2	Level-3	Level-4	Level-5
	[5, 10[	[5, 10[	[5, 10[	[5, 30[	[5, 99[
	[5, 10[	[5, 10[	[5, 10[	[5, 30[	[5, 99[
	[5, 10[	[5, 10[	[5, 10[	[5, 30[	[5, 99[
	[5, 10[	[5, 10[	[5, 10[	[5, 30[	[5, 99[
	[5, 10[	[5, 10[	[5, 10[	[5, 30[	[5, 99[
	[10, 15[	[10, 15[	[10, 20[	[5, 30[	[5, 99[
	[10, 15[	[10, 15[	[10, 20[	[5, 30[	[5, 99[
	[10, 15[	[10, 15[	[10, 20[	[5, 30[	[5, 99[
	[10, 15[	[10, 15[	[10, 20[	[5, 30[	[5, 99[
	[10, 15[	[10, 15[	[10, 20[	[5, 30[	[5, 99[
	[15, 20[	[15, 20[	[10, 20[	[5, 30[	[5, 99[
	[15, 20[	[15, 20[	[10, 20[	[5, 30[	[5, 99[
	[15, 20[	[15, 20[	[10, 20[	[5, 30[	[5, 99[
	[15, 20[	[15, 20[	[10, 20[	[5, 30[	[5, 99[
	[15, 20[	[15, 20[	[10, 20[	[5, 30[	[5, 99[
	[20, 25[	[20, 30[	[20, 30[	[5, 30[	[5, 99[
	[20, 25[	[20, 30[	[20, 30[	[5, 30[	[5, 99[
	[20, 25[	[20, 30[	[20, 30[	[5, 30[	[5, 99[
	[20, 25[	[20, 30[	[20, 30[	[5, 30[	[5, 99[
	[20, 25[	[20, 30[	[20, 30[	[5, 30[	[5, 99[
	[25, 30[	[20, 30[	[20, 30[	[5, 30[	[5, 99[

Εικόνα 9.6: ARX - Επίπεδα Γενίκευσης Ηλικίας

Τέλος έχουμε ένα ακόμα ψευδο-αναγνωριστικό που είναι ο ταχυδρομικός κώδικας κατοικίας του ασθενούς. Στη περίπτωση αυτή θα επιλέξουμε τη κάλυψη ως μέθοδο δημιουργώντας πέντε επίπεδα κάλυψης του Τ.Κ. όπως εμφανίζονται στην εικόνα 9.7 παρακάτω. Όπως μπορούμε να δούμε κάθε επίπεδο κάλυψης αποκρύβει ένα από τα τελευταία νούμερα του Τ.Κ.



Level-0	Level-1	Level-2	Level-3	Level-4	Level-5
11100	1110*	111**	11***	1****	*****
11101	1110*	111**	11***	1****	*****
11102	1110*	111**	11***	1****	*****
11103	1110*	111**	11***	1****	*****
11104	1110*	111**	11***	1****	*****
11105	1110*	111**	11***	1****	*****
11106	1110*	111**	11***	1****	*****
11107	1110*	111**	11***	1****	*****
11108	1110*	111**	11***	1****	*****
11109	1110*	111**	11***	1****	*****
11110	1111*	111**	11***	1****	*****
11111	1111*	111**	11***	1****	*****
11112	1111*	111**	11***	1****	*****
11113	1111*	111**	11***	1****	*****
11114	1111*	111**	11***	1****	*****
11115	1111*	111**	11***	1****	*****
11116	1111*	111**	11***	1****	*****
11117	1111*	111**	11***	1****	*****
11118	1111*	111**	11***	1****	*****
11119	1111*	111**	11***	1****	*****
11120	1112*	111**	11***	1****	*****
11121	1112*	111**	11***	1****	*****
11122	1112*	111**	11***	1****	*****
11123	1112*	111**	11***	1****	*****
11124	1112*	111**	11***	1****	*****

Εικόνα 9.7: ARX - Κάλυψη Τ. Κ.

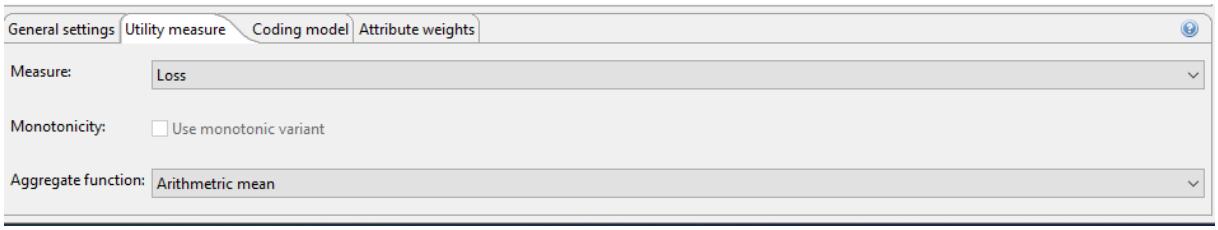
9.1.3 Διατήρηση Χρησιμότητας Δεδομένων

Συνεχίζοντας τη διαδικασία προετοιμασίας των δεδομένων μας πριν την ανωνυμοποίησή τους θα πρέπει να εξετάσουμε το επίπεδο διατήρησης της χρησιμότητας των δεδομένων μας έναντι της αποκάλυψης τους από επικείμενες επιθέσεις [67]. Αρχικά μέσα από το ARX όπως μπορούμε να δούμε (εικόνα 9.8) ότι μας δίνεται η δυνατότητα να επιλέξουμε το όριο κατάργησης ολόκληρων εγγραφών που επιθυμούμε να εφαρμόσουμε για τις ανάγκες της ανωνυμοποίησης.

Εικόνα 9.8: ARX – Όριο Κατάργησης Εγγραφών

Ορίσαμε το όριο αυτό στο 100% ως μέγιστο ποσοστό κατάργησης που είναι ανεκτό για τις ανάγκες της έρευνάς μας, ώστε να μας δώσει όλους τους πιθανούς μετασχηματισμούς προσεγγίζοντας τον αποτελεσματικότερο για τις ανάγκες της έρευνάς μας.

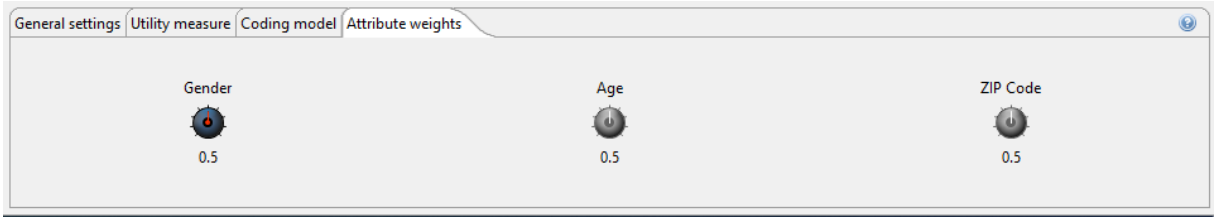
Συνεχίζοντας, στην επόμενη καρτέλα θα αφήσουμε το measure ως έχει στη προκαθορισμένη τιμή του, ενώ θα αλλάξουμε την επιλογή των αθροιστικών συναρτήσεων σε arithmetic mean. Οι συναρτήσεις αυτές χρησιμοποιούνται για τη συγκέντρωση των εκτιμήσεων που λαμβάνονται για μεμονωμένα γνωρίσματα ενός συνόλου δεδομένων σε μία συνολική τιμή [68].



Εικόνα 9.9: ARX – Μέτρο αναφορικά με την Χρησιμότητα των δεδομένων

Τέλος μία ακόμα δυνατότητα που μας δίνει το ARX είναι να δώσουμε ανάλογη βαρύτητα στα ψευδο-αναγνωριστικά μας, ορίζοντας έτσι αυτά στα οποία επιθυμούμε να δοθεί περισσότερο ή λιγότερο βάρος κατά την ανωνυμοποίηση διατηρώντας έτσι σε μεγαλύτερο βαθμό την ακεραιότητά τους. Οι τιμές που λαμβάνει το βάρος αυτό είναι από 0 έως 1 με προκαθορισμένη τιμή στο 0,5, ενώ αυξάνοντας τη τιμή αυτή για συγκεκριμένο γνώρισμα, περιορίζουμε την απώλεια πληροφορία για αυτό. Για να ορίσουμε το βάρος των ψευδο-αναγνωριστικών θα πρέπει να καθορίσουμε τις ανάγκες της έρευνας για την οποία θα χρησιμοποιηθούν τα ανώνυμα δεδομένα, και ανάλογα να εστιάσουμε στα ψευδο-αναγνωριστικά αυτά που θέλουμε να επηρεαστεί λιγότερο κατά τη διαδικασία [68].

Αποφασίσαμε ότι και τα τρία ψευδο-αναγνωριστικά μας έχουν τη ίδια σημασία για τις ανάγκες της έρευνας που ακολουθούμε και έτσι έλαβαν το ίδιο βάρος μεταξύ τους, χωρίς να δίνεται ιδιαίτερη προτεραιότητα σε κάποιο από αυτά.



Εικόνα 9.10: ARX – Απόδοση Βάρους Γνωρισμάτων

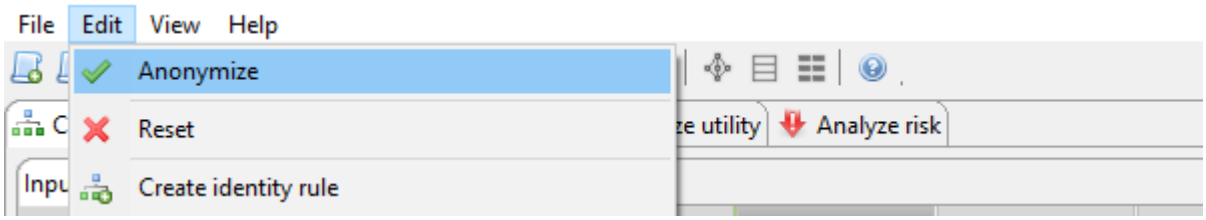
9.2 Ανωνυμοποίηση Δεδομένων

Έχοντας ολοκληρώσει το στάδιο της προετοιμασίας μέσα από το ARX, στο σημείο αυτό θα ξεκινήσουμε την ανωνυμοποίηση. Όπως έχουμε αναφέρει ήδη θα ακολουθήσουμε διαφορετικές προσεγγίσεις, καταγράφοντας τα αποτελέσματα τους, ώστε να έχουμε μία σύγκριση μεταξύ των προσεγγίσεων αυτών.

9.2.1 Προσεγγίσεις Ανωνυμοποίησης

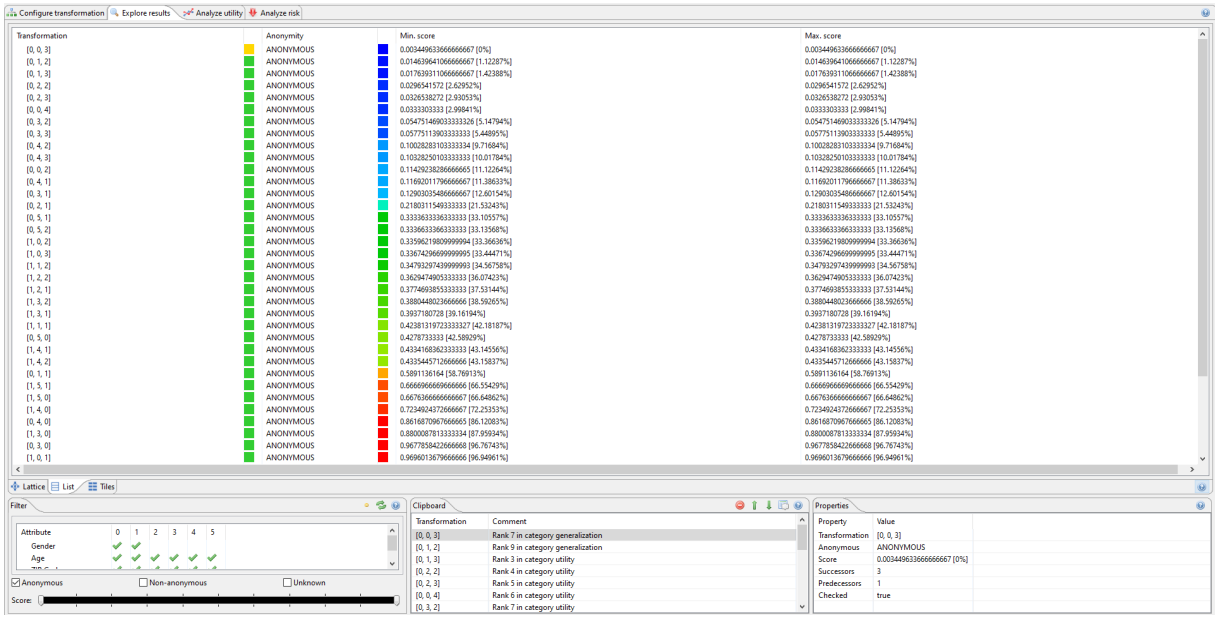
Στη πρώτη προσέγγιση της ανωνυμοποίησης, έχουμε επιλέξει μόνο τη νόσο HRS ως εμπιστευτικό γνώρισμα από τη προηγούμενη ενότητα. Στο σημείο αυτό θα επιλέξουμε τα μοντέλα απορρήτου που θα χρησιμοποιήσουμε αρχικά και είναι η k-ανωνυμία για τιμή k = 5 και l-διαφορετικότητα για τιμή l = 2. Για την l-διαφορετικότητα να τονίσουμε ότι η μέγιστη τιμή που μπορεί να πάρει το l είναι 2 καθώς το εμπιστευτικό γνώρισμα μπορεί να πάρει μόνο δύο τιμές (HRS+ ή HRS-). Όπως μπορούμε να δούμε στην εικόνα 9.11 εφαρμόζουμε τα μοντέλα απορρήτου και επιλέγουμε (εικόνα 9.12) ώστε να εκτελεστεί η ανωνυμοποίηση.

Εικόνα 9.11: ARX – Μοντέλα Απορρήτου

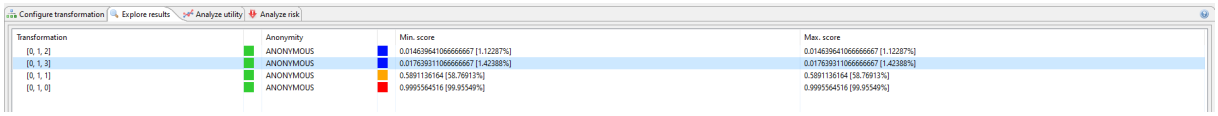


Εικόνα 9.12: ARX – Εκτέλεση Ανωνυμοποίησης

Αφού ολοκληρωθεί η διαδικασία, επιλέγουμε τη δεύτερη καρτέλα «εξερεύνηση αποτελεσμάτων», από όπου μπορούμε να ορίσουμε το μετασχηματισμό που θα εφαρμόσουμε. Παρατηρώντας την εικόνα 9.13 μπορούμε στο πάνω παράθυρο να δούμε όλους τους μετασχηματισμούς που μας δίνει το ARX και να επιλέξουμε βάσει της γενίκευσης που θέλουμε να δώσουμε στο κάθε ψευδο-αναγνωριστικό, τη συνολική απώλεια που είναι ανεκτή για το σύνολο των δεδομένων μας και το ρίσκο που επιφέρει κάθε μετασχηματισμός.



Εικόνα 9.13: ARX – Μετασχηματισμοί Ψευδο-Αναγνωριστικών

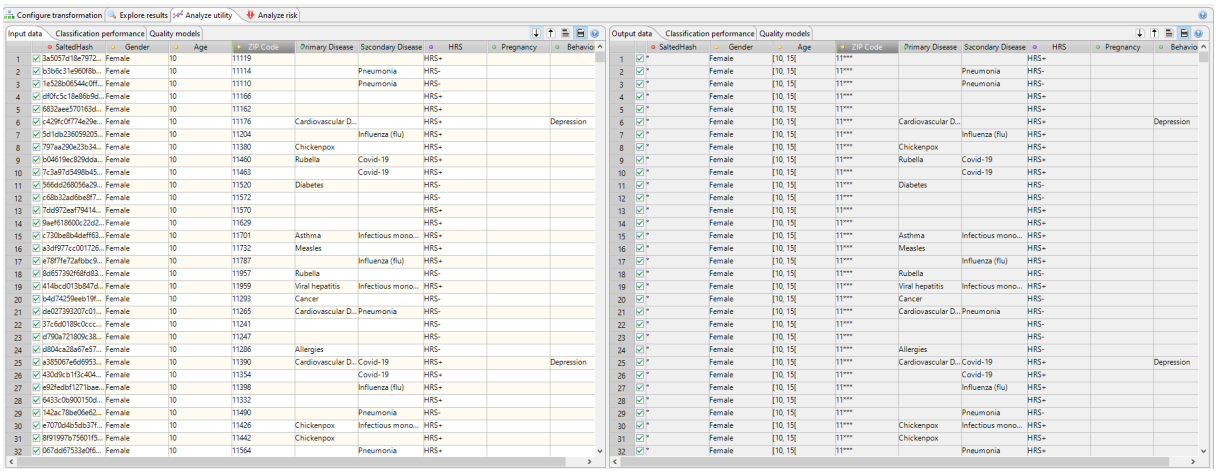


Εικόνα 9.14: ARX – Επιλογή Μετασχηματισμού

Ο μετασχηματισμός που εφαρμόζουμε αρχικά είναι [0, 1, 3] όπου ο κάθε αριθμός εκπροσωπεί το επίπεδο γενίκευσης ή κάλυψης του κάθε ψευδο-αναγνωριστικού στο σύνολο των δεδομένων μας.

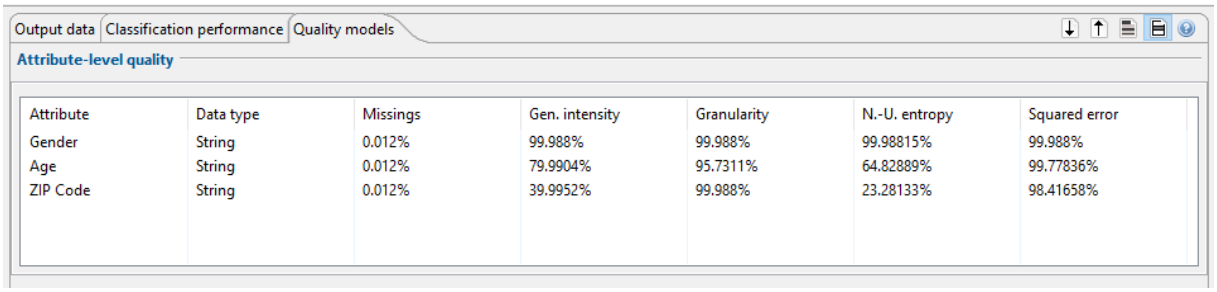
Ο λόγος που αρχικά επιλέξαμε τον μετασχηματισμό αυτό για τα δεδομένα μας έναντι των υπολοίπων είναι ότι για τις ανάγκες της έρευνάς μας κρατάει το φύλο ανεπηρέαστο, καθώς επίσης κρατάει την ηλικία του ασθενούς σε επίπεδο γενίκευσης πενταετίας κάτι που επαρκεί για τις ανάγκες της έρευνάς μας. Όσο για τον ΤΚ αφαιρεί τα τρία τελευταία του ψηφία κάτι που δεν επηρεάζει σε πολύ μεγάλο βαθμό την έρευνα, ωστόσο θα εξετάσουμε και άλλους χρήσιμους μετασχηματισμούς.

Στη τρίτη καρτέλα «ανάλυση χρησιμότητας», αφού έχουμε εφαρμόσει το μετασχηματισμό μας μπορούμε να δούμε, τα ανώνυμα πλέον δεδομένα όπως έχουν διαμορφωθεί και έχουν εξαχθεί από την ανωνυμοποίηση.



Εικόνα 9.15: ARX – Ανώνυμα Δεδομένα

Διερευνώντας την ανάλυση χρησιμότητας, μπορούμε να δούμε τα μοντέλα ποιότητας στα τρία ψευδο-αναγνωριστικά των ανώνυμων δεδομένων. Το κύριο κομμάτι που μας ενδιαφέρει (εικόνα 9.16) είναι η απώλεια πληροφορίας για το κάθε ψευδο-αναγνωριστικό που είναι στο 0,012%.

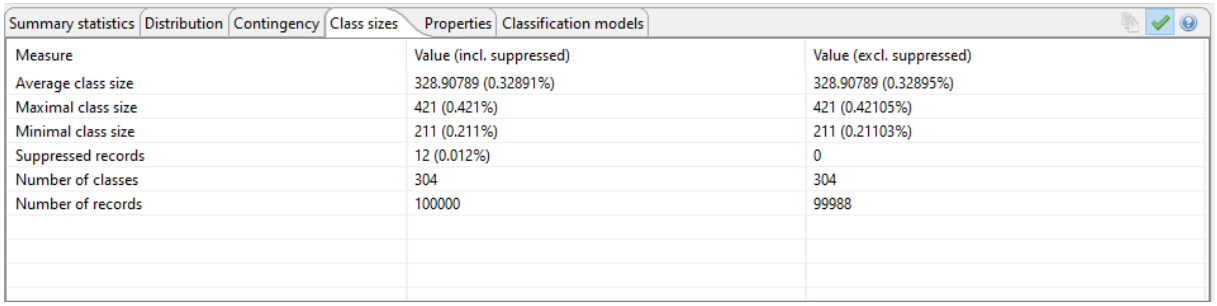

Output data | Classification performance | Quality models

Attribute-level quality

Attribute	Data type	Missings	Gen. intensity	Granularity	N.-U. entropy	Squared error
Gender	String	0.012%	99.988%	99.988%	99.98815%	99.988%
Age	String	0.012%	79.9904%	95.7311%	64.82889%	99.77836%
ZIP Code	String	0.012%	39.9952%	99.988%	23.28133%	98.41658%

Εικόνα 9.16: ARX – Απώλεια Πληροφορίας Ψευδο-Αναγνωριστικών

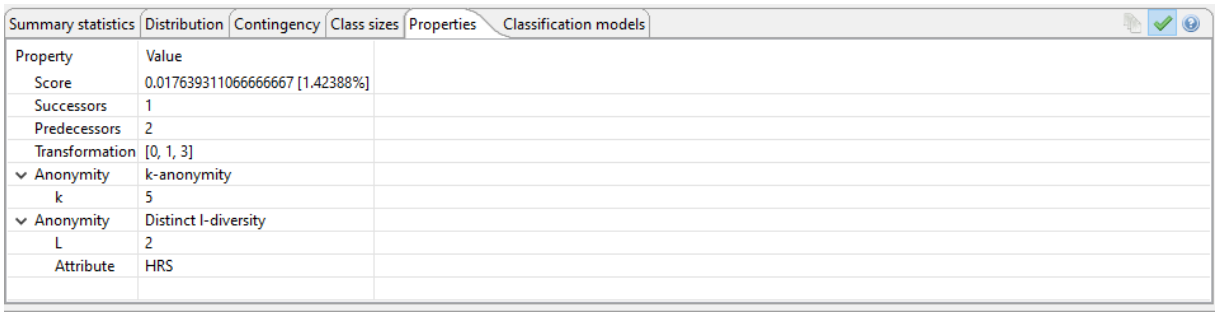
Μέσα από την ανάλυση της χρησιμότητα βλέπουμε επίσης το μέγεθος των κλάσεων ισοδυναμίας (μέσο, ελάχιστο, μέγιστο) καθώς και τον αριθμό των κλάσεων που δημιουργήθηκαν από την ανωνυμοποίηση. Ακόμη μπορούμε να δούμε τις εγγραφές που καταργήθηκαν για τις ανάγκες επίτευξης την ανωνυμοποίησης οι οποίες είναι δώδεκα, που είναι ένα αρκετά χαμηλό ποσοστό (0.012%). Για τη συγκεκριμένη προσέγγιση δεν υπάρχει λόγος να χρησιμοποιήσουμε άλλη τιμή για k καθώς η μικρότερη κλάση ισοδυναμίας που δημιουργείται είναι 211 εγγραφές.



Measure	Value (incl. suppressed)	Value (excl. suppressed)
Average class size	328.90789 (0.32891%)	328.90789 (0.32895%)
Maximal class size	421 (0.421%)	421 (0.42105%)
Minimal class size	211 (0.211%)	211 (0.21103%)
Suppressed records	12 (0.012%)	0
Number of classes	304	304
Number of records	100000	99988

Εικόνα 9.17: ARX – Διαμόρφωση Κλάσεων

Τέλος η καρτέλα 'properties' μας δίνει τη βαθμολογία του μετασχηματισμού που αφορά το συνολικό ποσοστό απώλειας πληροφορίας στα δεδομένα μας, καθώς και άλλες λεπτομέρειες όπως το μετασχηματισμό, τα μοντέλα απορρήτου που εφαρμόσαμε και τις τιμές για k και l.

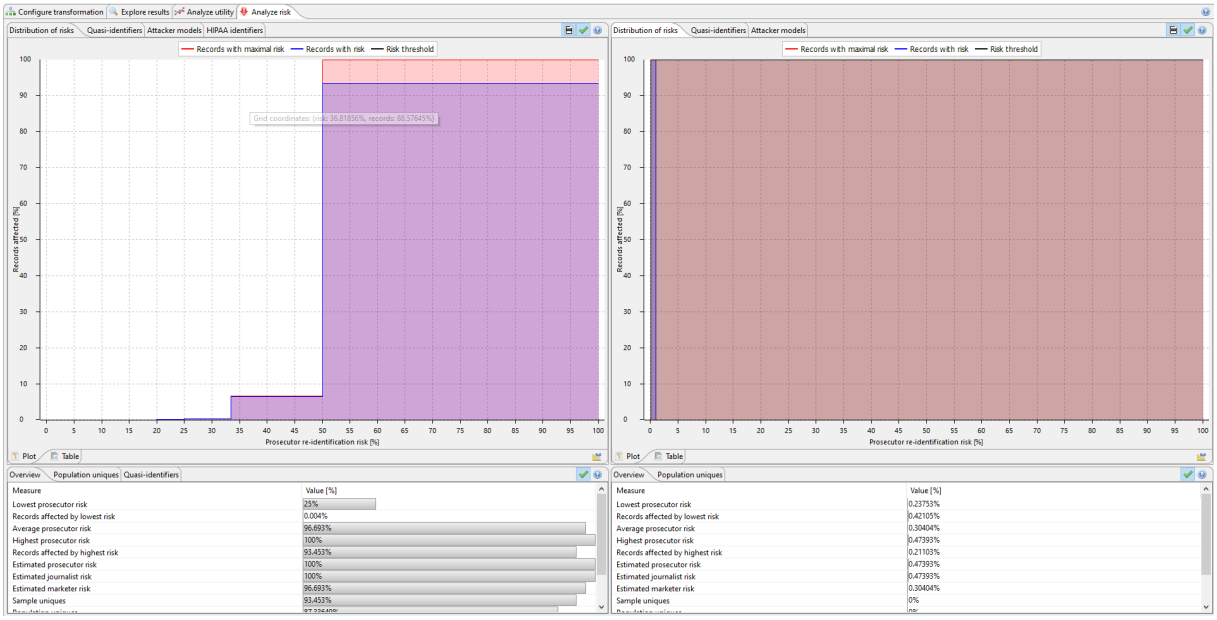


Property	Value
Score	0.017639311066666667 [1.42388%]
Successors	1
Predecessors	2
Transformation	[0, 1, 3]
Anonymity	k-anonymity
k	5
Anonymity	Distinct l-diversity
L	2
Attribute	HRS

Εικόνα 9.18: ARX – Ιδιότητες Μετασχηματισμού

Στο σημεία αυτό είναι καλό να αναφερθεί ότι η επιλογή του μετασχηματισμού αποτελεί κρίση του υπεύθυνου επεξεργασίας, καθώς αλλάζοντας τον μετασχηματισμό, αλλάζουμε και το βάθος της γενίκευσης του κάθε ψευδο-αναγνωριστικού, εστιάζοντας έτσι σε διαφορετικές πτυχές την έρευνα. Από την άλλη βέβαια, διαμορφώνεται διαφορετικά η χρησιμότητα των δεδομένων καθώς και το ρίσκο αποκάλυψής τους.

Επιστρέφοντας στην αρχική προσέγγιση μπορούμε να δούμε παρακάτω στην εικόνα 9.19, πώς διαμορφώνεται η διανομή του ρίσκου στα ανώνυμα δεδομένα μας με το μετασχηματισμό που επιλέξαμε. Όπως μπορούμε να δούμε σε μορφή γραφήματος παρουσιάζεται το ρίσκο αποκάλυψης των δεδομένων πριν και μετά την ανωνυμοποίηση. Μπορούμε να διαπιστώσουμε ότι το ρίσκο μετά την ανωνυμοποίηση είναι σε πολύ χαμηλά επίπεδα, κάτω του 0,5% για το μέγιστο ρίσκο.

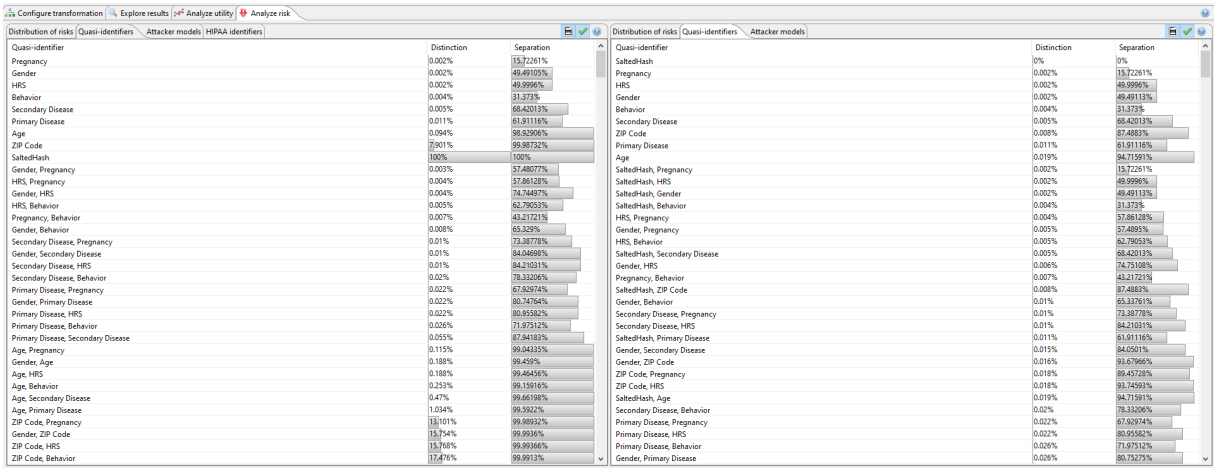


Εικόνα 9.19: ARX – Διανομή Ρίσκου

Επίσης μέσω του ARX έχουμε μία ακόμα καρτέλα παρουσίασης του ρίσκου για τα τρία μοντέλα επιτιθέμενων (prosecutor, journalist, marketer) καθώς και τα ποσοστά των εγγραφών που έχουν ρίσκο αποκάλυψης, το μέγιστο ρίσκο και το ποσοστό επιτυχίας μίας επίθεσης για κάθε ένα από τα μοντέλα αυτά [24] (τα εν λόγω μοντέλα έχουν περιγραφεί αναλυτικά στο κεφάλαιο 2). Όπως μπορούμε να διαπιστώσουμε (εικόνα 9.20) το μέγιστο ρίσκο είναι αρκετά χαμηλό στο 0,47%, κάτι που σε συνδυασμό με τη χαμηλή απώλεια πληροφορίας στο σύνολο το δεδομένων καθιστά μία πολύ καλή αρχική προσέγγιση.

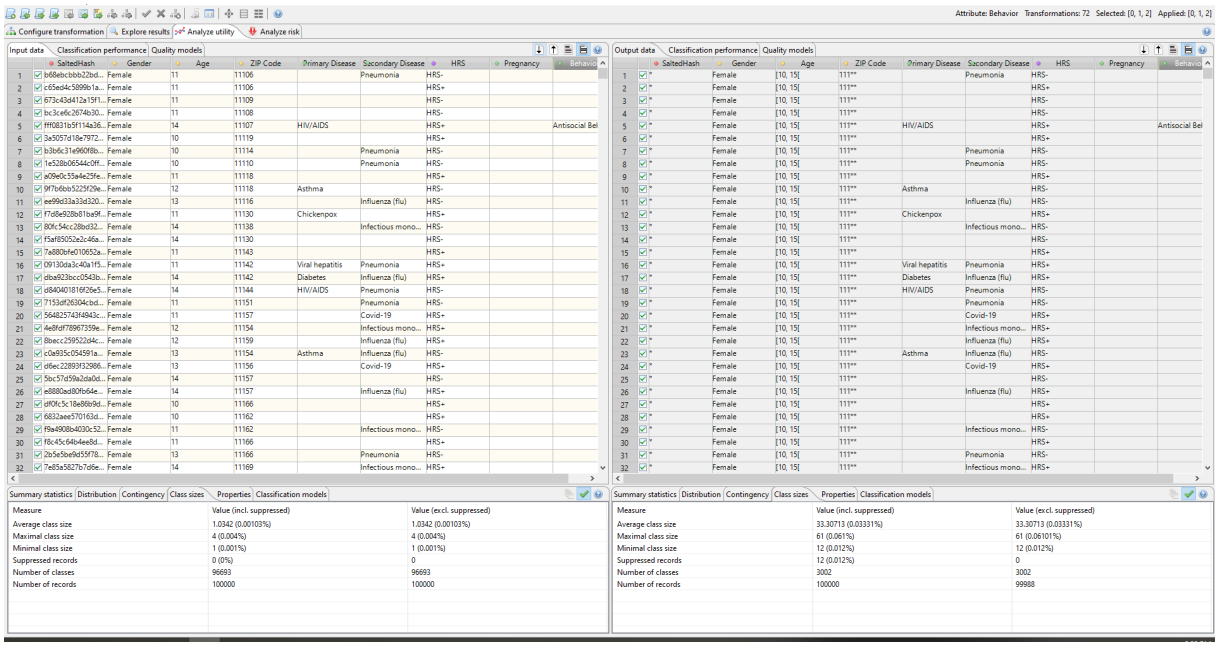
Εικόνα 9.20: ARX – Ρίσκο Αποκάλυψης

Τέλος μπορούμε να εξετάσουμε πώς διαμορφώνονται τα ποσοστά των γνωρισμάτων πριν και μετά την ανωνυμοποίηση. Παρατηρώντας την εικόνα 9.21 καταγράφεται για κάθε γνώρισμα αλλά και συνδυασμό γνωρισμάτων, το ποσοστό διαχωρισμού του, κάτι που επηρεάζεται από τις διαφορετικές τιμές που λαμβάνει το κάθε γνώρισμα.



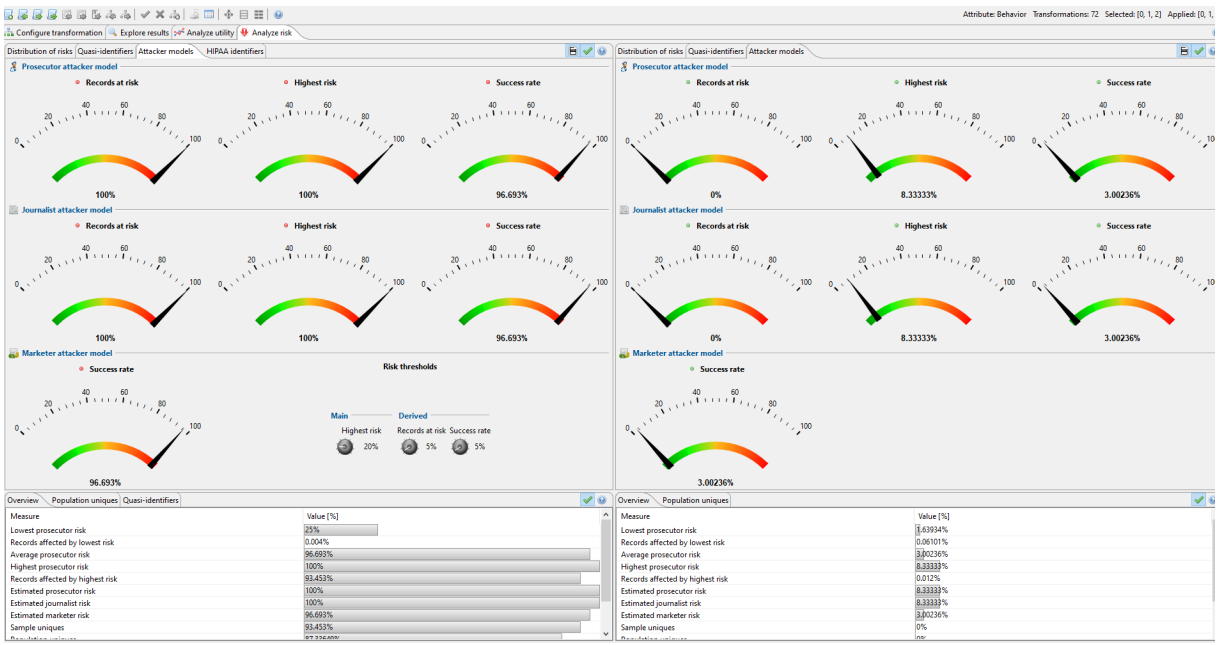
Εικόνα 9.21: ARX – Διαχωρισμός Γνωρισμάτων

Ένας ακόμα μετασχηματισμός που θα μπορούσε να φανεί πολύ χρήσιμος και αξίζει να σημειωθεί ως και προς τον ΤΚ για την έρευνά μας είναι ο [0, 1, 2] έχοντας μικρότερη απώλεια πληροφορίας, αφού διατηρεί τρία από τα ψηφία του ΤΚ, δημιουργώντας αρκετά μικρότερες κλάσεις ισοδυναμίας (εικόνα 9.22).



Εικόνα 9.22: ARX – Εξαχθέντα Δεδομένα

Ο μετασχηματισμός αυτός δημιουργεί αρκετά μεγαλύτερο μέγιστο ρίσκο αποκάλυψης στο σύνολο των δεδομένων και πιο συγκεκριμένα 8,33% έναντι 0,47%, ωστόσο οι εγγραφές που επηρεάζοντας από το μέγιστο ρίσκο όπως φαίνεται στην εικόνα κάτω δεξιά είναι μόλις στο 0.012%. Η προσέγγιση αυτή θα εξεταστεί περαιτέρω στη συνέχεια, καθώς μπορεί να δώσει πολύτιμα συμπεράσματα για την έρευνά μας, ως προς τη γεωγραφική επέκταση της επιδημίας, το ηλικιακό εύρος και το φύλο των ασθενών.



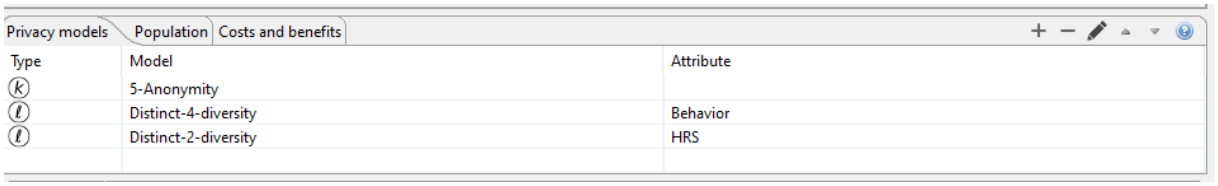
Εικόνα 9.23: ARX – Ανάλυση Ρίσκου Αποκάλυψης

Για ανωνυμία τάξης 10 δεν επωφελούμαστε ως προς τη διατήρηση της χρησιμότητα των δεδομένων αλλά ούτε και ως προς το ρίσκο αποκάλυψής τους καθώς σε κάθε μετασχηματισμό η μικρότερη κλάση ισοδυναμία που συναντήσαμε είναι δώδεκα εγγραφές και για αυτό δεν θα επεκταθούμε περαιτέρω.

Συνεχίζοντας με τη επόμενη προσέγγιση της ανωνυμοποίησης αποφασίσαμε να συμπεριλάβουμε ένα ακόμα ευαίσθητο γνώρισμα, αυτό της συμπεριφοράς του ασθενούς ως εμπιστευτικό και να εξετάσουμε πώς επηρεάζει τα έως τώρα δεδομένα μας.

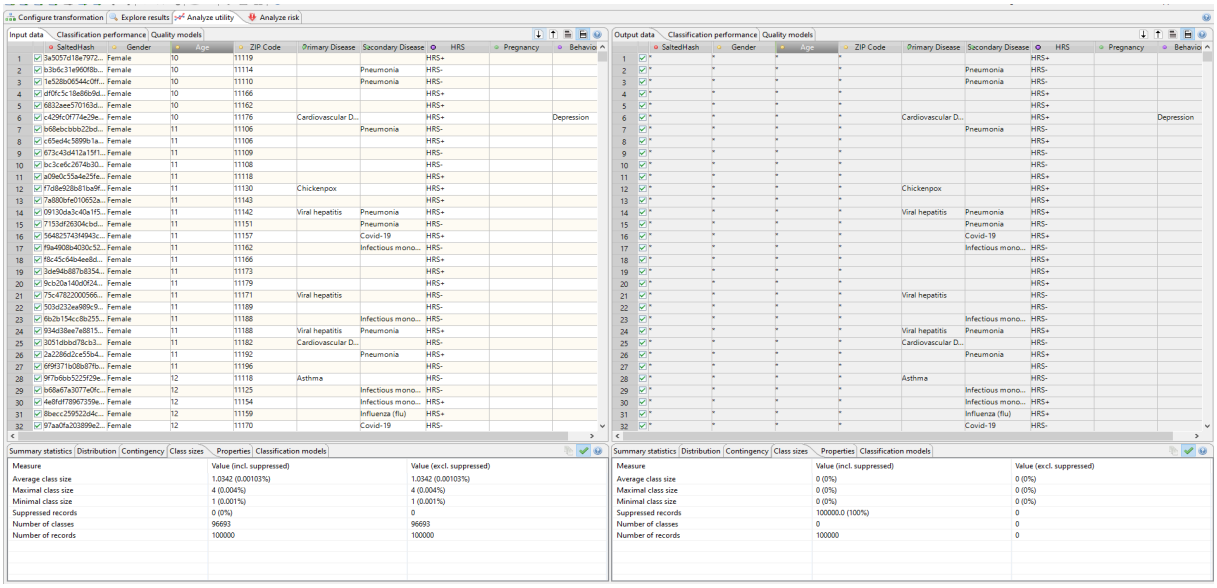
Για τις ανάγκες της έρευνάς μας ιδανικά στοχεύουμε να κρατήσουμε το φύλο του ασθενούς ανεπηρέαστο, χωρίς γενίκευση καθώς και την ηλικιακή ομάδα του σε ένα εύρος πενταετίας θεωρώντας ότι έτσι θα λάβουμε πολύτιμα συμπεράσματα για την νόσο HRS, διατηρώντας τη ποιότητα των δεδομένων μας σε υψηλά επίπεδα. Όσο για τον ΤΚ για τη γεωγραφική επέκταση της νόσου επιζητάμε ένα μέγιστο βάθος γενίκευσης 2ης ή 3ης τάξης σε κάποιες περιπτώσεις, διατηρώντας τουλάχιστον δύο ή τρία ψηφία του ανεπηρέαστα.

Το γνώρισμα της συμπεριφορά λαμβάνει τέσσερις διαφορετικές τιμές και έτσι η πρώτη απόπειρά μας, θα λάβει τιμή l=4 για το γνώρισμα αυτό.



Εικόνα 9.24: ARX – Μοντέλα Απορρήτου

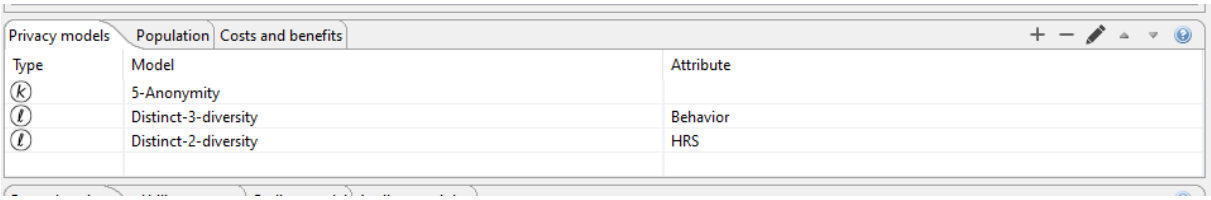
Ωστόσο αυτό που θα παρατηρήσουμε είναι ότι με αυτές τις τιμές εμφανίζονται τα πρώτα προβλήματα στα δεδομένα μας. Προκειμένου να κρατήσουμε χαμηλή γενίκευση στα δεδομένα μας, διατηρώντας χρήσιμα συμπεράσματα, για τις τιμές αυτές στα μοντέλα απορρήτου έχουμε πλήρη κατάργηση των δεδομένων μας σε όλους τους χρήσιμους μετασχηματισμούς.



Εικόνα 9.25: ARX – Πλήρης Κατάργηση Εγγραφών

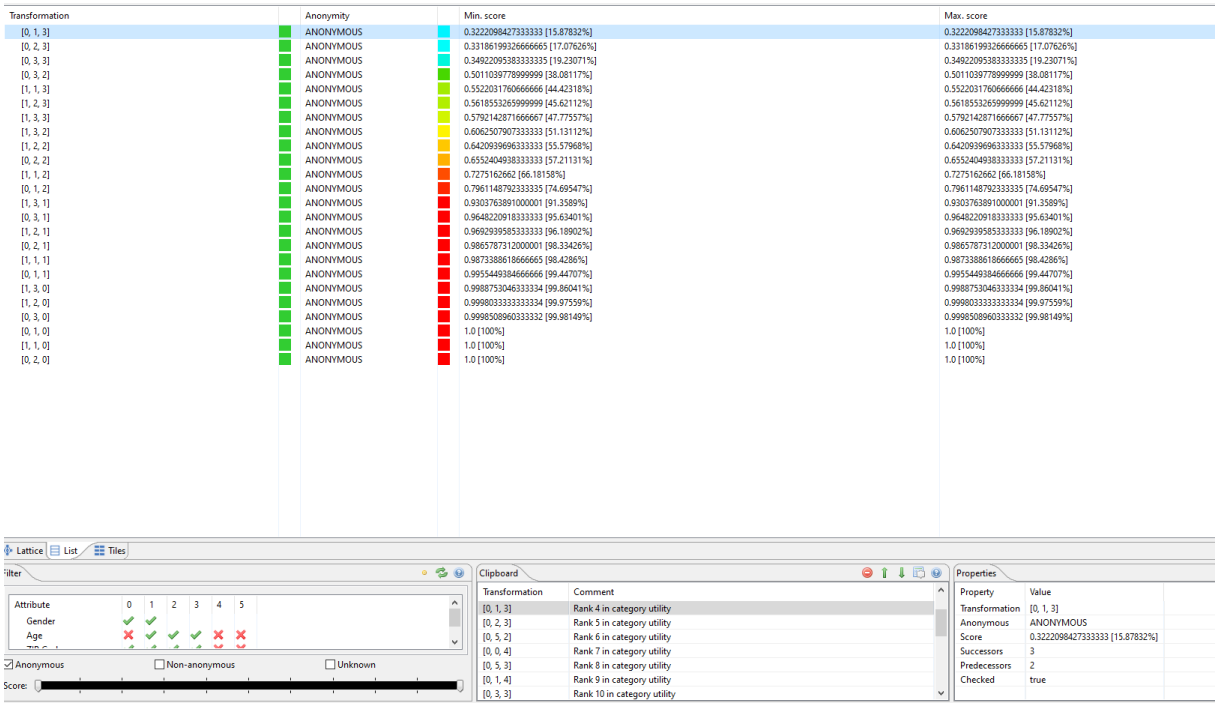
Το γεγονός αυτό οφείλεται στο πολύ μικρό ποσοστό αλλαγής της συμπεριφοράς των εξεταζόμενων ασθενών που αδυνατεί να δημιουργήσει ποικιλομορφία 4ης τάξης στο γνώρισμα 'behavior', έτσι απορρίπτουμε τη προσέγγιση αυτή.

Συνεχίζοντας δοκιμάζουμε μικρότερη τιμή για l στη συμπεριφορά του ασθενούς και συγκεκριμένα για l=3 αναμένοντας καλύτερα αποτελέσματα.



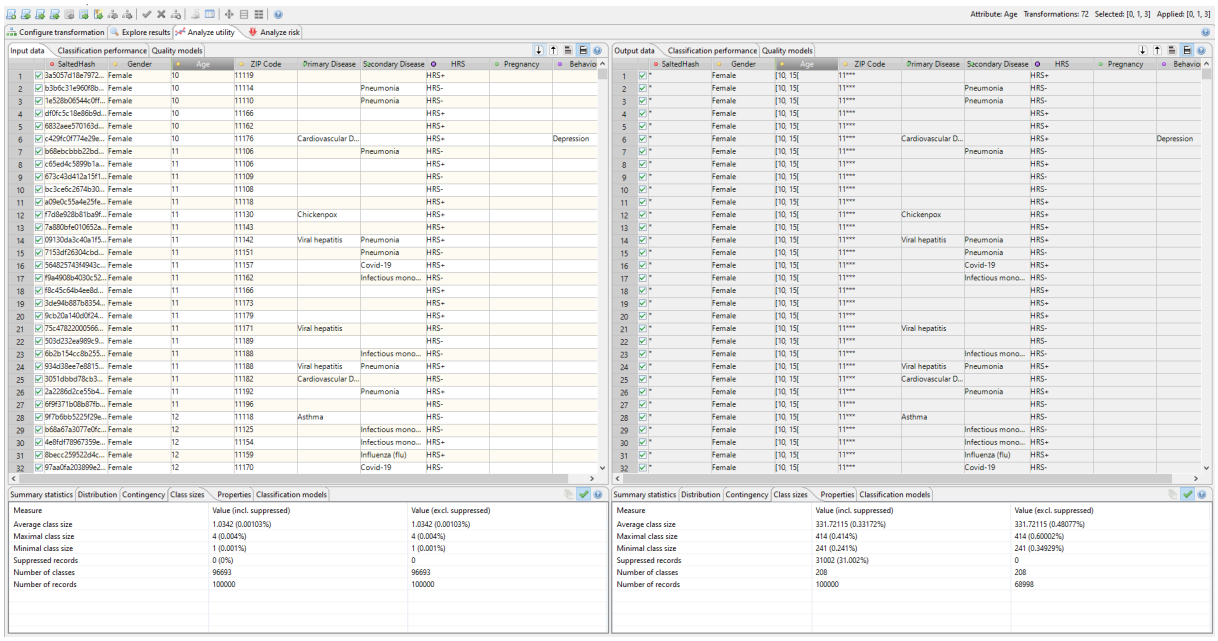
Εικόνα 9.26: ARX – Μοντέλα Απορρήτου

Προκειμένου να κρατήσουμε τα δεδομένα μας όσο πιο κοντά στην αρχική τους μορφή, θα δούμε ότι αρκετοί από τους μετασχηματισμούς επιφέρουν μεγάλο ποσοστό γενίκευσης που συνεπάγεται και απώλεια στα δεδομένα μας με το καλύτερο μετασχηματισμό όσον αφορά τη γενίκευση να ξεκινάει με απώλεια πληροφορίας στο 15%.



Εικόνα 9.27: ARX – Επιλογή Μετασχηματισμού

Εξετάζοντας τα ανώνυμα δεδομένα μας για το συγκεκριμένο μετασχηματισμό [0, 1, 3] βλέπουμε ότι δημιουργεί αρκετά μεγάλο ποσοστό κατάργησης εγγραφών, άνω του 30% από τη μία, αλλά από την άλλη, λόγω του μεγάλου αρχικού δείγματος, 68998 εγγραφές που απομένουν μπορούν να δώσουν σαφή αποτελέσματα για την έρευνά μας. Οι κλάσεις ισοδυναμίας παραμένουν στα ίδια περίπου επίπεδα καθώς και το ρίσκο παραμένει πολύ κοντά στα αποτελέσματα της πρώτης προσέγγισης αφού επηρεάζεται από το μέγεθος της μικρότερης κλάσης και τη χρήση του ίδιου μετασχηματισμού.



Εικόνα 9.28: ARX – Εξαχθέντα Δεδομένα

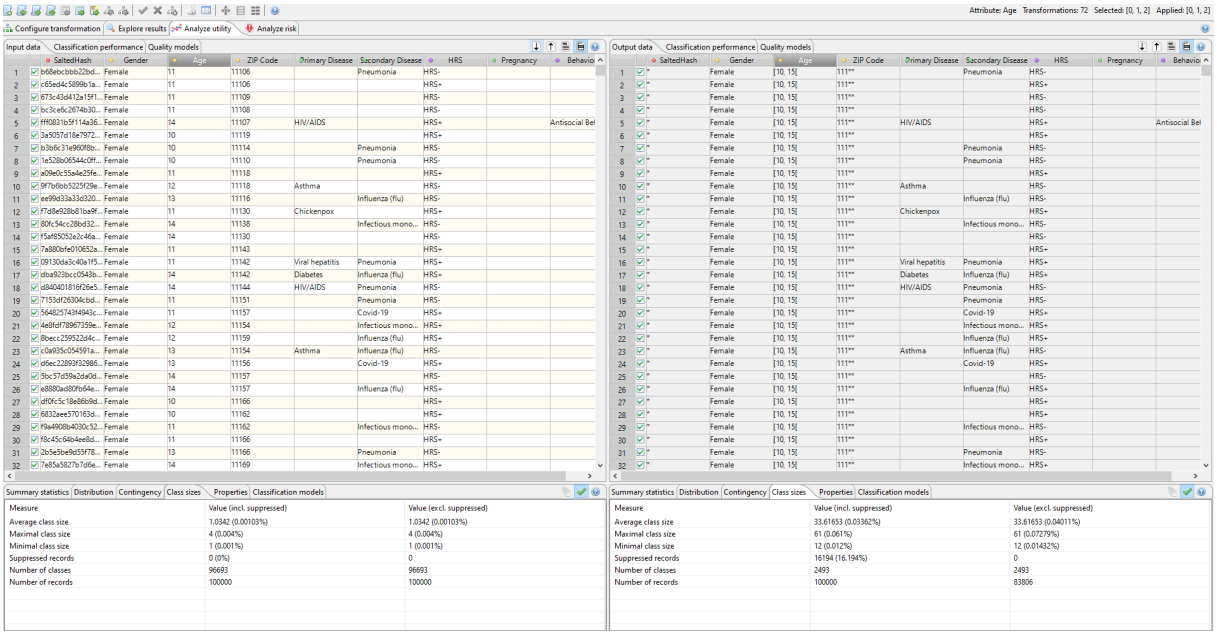
Εικόνα 9.29: ARX – Ανάλυση Ρίσκου

Για οποιοδήποτε άλλο μετασχηματισμό θα είχαμε πολύ μεγαλύτερη γενίκευση στα ψευδο-αναγνωριστικά μας κάτι που δεν επιθυμούμε καθώς αλλοιώνονται αρκετά από τα αρχικά δεδομένα ή δημιουργούμε μεγάλο ποσοστό κατάργησης εγγραφών, άνω του 50% στο σύνολο των δεδομένων μας έως και 80% σε κάποιες περιπτώσεις. Τα ποσοστά αυτά, τόσο της γενίκευσης αλλά και της απώλειας πληροφορίας δεν είναι ανεκτά για τις ανάγκες μας και δεν θα επεκταθούμε.

Για τιμή l=2 στο εμπιστευτικό γνώρισμα της συμπεριφοράς, λαμβάνουμε ακριβώς τα ίδια αποτελέσματα με την πρώτη μας προσέγγιση για το μετασχηματισμό [0, 1, 3]. Καθώς το μέγεθος της μικρότερης κλάσης ισοδυναμίας παραμένει το ίδιο, έχοντας 211 εγγραφές, κάθε κλάση

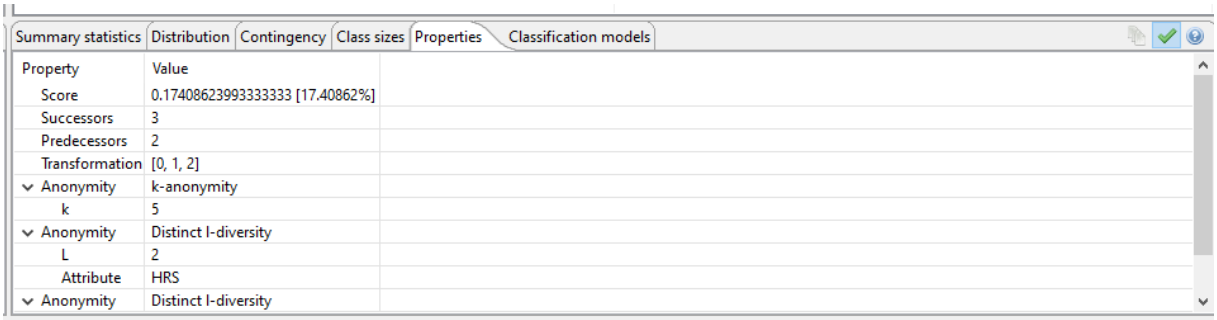
ισοδυναμία περιέχει από τη δημιουργία των δεδομένων τουλάχιστον δύο διαφορετικές τιμές για το γνώρισμα 'behavior'.

Διαφοροποιώντας το μετασχηματισμό σε [0, 1, 2] ώστε να επωφεληθούμε και από τον ΤΚ, έχουμε αρχικά τη κατάργηση των εγγραφών να διαμορφώνεται στο 16,19%, παρατηρώντας ότι έχει δημιουργήσει πολύ περισσότερες καταργήσεις εγγραφών ορίζοντας και το 'behavior' ως εμπιστευτικό γνώρισμα συγκριτικά με τη προσέγγιση που είχε μόνο τη νόσο HRS ως εμπιστευτικό γνώρισμα.



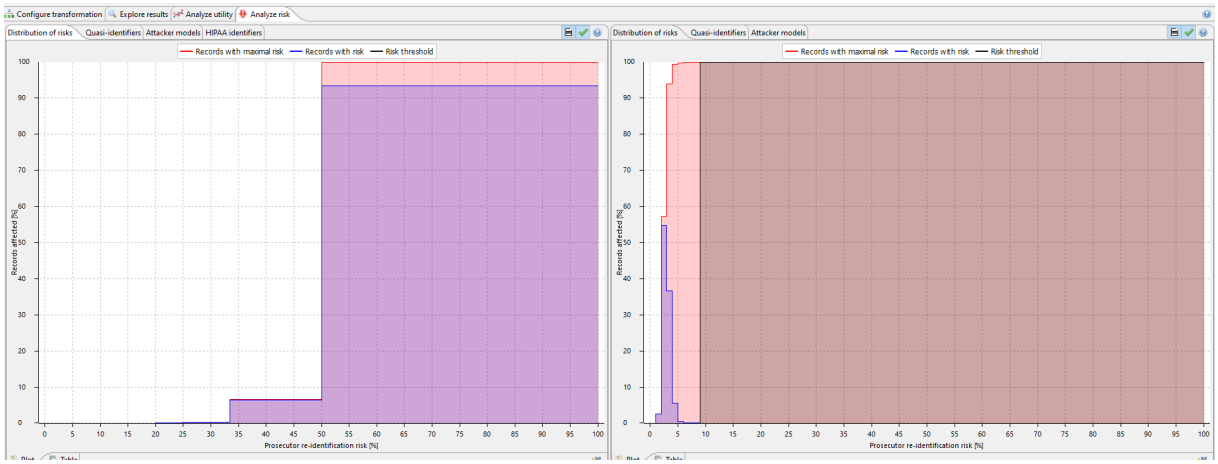
Εικόνα 9.30: ARX – Εξαχθέντα Δεδομένα

Όσον αφορά την απώλεια πληροφορίας γενικότερα με το μετασχηματισμό αυτό, βλέπουμε ότι είναι στο 17.4%, κάτι που θα μπορούσε να κριθεί ως ικανοποιητικό για τις ανάγκες μας, από τη σκοπιά της διατήρησης χαμηλής γενίκευσης στα δεδομένα.



Εικόνα 9.31: ARX – Ιδιότητες Μετασχηματισμού

Περνώντας στην ανάλυση του ρίσκου για το συγκεκριμένο μετασχηματισμό παρατηρούμε μέσω του γραφήματος πώς κατανέμεται το ρίσκο στα δεδομένα μας για τις εγγραφές με ρίσκο, αυτές με μέγιστο ρίσκο και το κατώφλι του ρίσκου.



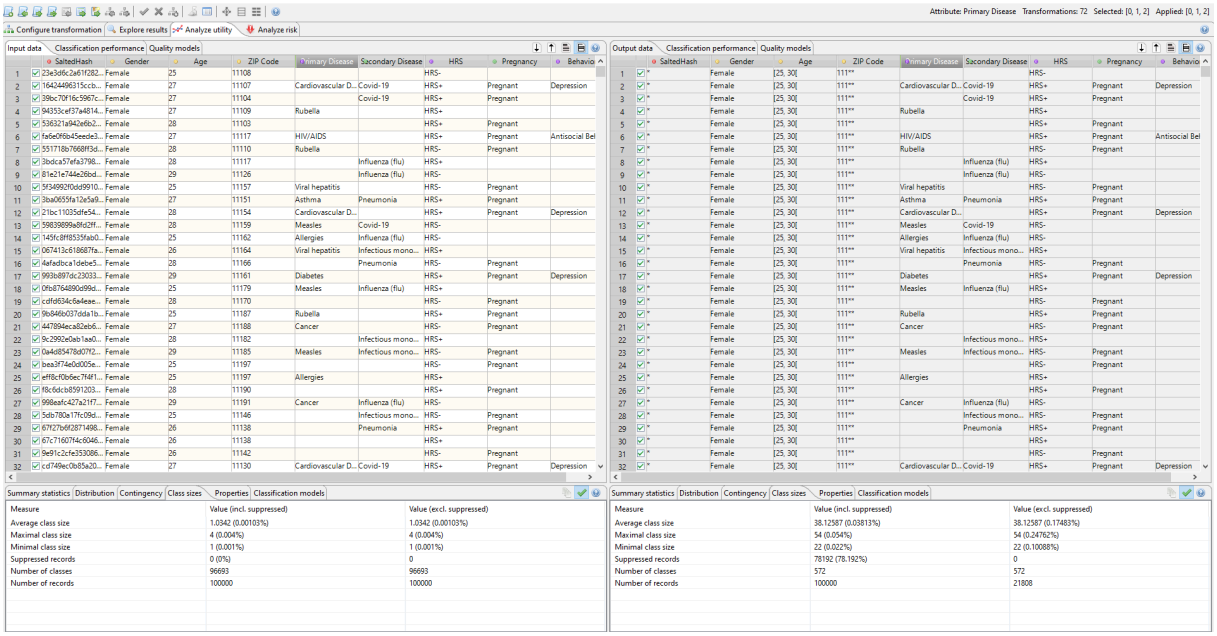
Εικόνα 9.32: ARX – Διανομή Ρίσκου

Παρακάτω βλέπουμε σε διαφορετική μορφή το πώς διαμορφώνεται το ρίσκο για τα τρία μοντέλα επιτιθέμενων παρατηρώντας το μέγιστο ρίσκο τα βρίσκεται στο 8,33%, ωστόσο αυτό επηρεάζει μόνο το 0,014% των εγγραφών.

Εικόνα 9.33: ARX – Ανάλυση Ρίσκου

Συνεχίζουμε με την επόμενη προσέγγιση της ανωνυμοποίησης όπου θα συμπεριλάβουμε στα δεδομένα μας τη πρωταρχική ασθένεια (Primary Disease) ως εμπιστευτικό γνώρισμα. Η ασθένεια αυτή λαμβάνει δέκα διαφορετικές τιμές στο σύνολο των δεδομένων μας, ωστόσο εμφανίζεται σε ένα σχετικά μικρό ποσοστό των ασθενών. Αφού ξεκινήσουμε την ανάλυσή μας θα διαπιστώσουμε ότι όσο το εμπιστευτικό γνώρισμα 'behavior' διατηρεί τιμή l=2, κρατώντας τον μετασχηματισμό [0, 1, 3], ότι τιμές και να λάβει το l για την κύρια ασθένεια θα έχουμε τα ίδια αποτελέσματα με την αρχική μας προσέγγιση λόγω του μεγέθους των κλάσεων ισοδυναμίας, όπως αναφέραμε και παραπάνω. Ως εκ τούτου το δείγμα μας θα εξεταστεί για διαφορετική τιμή l για το γνώρισμα 'behavior' ή για διαφορετικό μετασχηματισμό. Θα ξεκινήσουμε με τιμή l=10 για τη πρωταρχική ασθένεια και τιμή l=2 για τη συμπεριφορά.

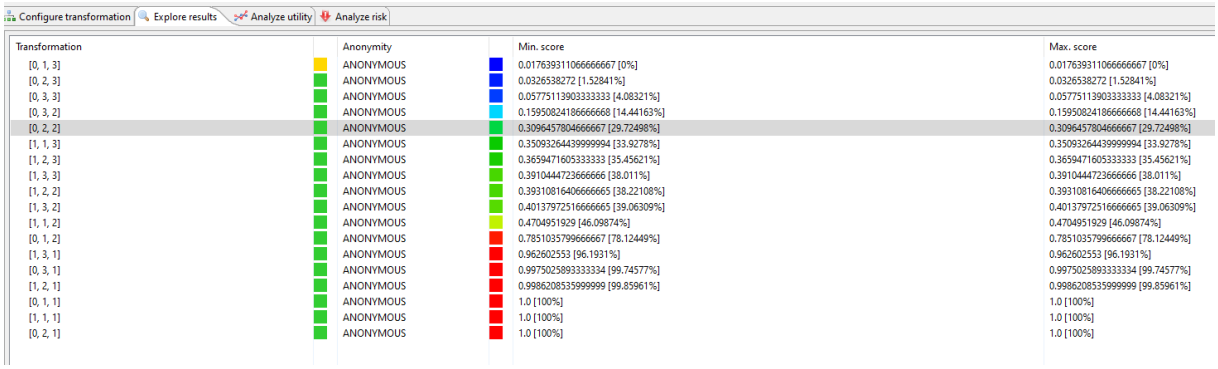
Προκειμένου να επιφέρουμε σχετικά μικρή γενίκευση στα δεδομένα μας θα δοκιμάσουμε τον μετασχηματισμό [0, 1, 2] όπου διαπιστώνουμε ότι έχουμε πολύ μεγάλη κατάργηση των εγγραφών περίπου στο 78% κάτι που δεν διατηρεί χρήσιμα τα δεδομένα μας για την έρευνα.



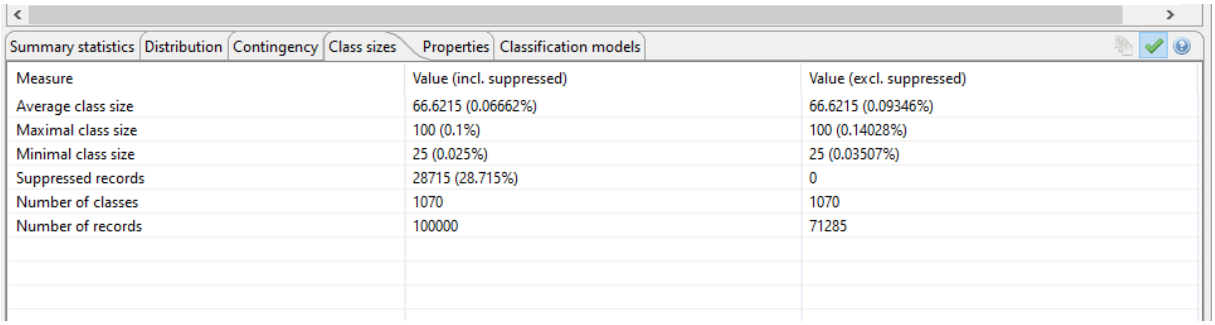
Εικόνα 9.34: ARX – Εξαχθέντα Δεδομένα

Προκειμένου να μη χάσουμε πληροφορία λόγω γενίκευσης ψευδο-αναγνωριστικών όπως το φύλο, δεν θα προβούμε σε μετασχηματισμούς που καταργούν το συγκεκριμένο γνώρισμα, επομένως ένα ακόμα μετασχηματισμός που θα μπορούσαμε ίσως να εξετάσουμε είναι ο [0, 2, 2], όπου γενικεύουμε λίγο περισσότερο την ηλικία σε εύρος δεκαετίας, ωστόσο θα δούμε ότι ο συνδυασμός μεγάλης απώλεια πληροφορίας και κατάργησης εγγραφών περίπου στο 30%

έκαστο, καθώς και του μεγάλου εύρους ηλικιών για τις ανάγκες της έρευνας δεν δημιουργούν μία καλή προσέγγιση.

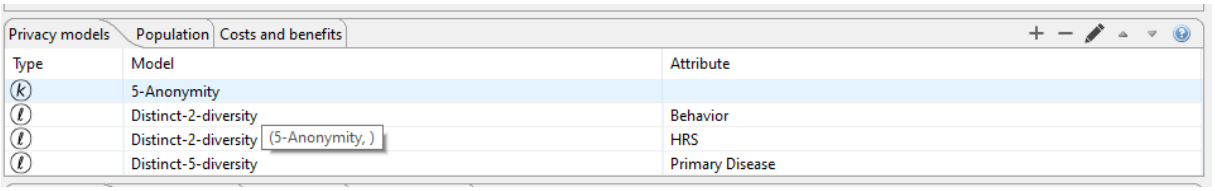


Εικόνα 9.35: ARX – Επιλογή Μετασχηματισμού



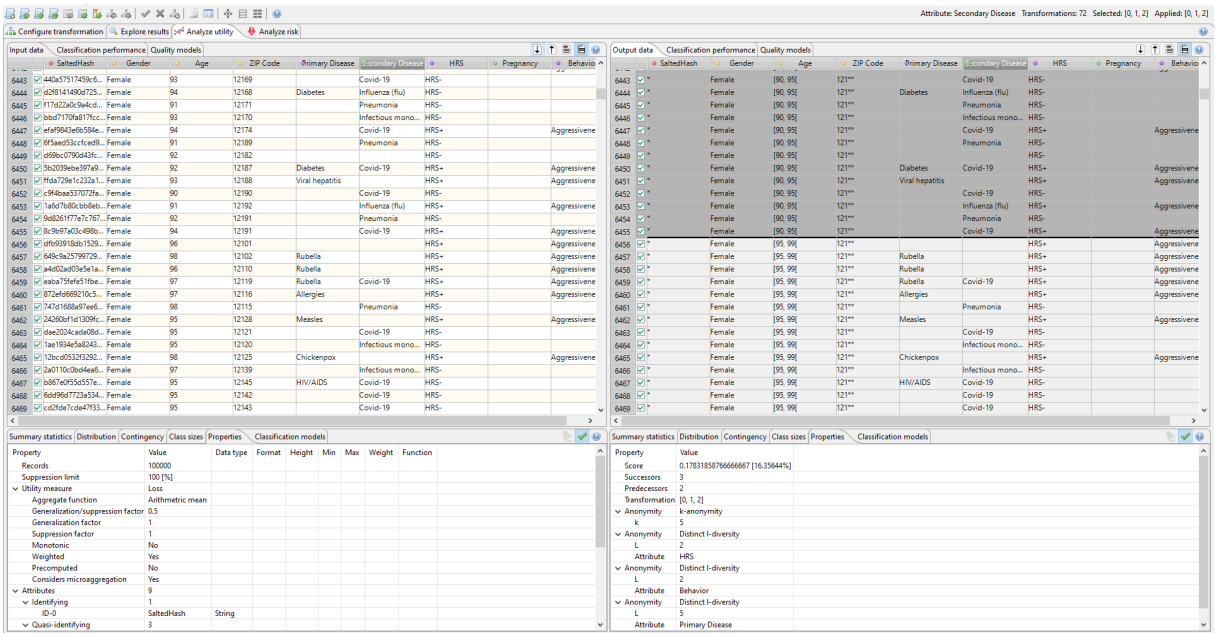
Εικόνα 9.36: ARX – Διαμόρφωση Κλάσεων

Δοκιμάζοντας για όλες τις τιμές l για τη πρωταρχική ασθένεια παρατηρούμε ότι τη προσέγγιση με τιμή l=5 παρουσιάζει τη μικρότερη κατάργηση, σε χαμηλά επίπεδα για τις ανάγκες της έρευνας, ενώ παράλληλα δημιουργεί μία καλή ποικιλομορφία στο γνώρισμα αυτό.



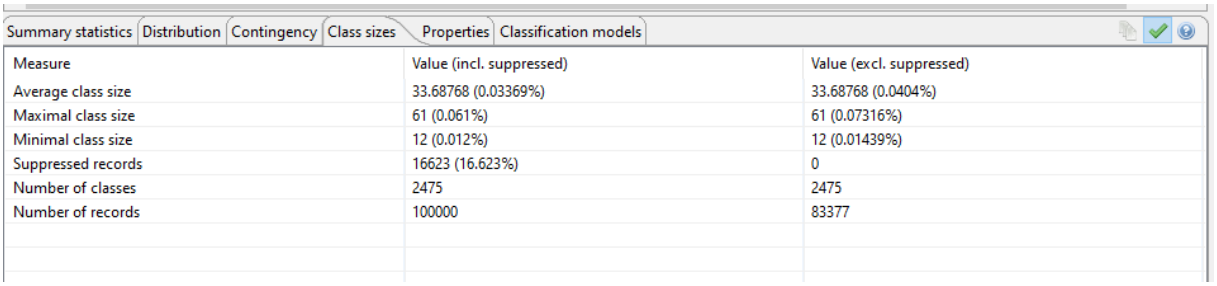
Εικόνα 9.37: ARX – Μοντέλα Απορρήτου

Διατηρώντας έτσι το μετασχηματισμό [0, 1, 2] αρχικά έχουμε μία συνολική απώλεια πληροφορίας στο 16,35% και παράλληλα διατηρούμε τα δείγμα τα μας σε καλά επίπεδα προς μελέτη για τα τρία ψευδο-αναγνωριστικά.



Εικόνα 9.38: ARX – Εξαχθέντα Δεδομένα

Οι εγγραφές μας στη περίπτωση αυτή για τις ανάγκες της ανωνυμοποίησης καταργούνται περίπου στο 16%, διατηρώντας 83377 εγγραφές προς μελέτη.



Measure	Value (incl. suppressed)	Value (excl. suppressed)
Average class size	33.68768 (0.03369%)	33.68768 (0.0404%)
Maximal class size	61 (0.061%)	61 (0.07316%)
Minimal class size	12 (0.012%)	12 (0.01439%)
Suppressed records	16623 (16.623%)	0
Number of classes	2475	2475
Number of records	100000	83377

Εικόνα 9.39: ARX – Διαμόρφωση Κλάσεων

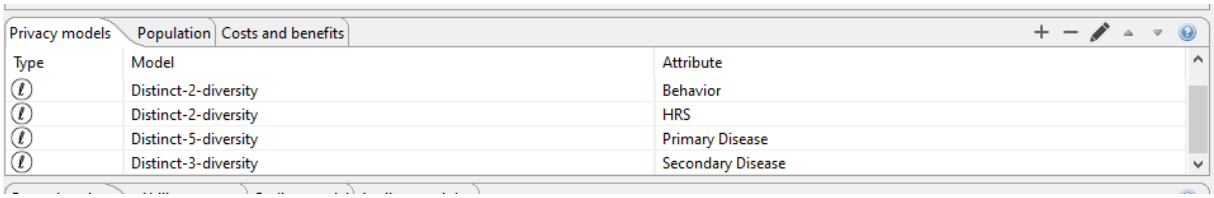
Περνώντας στο ρίσκο αναγνώρισης κάποιας εγγραφής θα δούμε ότι με τη συγκεκριμένη προσέγγιση έχουμε το μεγαλύτερο ρίσκο να είναι 8.33% ωστόσο αυτό επηρεάζει μόλις το 0,014% των εγγραφών μας. Σε ένα δείγμα 100000 εγγραφών, το ποσοστό αυτό συνεπάγεται την πιθανότητα αναγνώρισης 14^{ων} εγγραφών με πιθανότητα μόλις στο 8% περίπου.

Εικόνα 9.40: ARX – Ανάλυση Ρίσκου

Στη συνέχεια προσεγγίζουμε την ανωνυμοποίηση μας με τιμή για το 'behavior' l=3 ξανά, ωστόσο αυτό που θα παρατηρήσουμε είναι ότι ακόμα και για τη μικρότερη τιμή l του γνωρίσματος της πρωταρχικής ασθένειας (l=2), λαμβάνουμε τα ίδια αποτελέσματα με τις προηγούμενες προσεγγίσεις (όπου χρησιμοποιήσαμε ξανά τιμή l=3 για το 'behavior') τόσο για το ρίσκο αναγνώρισης αλλά και στην απώλεια πληροφορίας για όλους τους μετασχηματισμούς, ως εκ τούτου η περαιτέρω διερεύνηση στη προσέγγιση αυτή δεν έχει να μας δώσει κάτι χρήσιμο.

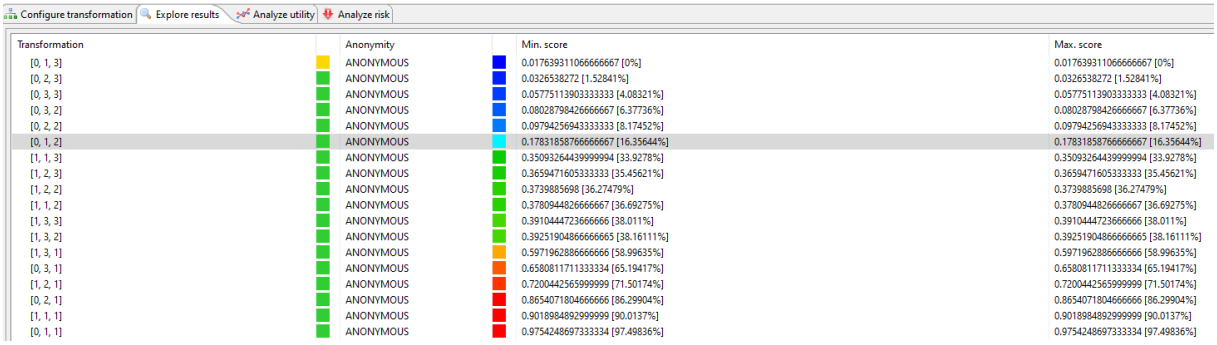
Κλείνοντας την ανωνυμοποίηση των δεδομένων μας, θα συμπεριλάβουμε και το τελευταίο ευαίσθητο γνώρισμα ως εμπιστευτικό και είναι η δευτερεύουσα ασθένεια (Secondary Disease), πού όπως έχουμε αναφέρει είναι και ο πιθανός λόγος επίσκεψης του ασθενούς στο νοσοκομείο. Το γνώρισμα αυτό καθώς εμφανίζεται επίσης σε μικρό ποσοστό στους ασθενείς λαμβάνει πέντε διαφορετικές τιμές μέσα στο σύνολο δεδομένων. Επίσης το γνώρισμα αυτό δείχνει πως δεν επηρεάζει ή επηρεάζεται από τη νόσο HRS.

Καθώς εισάγουμε και αυτό το γνώρισμα ως εμπιστευτικό στην ανωνυμοποίηση μας θα λάβει τιμή l = 3.



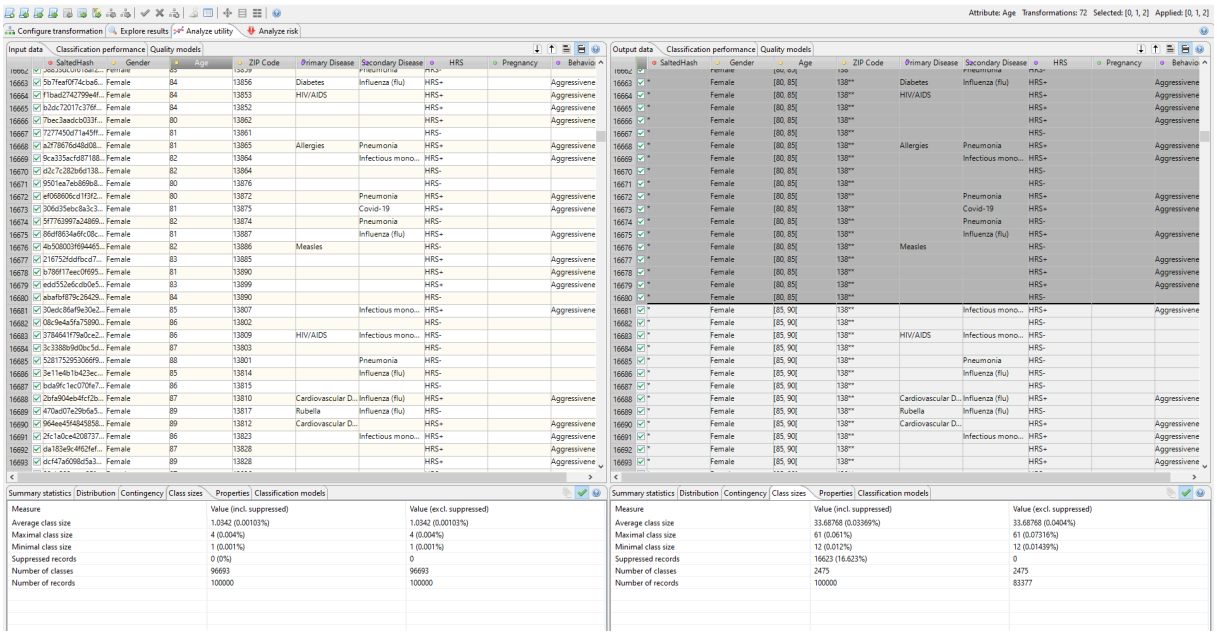
Εικόνα 9.41: ARX – Μοντέλα Απορρήτου

Εξετάζοντας του μετασχηματισμούς μας θα δούμε ότι ο πρώτος [0, 1, 3] μας δίνει τα ίδια αποτελέσματα με προηγούμενες προσεγγίσεις, καθώς ο [0, 1, 2] έχει κάποιες διαφοροποιήσεις.



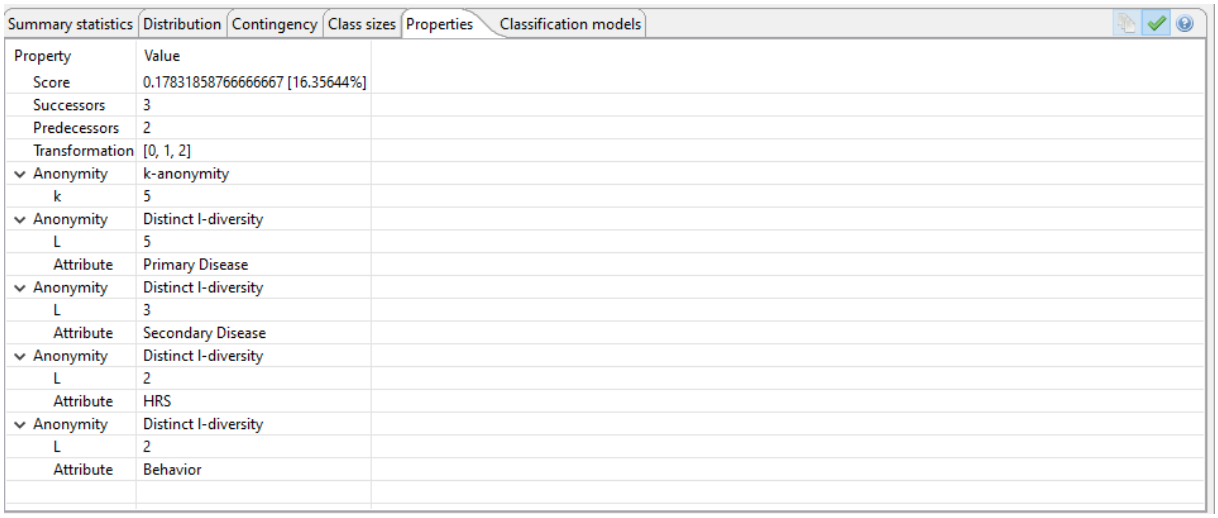
Εικόνα 9.42: ARX – Επιλογή Μετασχηματισμού

Ως εκ τούτου θα εστιάσουμε στο δεύτερο μετασχηματισμό παρατηρώντας πώς διαμορφώνονται τα δεδομένα μας.



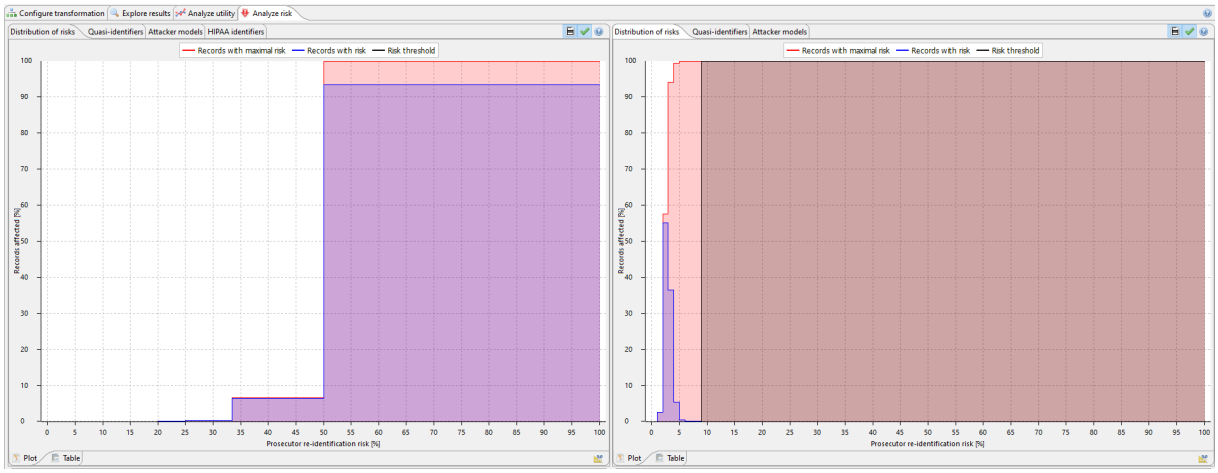
Εικόνα 9.43: ARX – Εξαχθέντα Δεδομένα

Θα δούμε ότι η κατάργηση των εγγραφών είναι περίπου στο 16%, ένα καλό σχετικά ποσοστό για τις ανάγκες της έρευνας, καθώς και η απώλεια της πληροφορίας διακυμαίνεται στο 16% περίπου επίσης.



Εικόνα 9.44: ARX – Ιδιότητες Μετασχηματισμού

Παρακάτω (εικόνα 9.45) βλέπουμε σε μορφή γραφήματος πώς κατανέμεται το ρίσκο για τα αρχικά και τα ανώνυμα δεδομένα μας.



Εικόνα 9.45: ARX – Διανομή Ρίσκου

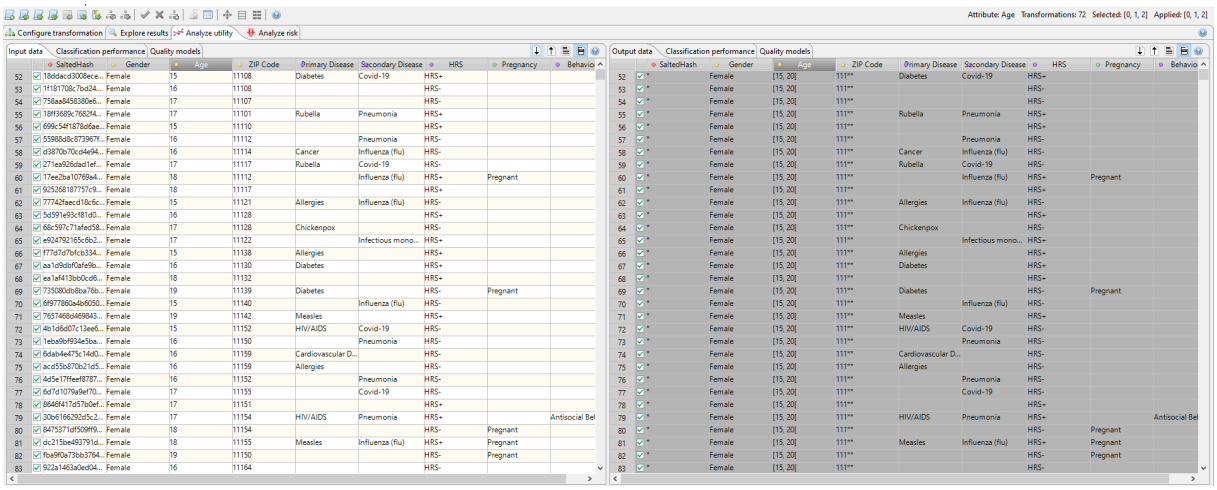
Όπως θα δούμε και στην εικόνα 9.X κάτω, το μέγιστο ρίσκο για αυτή τη προσέγγιση είναι στο 8.33% ενώ οι επηρεαζόμενες εγγραφές από αυτό είναι σε ποσοστό 0,014%.

Εικόνα 9.46: ARX – Ανάλυση Ρίσκου

Έχοντας εξετάσει όλες τις χρήσιμες για τα δεδομένα μας, από πλευράς χρησιμότητας άλλα και ρίσκου, προσεγγίσεις θα πρέπει να αναλύσουμε σε επόμενη ενότητα τα αποτελέσματα που λάβαμε καταλήγοντας στους αποτελεσματικότερους συνδυασμούς για τις ανάγκες της έρευνας. Η χρήση του μοντέλου t-closeness δεν εφαρμόστηκε λόγω της πολυπλοκότητας των δεδομένων και του μικρού ποσοστού εμφάνισης των ασθενειών καθώς και της συμπεριφοράς στις εγγραφές μας, διότι λόγω των παραπάνω δημιουργεί μεγάλη απώλεια στα δεδομένα μας χωρίς να προσφέρει περισσότερη ασφάλεια στη διαδικασία.

Επίσης κάτι σημαντικό που πρέπει να εξετάσουμε, ακόμα και αν δεν αποτελεί εμπιστευτικό γνώρισμα όπως αναφέραμε και σε προηγούμενο κεφάλαιο, είναι η πιθανότητα εγκυμοσύνης η οποία παίζει καθοριστικό ρόλο στην αποκάλυψη κάποιας εγγραφής μέσα στο σύνολο των ανώνυμων δεδομένων. Όπως μπορούμε να δούμε από την εικόνα 9.47, υπάρχει τουλάχιστον μία κλάση ισοδυναμίας όπου εμφανίζεται έγκυος γυναίκα η οποία είναι η μοναδική στη κλάση αυτή που πάσχει από σακχαρώδη διαβήτη. Το πρόβλημα αυτό εμφανίζεται και σε άλλες κλάσεις για διαφορετικές ασθένειες. Έχοντας πρόσβαση κάποιος στα ανώνυμα δεδομένα αλλά και στον αρχικό πίνακα, μπορεί με βεβαιότητα να αποκαλύψει αυτή την εγγραφή, εφόσον βέβαια γνωρίζει το χαρακτηριστικό της εγκυμοσύνης για τη συγκεκριμένη γυναίκα που έχει συμπεριληφθεί στην εν λόγω κλάση ισοδυναμίας. Το γνώρισμα αυτό πρακτικά δεν αποτελεί ούτε ευαίσθητο αλλά ούτε και ψευδο-αναγνωριστικό (παρόλο που, στο σενάριο αναγνώρισης που περιγράψαμε, θεωρήθηκε κατά μία έννοια ως ψευδο-αναγνωριστικό) και δεν μπορεί να αντιμετωπιστεί ως τέτοιο, καθώς θα

επέφερε πολύ μεγάλη απώλεια πληροφορίας στα δεδομένα μας καθιστώντας τα αδύνατα προς μελέτη.



Εικόνα 9.47: ARX – Κίνδυνος Αποκάλυψης Εγγραφής

Το παράδειγμα αυτό καταδεικνύει τη δυσκολία μίας αποτελεσματικής προσπάθειες ανωνυμοποίησης ιδίως σε ρεαλιστικά σενάρια δεδομένων με μεγάλο πλήθος γνωρισμάτων. Ένας τρόπος αντιμετώπισης τέτοιου ζητήματος, θα ήταν η ακολουθία αρκετά μεγαλύτερων κλάσεων ισοδυναμίας μέσω μεγαλύτερης γενίκευσης που θα αποτελούσε και μεγαλύτερη απώλεια στα δεδομένα μας. Ωστόσο από τη στιγμή που τα δεδομένα αυτά δεν δημοσιεύονται, αλλά παραχωρούνται σε ερευνητικό κέντρο προς μελέτη είναι ένα ρίσκο που υπό προϋποθέσεις μπορεί να γίνει αποδεκτό για τις ανάγκες της έρευνας αν συνυπολογίσουμε και το γεγονός ότι τα αναγνωριστικά του αρχικού πίνακα είναι ψευδωνυμοποιημένα. Σε άλλη περίπτωση θα έπρεπε να σχηματιστούν σαφώς μεγαλύτερες κλάσεις ισοδυναμίας, όπως αναφέραμε, που θα εξάλειφαν το πρόβλημα αυτό.

9.3 Ανάλυση Αποτελεσμάτων

Στο σημείο αυτό έχουμε ολοκληρώσει τις προσεγγίσεις της ανωνυμοποίησης. Για κάθε μία από αυτές καταγράψαμε τα αποτελέσματα τόσο για την ανάλυση της χρησιμότητας των ανώνυμων δεδομένων έτσι όπως προέκυψαν από τη διαδικασία, όσο και του ρίσκου αποκάλυψης, μέσα από τα τρία μοντέλα επιτιθέμενων. Όπως είδαμε μέσα από τις προσεγγίσεις αυτές η ανωνυμοποίηση είναι μία απαιτητική διαδικασία που χρειάζεται από το σχεδιασμό ένα καλά ορισμένο στόχο αξιοποίησης των ανώνυμων δεδομένων για τις ανάγκες της έρευνας. Ακόμη διακρίναμε ότι

κάποιες προσεγγίσεις εστιάζουν περισσότερο στη κάλυψη του ενδεχόμενου ρίσκου, ενώ κάποιες άλλες ως προς την απώλεια λιγότερης κατά το δυνατό πληροφορίας.

9.3.1 Ανάλυση Χρησιμότητας

Αναφορικά με τη χρησιμότητα των δεδομένων είδαμε ότι σημαντικό ρόλο παίζει η γενίκευση των γνωρισμάτων μέσω των μετασχηματισμών που δημιουργούμε. Επίσης παρατηρήσαμε ότι αρκετές προσεγγίσεις επιφέρουν αρκετά μεγάλη, μη αποδεκτή σε κάποιες περιπτώσεις, κατάργηση εγγραφών.

Για τις ανάγκες της ερευνάς μας κινηθήκαμε σε ένα συγκεκριμένο πλαίσιο, μέσω του οποίου τα δεδομένα μας θα παραμείνουν χρήσιμα κατά το μέγιστο δυνατό προς μελέτη και για τα τρία ψευδο-αναγνωριστικά, κάτι που ήταν εξ' αρχής ο στόχος μας. Πιο συγκεκριμένα από τα εξαχθέντα ανώνυμα δεδομένα επιθυμούσαμε να λάβουμε συμπεράσματα για την επιρροή της επιδημίας HRS ως προς το φύλο του ασθενούς, τις ηλικιακές ομάδες, διατηρώντας το εύρος τους σε χαμηλό επίπεδο αλλά και τη γεωγραφική επέκταση της νόσου. Έτσι εστιάσαμε σε συγκεκριμένους μετασχηματισμούς όπου διατηρούν τις πληροφορίες αυτές όσο πιο κοντά στα αρχικά δεδομένα.

Ένα από τους κύριους λόγους της πολυπλοκότητας του έργου μας ήταν ότι περιείχε αρκετά εμπιστευτικά γνωρίσματα, κάτι που έπαιξε σημαντικό ρόλο στη πορεία της διαδικασίας.

Κατά την ανωνυμοποίηση ακολουθήσαμε δύο μετασχηματισμούς που θεωρήσαμε χρήσιμους για την έρευνα. Ο πρώτος [0, 1, 3] επιφέρει μεγαλύτερη γενίκευση στα ψευδο-αναγνωριστικά και συγκεκριμένα σε αυτό του ΤΚ, ωστόσο έχει και ως αποτέλεσμα σαφώς μικρότερα ποσοστά κατάργησης στις εγγραφές. Ο δεύτερος μετασχηματισμός [0, 1, 2] αποτελεί μία πολύ καλή προσέγγιση έρευνας και για τα τρία ψευδο-αναγνωριστικά μας, πετυχαίνοντας μικρότερη γενίκευση σε αυτά, ωστόσο λόγω του μεγαλύτερου ποσοστού καταργήσεων στις εγγραφές του συνόλου, έχουμε και μεγαλύτερη απώλεια στη συνολική πληροφορία μας. Εξετάζοντας όλες τις προσεγγίσεις έχουμε τα παρακάτω αποτελέσματα.

Κατά τη πρώτη προσέγγιση συμπεριλάβαμε μόνο τη νόσο HRS ως εμπιστευτικό γνώρισμα.

Προσέγγιση 1	**k = 5**	**HRS**	**l = 2**
Μετασχηματισμός	**Απώλεια**	**Κατάργηση**	
[0, 1, 3]	1,42%	0,012%	

	Κλάσεις Ισοδυναμίας		
	Min	**Max**	**Mean**
	211	421	328
[0, 1, 2]	**Απώλεια**	**Κατάργηση**	
	1,12%	0,012%	
	Κλάσεις Ισοδυναμίας		
	Min	**Max**	**Mean**
	12	61	33

Πίνακας 9.1: Προσέγγιση 1 - Χρησιμότητα Δεδομένων

Όπως βλέπουμε από τον πίνακα 9.1 παραπάνω, έχοντας ένα μόνο εμπιστευτικό γνώρισμα, και λόγω της εντροπίας που παρουσιάζει καθώς λαμβάνει δύο τιμές (θετικό ή αρνητικό) με ποσοστό περίπου κοντά στο 50% και οι δύο μετασχηματισμοί έχουν την ίδια κατάργηση εγγραφών και αρκετά κοντινή απώλεια χρήσιμης πληροφορίας. Ο μετασχηματισμός [0, 1, 2] μάλιστα μας δίνει και αρκετά πιο σαφή συμπεράσματα για τη γεωγραφική επέκταση της νόσου.

Στη δεύτερη προσέγγιση της ανωνυμοποίησης συμπεριλάβαμε ένα ακόμα εμπιστευτικό πεδίο, αυτό της συμπεριφοράς του ασθενούς το οποίο λαμβάνει τέσσερις διαφορετικές τιμές, επηρεαζόμενη θεωρητικά από το αποτέλεσμα της νόσου HRS. Για αρχική τιμή l = 4 έχουμε πλήρη απώλεια πληροφορίας και κατάργηση των εγγραφών. Οπότε επιλέξαμε τιμή l = 3.

Προσέγγιση 2	**k = 5**	**HRS**	**l = 2**
		Behavior	**l = 3**
Μετασχηματισμός	**Απώλεια**	**Κατάργηση**	
[0, 1, 3]	15,87%	31,02%	
	Κλάσεις Ισοδυναμίας		
	Min	**Max**	**Mean**
	241	414	331
[0, 1, 2]	**Απώλεια**	**Κατάργηση**	
	74,69%	79,30%	

Πίνακας 9.2: Προσέγγιση 2 - Χρησιμότητα Δεδομένων

Όπως μπορούμε να διακρίνουμε στη προσέγγιση αυτή καθώς και όλες τις επόμενες για τιμή της συμπεριφοράς l = 3 δεν μπορεί να χρησιμοποιηθεί ο μετασχηματισμός [0, 1, 2] καθώς χάνουμε σχεδόν όλα τα δεδομένα της έρευνας. Ωστόσο αξίζει να σημειωθεί για τον μετασχηματισμό [0, 1, 3], παρά το ότι το ποσοστό της κατάργησης είναι επίσης αρκετό, παραμένουν αρκετές εγγραφές στα ανώνυμα δεδομένα που θα μπορούσαν να μας δώσουν πολύτιμα συμπεράσματα.

Η τρίτη προσέγγιση έλαβε τιμή l=2 για τη συμπεριφορά και όπως θα δούμε μας δίνει τα ίδια αποτελέσματα με αυτά της πρώτης για το μετασχηματισμό [0, 1, 3] τόσο στην απώλεια πληροφορίας, όσο και στη συνολική κατάργηση εγγραφών. Ο λόγος οφείλεται στο γεγονός ότι όλες οι κλάσεις ισοδυναμίας για το μετασχηματισμό αυτό περιέχουν κατ' ελάχιστο 211 εγγραφές με αποτέλεσμα να εμφανίζονται τουλάχιστον δύο διαφορετικές τιμές της συμπεριφοράς σε κάθε κλάση. Για το μετασχηματισμό [0, 1, 2] ωστόσο παρατηρούμε ότι έχουμε μεγαλύτερη απώλεια στα δεδομένα μας καθώς και στο σύνολο των εγγραφών.

Προσέγγιση 3	**k = 5**	**HRS**	**l = 2**
		Behavior	**l = 2**
Μετασχηματισμός	**Απώλεια**	**Κατάργηση**	
[0, 1, 3]	1,42%	0,012%	
	Κλάσεις Ισοδυναμίας		
	Min	**Max**	**Mean**
	211	421	328
[0, 1, 2]	**Απώλεια**	**Κατάργηση**	
	15,92%	16,19%	
	Κλάσεις Ισοδυναμίας		
	Min	**Max**	**Mean**
	12	61	33

Πίνακας 9.3: Προσέγγιση 3 - Χρησιμότητα Δεδομένων

Στην τέταρτη προσέγγιση συμπεριλάβαμε και το γνώρισμα 'Primary Disease' ως εμπιστευτικό, το οποίο λαμβάνει δέκα διαφορετικές τιμές. Για το μετασχηματισμό [0, 1, 3] όπως θα δούμε και στον πίνακα 9.4 παρακάτω δεν υπάρχει νόημα να εξετάζουμε περεταίρω για τιμή του 'behavior' l = 2 καθώς λαμβάνει τα ίδια αποτελέσματα.

Προσέγγιση 4	**k = 5**	**HRS**	**l = 2**

		Behavior	**l = 2**
		Primary Disease	**l = 10**
Μετασχηματισμός	**Απώλεια**	**Κατάργηση**	
[0, 1, 3]	1,42%	0,012%	
	Κλάσεις Ισοδυναμίας		
	Min	**Max**	**Mean**
	211	421	328
[0, 1, 2]	**Απώλεια**	**Κατάργηση**	
	78,12%	78,19%	

Πίνακας 9.4: Προσέγγιση 4 - Χρησιμότητα Δεδομένων

Όπως βλέπουμε στη προσέγγιση αυτή ο μετασχηματισμός [0, 1, 2] επίσης δεν μπορεί να λειτουργήσει καθώς έχουμε υπερβολικά μεγάλη απώλεια στα δεδομένα μας.

Στη συνέχεια ακολουθήσαμε μία άλλη προσέγγιση αφού δοκιμάσαμε για διάφορες τιμές του l για το γνώρισμα 'Primary Disease' καταλήγοντας στη τιμή l = 5.

Προσέγγιση 5	**k = 5**	**HRS**	**l = 2**
		Behavior	**l = 2**
		Primary Disease	**l = 5**
Μετασχηματισμός	**Απώλεια**	**Κατάργηση**	
[0, 1, 3]	1,42%	0,012%	
	Κλάσεις Ισοδυναμίας		
	Min	**Max**	**Mean**
	211	421	328
[0, 1, 2]	**Απώλεια**	**Κατάργηση**	
	16,35%	16,62%	
	Κλάσεις Ισοδυναμίας		
	Min	**Max**	**Mean**
	12	61	33

Πίνακας 9.5: Προσέγγιση 5 - Χρησιμότητα Δεδομένων

Στη συνέχεια δοκιμάσαμε για τιμή του γνωρίσματος 'behavior' l=3 ξανά, ωστόσο η προσέγγιση αυτή δεν μπορεί να λειτουργήσει για μετασχηματισμό [0, 1, 2].

Προσέγγιση 6	**k = 5**	**HRS**	**l = 2**
		Behavior	**l = 3**
		Primary Disease	**l = 5**
Μετασχηματισμός	**Απώλεια**	**Κατάργηση**	
[0, 1, 3]	15,87%	31,002%	
	Κλάσεις Ισοδυναμίας		
	Min	**Max**	**Mean**
	241	414	331
[0, 1, 2]	**Απώλεια**	**Κατάργηση**	
	74,71%	79,32%	

Πίνακας 9.6: Προσέγγιση 6 - Χρησιμότητα Δεδομένων

Τέλος εφαρμόσαμε τη τελευταία προσέγγιση όπου έχουν συμπεριληφθεί όλα τα εμπιστευτικά γνωρίσματα.

Προσέγγιση 7	**k = 5**	**HRS**	**l = 2**
		Behavior	**l = 2**
		Primary Disease	**l = 5**
		Secondary Disease	**l = 3**
Μετασχηματισμός	**Απώλεια**	**Κατάργηση**	
[0, 1, 3]	1,42%	0,012%	
	Κλάσεις Ισοδυναμίας		
	Min	**Max**	**Mean**
	211	421	328
[0, 1, 2]	**Απώλεια**	**Κατάργηση**	
	16,35%	16,62%	
	Κλάσεις Ισοδυναμίας		
	Min	**Max**	**Mean**
	12	61	33

Πίνακας 9.7: Προσέγγιση 7 - Χρησιμότητα Δεδομένων

9.3.2 Ανάλυση Ρίσκου

Περνώντας στην ανάλυση του ρίσκου, παρατηρήσαμε ότι βάσει των γραφημάτων και των αποτελεσμάτων, το ρίσκο αποκάλυψης για όλες τις προσεγγίσεις κυμαινόταν σε αρκετά χαμηλά επίπεδα.

Προσέγγιση 1	**k = 5**	**HRS**	**l = 2**
Μετασχηματισμός	**Ρίσκο**	**Μέγιστο Ρίσκο**	**Επηρεαζόμενες Εγγραφές**
[0, 1, 3]	0%	0,47%	0,21%
[0, 1, 2]	0%	8,33%	0,012%

Πίνακας 9.8: Προσέγγιση 1 – Ανάλυση Ρίσκου

Από πλευράς ρίσκου αποκάλυψης κάποιας εγγραφής θα δούμε ότι κατά τη πρώτη προσέγγιση το μέγιστο ρίσκο είναι αρκετά χαμηλότερο με τον μετασχηματισμό [0, 1, 3], ωστόσο με τον μετασχηματισμό [0, 1, 2] αν και παρουσιάζεται μεγαλύτερο μέγιστο ρίσκο, από αυτό επηρεάζονται αρκετά λιγότερες εγγραφές.

Στη δεύτερη προσέγγιση εξετάσαμε το ρίσκο μόνο για το μετασχηματισμό [0, 1, 3] καθώς είναι ο μόνος που δεν καταργεί σχεδόν πλήρως τα δεδομένα.

Προσέγγιση 2	**k = 5**	**HRS**	**l = 2**
		Behavior	**l = 3**
Μετασχηματισμός	**Ρίσκο**	**Μέγιστο Ρίσκο**	**Επηρεαζόμενες Εγγραφές**
[0, 1, 3]	0%	7,14%	0,067%

Πίνακας 9.9: Προσέγγιση 2 – Ανάλυση Ρίσκου

Όπως μπορούμε να δούμε το ρίσκο διαμορφώνεται κατά μέγιστο στο 7% περίπου επηρεάζοντας μόλις το 0,067% του συνολικού δείγματος.

Από πλευράς ρίσκου η τρίτη προσέγγιση κυμαίνεται στα ίδια επίπεδα με αυτά της πρώτης με τη μόνη διαφορά ότι έχει ελαφρώς αυξημένο το ποσοστό των εγγραφών που επηρεάζονται από το μέγιστο ρίσκο, στο 0,014%.

Η τέταρτη προσέγγιση έχει το ίδιο ρίσκο με τη δεύτερη για το μετασχηματισμό [0, 1, 3] ωστόσο για το μετασχηματισμό [0, 1, 2] όπως είδαμε έχουμε πολύ μεγάλη απώλεια στα δεδομένα μας και δεν υπάρχει λόγος να εξεταστεί περεταίρω.

Για τη πέμπτη προσέγγιση βλέπουμε το ρίσκο στον πίνακα 9.10. Όπως βλέπουμε το ρίσκο δεν αλλάζει στο πρώτο μετασχηματισμό, ενώ στον δεύτερο διαμορφώνεται διαφορετικά επηρεάζοντας το 0,014% των εγγραφών.

Προσέγγιση 5	**k = 5**	**HRS**	**l = 2**
		Behavior	**l = 2**
		Primary Disease	**l = 5**
Μετασχηματισμός	**Ρίσκο**	**Μέγιστο Ρίσκο**	**Επηρεαζόμενες Εγγραφές**
[0, 1, 3]	0%	0,47%	0,21%
[0, 1, 2]	0%	8,33%	0,014%

Πίνακας 9.10: Προσέγγιση 5 – Ανάλυση Ρίσκου

Το ρίσκο για την έκτη προσέγγιση διαμορφώνεται ως εξής. Καθώς ο μετασχηματισμός [0, 1, 2] επιφέρει μεγάλη κατάργηση δεν εξετάζεται ως προς το ρίσκο. Ωστόσο για το μετασχηματισμό [0, 1, 3] έχουμε το χαμηλότερο μέγιστο ρίσκο.

Προσέγγιση 6	**k = 5**	**HRS**	**l = 2**
		Behavior	**l = 3**
		Primary Disease	**l = 5**
Μετασχηματισμός	**Ρίσκο**	**Μέγιστο Ρίσκο**	**Επηρεαζόμενες Εγγραφές**
[0, 1, 3]	0%	0,41%	0,34%

Πίνακας 9.11: Προσέγγιση 6 – Ανάλυση Ρίσκου

Τέλος, παρακάτω βλέπουμε τη διαμόρφωση του ρίσκου για τη τελευταία προσέγγιση που συμπεριλαμβάνει όλα τα εμπιστευτικά γνωρίσματα.

Προσέγγιση 7	**k = 5**	**HRS**	**l = 2**
		Behavior	**l = 2**
		Primary Disease	**l = 5**

		Secondary Disease	l = 3
Μετασχηματισμός	**Ρίσκο**	**Μέγιστο Ρίσκο**	**Επηρεαζόμενες Εγγραφές**
[0, 1, 3]	0%	0,47%	0,21%
[0, 1, 2]	0%	8,33%	0,014%

Πίνακας 9.12: Προσέγγιση 7 – Ανάλυση Ρίσκου

9.3.3 Επιλογή Προσέγγισης

Έχοντας λάβει τα αποτελέσματά μας αναφορικά με τη χρησιμότητα των ανώνυμων δεδομένων και του ρίσκου αποκάλυψής τους, μπορούμε πλέον να καταλήξουμε στην επιλογή της προσέγγισης που θα ακολουθήσουμε. Καθώς τα αναγνωριστικά των αρχικών δεδομένων είναι ψευδωνυμοποιημένα, προσφέροντας ένα πρώτο ισχυρό επίπεδο ασφάλειας, καθώς και ο ανωνυμοποιημένος πίνακας προορίζεται για μελέτη από ερευνητικό κέντρο αποκλείοντας τη πιθανότητα δημοσίευσής του, μας δίνεται η ευχέρεια να αποδεχτούμε κάποια ρίσκα που επιφέρει η διαδικασία. Ως εκ τούτου μας επιτρέπει να εστιάσουμε σε μικρότερο βάθος γενίκευσης, λαμβάνοντας πολύ χρήσιμα αποτελέσματα για την έρευνα. Όπως είδαμε όλα τα εμπιστευτικά γνωρίσματα είναι σημαντικά για τις ανάγκες μας, έτσι θα εστιάσουμε στη τελευταία προσέγγιση που περιέχει όλα τα ευαίσθητα πεδία ως εμπιστευτικά. Όσον αφορά το μετασχηματισμό, θα επιλέξουμε τον [0, 1, 2] καθώς η απώλεια πληροφορίας στο 16,35% είναι ένα αποδεκτό ποσοστό για την έρευνα, ενώ διατηρεί 83377 εγγραφές προς μελέτη που μπορούν να μας δώσουν χρήσιμα συμπεράσματα. Από την οπτική του ρίσκου, το μέγιστο ρίσκο για όλα τα μοντέλα επιτιθέμενων ορίζεται στο 8.33%, ωστόσο επηρεάζει με τη πιθανότητα αυτή, μόλις το 0,014% στο σύνολο των δεδομένων το οποίο είναι επίσης αποδεκτό για τα ανάγκες μας. Τέλος καθώς παρατηρήσαμε ένα ακόμα γνώρισμα (αυτό της εγκυμοσύνης) να αποτελεί κίνδυνο αναγνώρισης κάποιας εγγραφής, η φύση των δεδομένων, όπως και ο τρόπος αξιοποίησης τους μας επιτρέπουν να αποδεχτούμε το ρίσκο αυτό και να προχωρήσουμε έτσι την έρευνα. Παρακάτω (πίνακας 9.13) παρουσιάζονται όλες οι χρήσιμες για τα δεδομένα μας προσεγγίσεις, καθώς και η προσέγγιση και ο μετασχηματισμός που επιλέξαμε (εικόνα 9.43).

Προσέγγιση 1	**k = 5**	**HRS**	**l = 2**
Μετασχηματισμός	**Απώλεια**	**Κατάργηση**	
[0, 1, 3]	1,42%	0,012%	
	Ρίσκο	**Μέγιστο Ρίσκο**	**Επηρεαζόμενες Εγγραφές**
	0%	0,47%	0,21%

	Κλάσεις Ισοδυναμίας		
	Min	**Max**	**Mean**
	211	421	328
[0, 1, 2]	**Απώλεια**	**Κατάργηση**	
	1,12%	0,012%	
	Ρίσκο	**Μέγιστο Ρίσκο**	**Επηρεαζόμενες Εγγραφές**
	0%	8,33%	0,012%
	Κλάσεις Ισοδυναμίας		
	Min	**Max**	**Mean**
	12	61	33
Προσέγγιση 2	**k = 5**	**HRS**	**l = 2**
		Behavior	**l = 3**
Μετασχηματισμός	**Απώλεια**	**Κατάργηση**	
[0, 1, 3]	15,87%	31,02%	
	Ρίσκο	**Μέγιστο Ρίσκο**	**Επηρεαζόμενες Εγγραφές**
	0%	7,14%	0,067%
	Κλάσεις Ισοδυναμίας		
	Min	**Max**	**Mean**
	241	414	331
Προσέγγιση 3	**k = 5**	**HRS**	**l = 2**
		Behavior	**l = 2**
Μετασχηματισμός	**Απώλεια**	**Κατάργηση**	
[0, 1, 3]	1,42%	0,012%	
	Ρίσκο	**Μέγιστο Ρίσκο**	**Επηρεαζόμενες Εγγραφές**
	0%	0,47%	0,21%
	Κλάσεις Ισοδυναμίας		
	Min	**Max**	**Mean**
	211	421	328
[0, 1, 2]	**Απώλεια**	**Κατάργηση**	
	15,92%	16,19%	
	Ρίσκο	**Μέγιστο Ρίσκο**	**Επηρεαζόμενες Εγγραφές**
	0%	8,33%	0,014%
	Κλάσεις Ισοδυναμίας		

	Min	**Max**	**Mean**
	12	61	33
Προσέγγιση 5	**k = 5**	**HRS**	**l = 2**
		Behavior	**l = 2**
		Primary Disease	**l = 5**
Μετασχηματισμός	**Απώλεια**	**Κατάργηση**	
[0, 1, 3]	1,42%	0,012%	
	Ρίσκο	**Μέγιστο Ρίσκο**	**Επηρεαζόμενες Εγγραφές**
	0%	0,47%	0,21%
	Κλάσεις Ισοδυναμίας		
	Min	**Max**	**Mean**
	211	421	328
[0, 1, 2]	**Απώλεια**	**Κατάργηση**	
	16,35%	16,62%	
	Ρίσκο	**Μέγιστο Ρίσκο**	**Επηρεαζόμενες Εγγραφές**
	0%	8,33%	0,014%
	Κλάσεις Ισοδυναμίας		
	Min	**Max**	**Mean**
	12	61	33
Προσέγγιση 6	**k = 5**	**HRS**	**l = 2**
		Behavior	**l = 3**
		Primary Disease	**l = 5**
Μετασχηματισμός	**Απώλεια**	**Κατάργηση**	
[0, 1, 3]	15,87%	31,002%	
	Ρίσκο	**Μέγιστο Ρίσκο**	**Επηρεαζόμενες Εγγραφές**
	0%	0,41%	0,34%
	Κλάσεις Ισοδυναμίας		
	Min	**Max**	**Mean**
	241	414	331
Προσέγγιση 7	**k = 5**	**HRS**	**l = 2**
		Behavior	**l = 2**
		Primary Disease	**l = 5**

		Secondary Disease	**l = 3**
Μετασχηματισμός	**Απώλεια**	**Κατάργηση**	
[0, 1, 3]	1,42%	0,012%	
	Ρίσκο	**Μέγιστο Ρίσκο**	**Επηρεαζόμενες Εγγραφές**
	0%	0,47%	0,21%
	Κλάσεις Ισοδυναμίας		
	Min	**Max**	**Mean**
	211	421	328
[0, 1, 2]	**Απώλεια**	**Κατάργηση**	
	16,35%	16,62%	
	Ρίσκο	**Μέγιστο Ρίσκο**	**Επηρεαζόμενες Εγγραφές**
	0%	8,33%	0,014%
	Κλάσεις Ισοδυναμίας		
	Min	**Max**	**Mean**
	12	61	33

Πίνακας 9.13: Συνολικός πίνακας προσεγγίσεων

9.4 Σύνοψη

Ολοκληρώνοντας την διαδικασία της ανωνυμοποίησης, αξιοποιώντας κάποιες από τις βασικές τεχνικές της, προσεγγίσαμε τα δεδομένα από διαφορετικές πτυχές, κάτι που ανέδειξε τις προκλήσεις της διαδικασίας αυτής. Παρατηρήσαμε ότι αυξάνοντας το πλήθος των εμπιστευτικών γνωρισμάτων, επιδιώκοντας παράλληλα τη διατήρηση των ανώνυμων δεδομένων πολύ κοντά στα αρχικά, παρουσιάζονται ιδιαίτερες δυσκολίες. Ο λόγος αξιοποίησης των ανώνυμων δεδομένων θα πρέπει να είναι εξ' ορισμού σαφής, ώστε να μπορέσουμε να εκμεταλλευτούμε όλες τις δυνατότητές τους, καθώς επίσης και να είμαστε σε θέση να «θυσιάσουμε» σε κάποιες περιπτώσεις ως προς την αντιστάθμιση ρίσκου αποκάλυψης/απώλειας πληροφορίας. Επίσης κατά τη πορεία της διαδικασίας ήρθαμε αντιμέτωποι με αρκετά διλλήματα που αφορούσαν το σύνολο των γνωρισμάτων που θα συμπεριληφθούν στη διαδικασία της ανωνυμοποίησης καθώς και γνωρίσματα που πρακτικά δεν μπορούν να οριστούν ως ψευδο-αναγνωριστικά ή εμπιστευτικά, ωστόσο υπό προϋποθέσεις μπορούν έμμεσα να ταυτοποιήσουν κάποια εγγραφή. Η ανωνυμοποίηση χρησιμοποιήθηκε, ως επέκταση της αρχικής ψευδωνυμοποίησης και αποτελεί

μία επιπλέον δικλείδα ασφαλείας στα δεδομένα υγείας και έτσι περιοριστήκαμε σε βασικές τεχνικές υλοποίησής της.

Κεφάλαιο 10
Επίλογος

Μέσα από την έρευνα που πραγματοποιήσαμε, εξετάσαμε αρκετές πτυχές των προσωπικών δεδομένων και πιο συγκεκριμένα των ευαίσθητων προσωπικών «ηλεκτρονικών» δεδομένων υγείας. Μελετήσαμε τρόπους με τους οποίους μπορούμε να συλλέξουμε και να αξιοποιήσουμε τα δεδομένα αυτά καθώς και κανονισμούς που τα προστατεύουν, εκμεταλλεύοντας τεχνικές που μπορούν να τα τροποποιήσουν κατάλληλα σε μία προσπάθεια να αξιοποιηθούν με ασφαλή τρόπο για ανάγκες έρευνας και να προσφέρουν πολύτιμα συμπεράσματα.

10.1 Ανασκόπηση

Ξεκινώντας το έργο της Μεταπτυχιακής Διατριβής μας θέσαμε κάποια ερευνητικά ερωτήματα πάνω στα οποία θα προσπαθούσαμε να αναπτύξουμε την έρευνα που θα πραγματοποιήσουμε και σιγά σιγά να απαντήσουμε στα ερωτήματα αυτά.

1. Θα μπορούσαν να λειτουργήσουν αποδοτικά οι τεχνικές ανωνυμοποίησης και ψευδωνυμοποίησης των δεδομένων υγείας;

2. Τι ειδικότερες απαιτήσεις προκύπτουν για τη ψευδωνυμοποίηση των δεδομένων; Μήπως κάθε τεχνική ψευδωνυμοποίησης πρέπει να συνδυάζεται και με τεχνικές ανωνυμοποίησης;

3. Η εφαρμογή τέτοιων τεχνικών διατηρεί τα δεδομένα σε μορφή τέτοια ώστε να παράγουν χρήσιμη, για τους εκάστοτε επιδιωκόμενους σκοπούς, πληροφορία;

4. Θα καλύπτουν τη συμμόρφωση με βάση τον GDPR;

Για να μπορέσουμε να δώσουμε σαφή απάντηση στα ερωτήματα αυτά και να εξάγουμε χρήσιμα συμπεράσματα, αρχικά χρειάστηκε να μελετήσουμε σε μεγάλο βαθμό τα ίδια τα προσωπικά δεδομένα ως ευρύτερη έννοια αλλά και πιο συγκεκριμένες περιπτώσεις δεδομένων όπως αυτά της υγείας που είναι και ο κύριος στόχος της έρευνας.

Από τα πρώτα βήματα που κάναμε ήταν να μελετήσουμε τα δεδομένα αυτά από διάφορες πτυχές τους, όπως το πόσο ωφέλιμα θα μπορούσαν να είναι για τις ανάγκες μιας αποτελεσματικότερης παροχής υγειονομικής περίθαλψης καθώς και για ανάγκες επιστημονικής έρευνας, αλλά και τους περιορισμούς αξιοποίησής τους υπό την αιγίδα των κανονισμών που τα προστατεύουν. Από νομικής αλλά και ηθικής πλευρά, εξετάσαμε έννοιες θεμελιωδών δικαιωμάτων όπως αυτό της ιδιωτικότητας και των προσωπικών δεδομένων και ακολουθήσαμε κανονισμούς όπως ιδίως ο Γενικός Κανονισμός Προστασίας Δεδομένων, με όλες τις προϋποθέσεις νόμιμης επεξεργασίας που αυτός θέτει, όπου μέσα από τα άρθρα και τις διατάξεις τους δημιουργούν τις κατάλληλες κατευθυντήριες γραμμές τις οποίες πρέπει να ακολουθήσουμε με συνέπεια κατά τη χρήση τέτοιων δεδομένων.

Στη συνέχεια «περνώντας» στο πρακτικό σκέλος της έρευνας, αρχικά δημιουργήσαμε τα δεδομένα υγείας βάσει των αναγκών της έρευνάς μας με κατάλληλα εργαλεία για το σκοπό αυτό και στη συνέχεια εξετάσαμε δύο τεχνικές με τις οποίες μπορούμε να αξιοποιήσουμε τα δεδομένα υγείας με ασφάλεια χωρίς να διακυβεύουμε προσωπικές πληροφορίες ατόμων και ακολουθώντας ρητά τους κανονισμούς που προστατεύουν τα άτομα αυτά και τα δεδομένα τους, την ψευδωνυμοποίηση και την ανωνυμοποίηση.

Κατά τη ψευδωνυμοποίηση εξετάσαμε αρκετές τεχνικές καταλήγοντας στην ασφαλέστερη κατά τις ανάγκες μας και τη κρίση μας, προκειμένου να εφαρμόσουμε ένα πρώτο επίπεδο ασφάλειας τους, αυτό της προστασίας από άμεση αναγνώριση. Καθώς η τεχνική αυτή αποδείχθηκε αρκετά χρήσιμη ως προς το διαμοιρασμό των δεδομένων εντός του περιβάλλοντος του νοσοκομείου, διαπιστώσαμε ότι δεν μπορεί να καταστήσει τα δεδομένα υγείας ως ανώνυμα και κατ' επέκταση δεν ικανοποιεί τις προϋποθέσεις διαμοιρασμού των δεδομένων αυτών σε τρίτες οντότητες όπως ερευνητικά κέντρα. Έτσι γεννήθηκαν οι πρώτες σημαντικές προκλήσεις, καθώς η ανάγκη εφαρμογής και της ανωνυμοποίησης κρίνεται απολύτως απαραίτητη ώστε τα δεδομένα υγείας να μπορούν να αξιοποιηθούν με ασφάλεια για τις ανάγκες έρευνας.

Έτσι, προκειμένου να μπορέσουμε να αξιοποιήσουμε τα δεδομένα υγείας για ανάγκες έρευνας, αποστέλλοντάς τα σε Ερευνητικό κέντρο, προχωρήσαμε σε διαδικασίες ανωνυμοποίησης, ως ένα επιπρόσθετο μέτρο ασφάλειας των ήδη ψευδωνυμοποιημένων δεδομένων υγείας. Κατά τη διαδικασία αυτή χρησιμοποιήσαμε κάποιες από τις βασικές τεχνικές της ανωνυμοποίησης, εξάγοντας εν΄ τέλει τα συμπεράσματά μας.

10.2 Συμπεράσματα

Μέσα από τα αποτελέσματά μας βγάλαμε πολύτιμα συμπεράσματα όπως την ιδιαίτερη προσοχή που απαιτεί η χρήση και η αξιοποίηση ευαίσθητων δεδομένων υγείας για τις ανάγκες έρευνας σε μία προσπάθεια να εξάγουμε για την έρευνα αυτή πολύτιμα συμπεράσματα και παράλληλα να διατηρούμε συμμόρφωση, νομική αλλά και ηθική, με τους κανονισμούς που προστατεύουν τα δεδομένα αυτά όπως τον GDPR. Η συμμόρφωση αυτή μπορεί να καλυφθεί μέσω των τεχνικών της ψευδωνυμοποίησης και της ανωνυμοποίησης, προστατεύοντας έτσι τα υποκείμενα των δεδομένων αυτών, ωστόσο σε πολλές περιπτώσεις (ειδικά σε μεγάλα σύνολα ρεαλιστικών δεδομένων) απαιτεί «θυσίες» ώστε να μπορέσει να επιτευχθεί. Συνεχίζοντας καταλήξαμε στο συμπέρασμα ότι η ψευδωνυμοποίηση από μόνη της φαίνεται ότι δεν αρκεί, τουλάχιστον για περιπτώσεις δεδομένων μεγάλης κλίμακας με υψηλούς κινδύνους, ώστε να καταστήσει τα δεδομένα αυτά σε μορφή που μπορούν να αξιοποιηθούν από τρίτες οντότητες καθιστώντας αναγκαίο το συνδυασμού της με τεχνικές ανωνυμοποίησης για τον ασφαλή διαμοιρασμό τους (ή, ισοδύναμα, για να είναι μία ψευδωνυμοποίηση ισχυρή φαίνεται ότι πρέπει να συνδυαστεί με τεχνικές ανωνυμοποίησης). Η ανωνυμοποίηση αποδείχτηκε μία αρκετά απαιτητική διαδικασία καθώς η αντιστάθμιση μεταξύ χρησιμότητας των δεδομένων και προστασίας από το ρίσκο αποκάλυψης απαιτεί ιδιαίτερη προσοχή και ειδικά σε περιπτώσεις μεγάλου όγκου δεδομένων

όπως αυτών της έρευνας μας όπως προαναφέραμε. Εν' κατακλείδι, λόγω της φύσης και του μεγέθους των δεδομένων που αξιοποιήσαμε και τις ανάγκες της έρευνάς μας, καθώς τα δεδομένα αυτά δεν επρόκειτο να διατεθούν δημόσια, πέραν του Ερευνητικού Κέντρου, είχαμε την ευχέρεια, να «θυσιάσουμε», με σχετικά μικρή απώλεια πληροφορίας και σχεδόν ελάχιστο συνολικό ρίσκο αποκάλυψης, κάποιες πιθανότητες αναγνώρισης, καθώς τα οφέλη έναντι των κινδύνων ήταν πολύ μεγαλύτερα.

Το βασικό συμπέρασμα που αναδεικνύεται είναι ότι δεν υπάρχει μία συγκεκριμένη λύση που να μπορεί να θεωρηθεί ως πανάκεια: κάθε περίπτωση είναι ξεχωριστή, με τις δικές της προκλήσεις και απαιτήσεις, οπότε απαιτείται μία «ad-hoc» προσέγγιση, προκειμένου η εκάστοτε περίπτωση να μελετηθεί συστηματικά για να καταλήξει κανείς στη βέλτιστη επιλογή (ή ενδεχομένως και για να κρίνει ότι δεν υπάρχει κατάλληλη επιλογή που να παρέχει τις επιθυμητές διασφαλίσεις). Στο πλαίσιο αυτό, είναι σημαντικό να ανακαλέσουμε την έννοια της εκτίμησης αντικτύπου ως προς τα προσωπικά δεδομένα που προβλέπει ο ΓΚΠΔ: επεξεργασίες δεδομένων υγείας μεγάλης κλίμακας εμπίπτουν σε αυτές που μπορούν να επιφέρουν κινδύνους για τα υποκείμενα των δεδομένων, οπότε μία εκτίμηση αντικτύπου, πριν την επεξεργασία, είναι υποχρεωτική: κατά την εκτίμηση αυτή, για περιπτώσεις όπως αυτή που εξετάστηκε στην έρευνά μας, διαφαίνεται ότι η εκτίμηση αντικτύπου θα πρέπει να ενσωματώνει μία αντίστοιχη προσέγγιση προκειμένου να τεκμηριωθεί αν υπάρχει (και, εάν ναι, ποια) μία κατάλληλη προσέγγιση ψευδωνυμοποίησης ή ανωνυμοποίησης.

10.3 Προκλήσεις Και Περιορισμοί

Κατά την έρευνά μας, ήρθαμε αντιμέτωποι με κάποιους περιορισμούς και αρκετές προκλήσεις. Οι περιορισμοί αυτοί οφείλονται κυρίως σε ηθικά και νομικά ζητήματα, καθώς το αντικείμενο της έρευνας είναι τα ευαίσθητα προσωπικά δεδομένα υγείας. Τα δεδομένα αυτά (αναφερόμενοι σε πραγματικά δεδομένα υγείας) δεν θα μπορούσαμε να τα εκμεταλλευτούμε, λόγω της δυσκολίας εύρεσής τους αρχικά, καθώς επίσης θα απαιτούσαν πολύ ιδιαίτερη προσοχή, ενώ η πιθανότητα να εμπίπταμε σε καταπάτηση της ιδιωτικότητας πραγματικών ασθενών, ήταν μεγάλη. Έτσι χρειάστηκε να δημιουργήσουμε εμείς τα δεδομένα πάνω στα οποία θα ξεκινούσαμε την έρευνα, το οποίο απαιτούσε αρκετό χρόνο μελέτης και εκπαίδευσης, τόσο των κατάλληλων εργαλείων για το σκοπό αυτό, όσο και της προσπάθειας δημιουργίας ρεαλιστικών δεδομένων υγείας που θα αντικατόπτριζαν σε μεγάλο βαθμό τα πραγματικά. Επίσης καθ' όλη τη διαδικασία της έρευνας

ήταν απαραίτητο να είμαστε σε πλήρη συμμόρφωση με του κανονισμούς του GDPR, κάτι που απαιτούσε ιδιαίτερη προσοχή.

Κατά την υλοποίηση της ψευδωνυμοποίησης καθώς και της ανωνυμοποίησης των δεδομένων υγείας συναντήσαμε αρκετές προκλήσεις. Αρχικά, η ψευδωνυμοποίηση απαιτούσε ιδιαίτερη προσοχή ως προς την επιλογή της κατάλληλης τεχνικής που θα δημιουργούσε ψευδώνυμα τα οποία θα ήταν αδύνατο να επαναπροσδιορίσουν το αρχικό γνώρισμα από μη εξουσιοδοτημένες οντότητες, ενώ παράλληλα θα μας έδινε τη δυνατότητα να τα αντιστοιχίσουμε σε περίπτωση ανάγκης. Εξετάζοντας ενδελεχώς τα δεδομένα μας μετά τη διαδικασία της ψευδωνυμοποίησης, ήρθαμε αντιμέτωποι με ένα ακόμα ζήτημα, διαπιστώνοντας ότι απαιτούνται περισσότερες ενέργειες ακόμα, πέραν της ψευδωνυμοποίησης, ώστε να μπορέσουμε να αξιοποιήσουμε τα δεδομένα αυτά για τις ανάγκες έρευνας. Ο προβληματισμός αυτός κατέστησε απαραίτητη την ανάγκη, τα δεδομένα αυτά να ανωνυμοποιηθούν, καθώς ακόμα και αν δεν μπορούσαν να ταυτοποιήσουν άμεσα κάποιο άτομο (λόγω της αντικατάστασης των αναγνωριστικών τους με ψευδώνυμα), υπήρχε σοβαρός κίνδυνος αποκάλυψής τους μέσω του συνδυασμού των ψευδο-αναγνωριστικών. Τέλος κάποια σημαντικά δίλλημα που παρουσιάστηκαν κατά τη διαδικασία της ανωνυμοποίησης, είχαν να κάνουν με τους στόχους της έρευνας των δεδομένων αυτών, την επιλογή των γνωρισμάτων που θα συμπεριληφθούν στη διαδικασία ανωνυμοποίησης ή όχι βάσει του αν προσφέρουν σε αυτή ή θα δημιουργούσαν απλά δυσκολίες και τον τρόπο αξιοποίησης κάποιων μη εμπιστευτικών γνωρισμάτων, τα οποία από τη μία δημιουργούσαν ζητήματα ταυτοποίησης ατόμων μέσα στο σύνολο δεδομένων, αλλά από την άλλη (αν τα ορίζαμε ως εμπιστευτικά ή ψευδο-αναγνωριστικά) θα επέφεραν μεγάλη απώλεια πληροφορίας καθιστώντας ανέφικτή τη διαδικασία ανωνυμοποίησης.

10.4 Θέματα Προς Μελλοντική Έρευνα

Κατά την έρευνα που πραγματοποιήσαμε, επικεντρωθήκαμε περισσότερο στην ψευδωνυμοποίηση δεδομένων υγείας, ως κύρια μέθοδο προστασίας των δεδομένων αυτών, αξιοποιώντας την ανωνυμοποίηση ως ένα πρόσθετο μέτρο που προσφέρει ακόμα μεγαλύτερη ασφάλεια, μελετώντας και χρησιμοποιώντας κυρίως τις πολύ βασικές τεχνικές της. Καθώς το ερευνητικό πεδίο των ηλεκτρονικών δεδομένων υγείας συνολικά, υπόσχεται «επαναστατικές» εξελίξεις, θα μπορούσαν να διερευνηθούν σε μελλοντικές εργασίες αρκετές και πιο προηγμένες τεχνικές και της ανωνυμοποίησης, που πιθανό να μπορούν να προσφέρουν ακόμη καλύτερα αποτελέσματα, ως προς την ασφαλή αξιοποίηση τέτοιων δεδομένων. Πέραν αυτού, αναλόγως της

περίπτωσης μπορεί να πρέπει να χρησιμοποιούνται προηγμένες τεχνικές κρυπτογράφησης, όπως π.χ. ασφαλείς υπολογισμοί (secure computations), είτε μόνες τους είτε συνδυαστικά με τεχνικές ανωνυμοποίησης. Μάλιστα, δεδομένου ότι πλέον ζούμε στην εποχή της τεχνητής νοημοσύνης και της μηχανικής μάθησης, όπου επιστημονικές έρευνες και στον τομέα της υγείας βασίζονται σε τέτοιους αλγορίθμους οι οποίοι «εκπαιδεύονται» από κατάλληλα σύνολα δεδομένων, αποτελεσματικές τεχνικές ανωνυμοποίησης χωρίς να επηρεάζεται δυσμενώς η ακρίβεια των αλγορίθμων πρέπει να εφαρμόζονται και στα δεδομένα αυτά.

Τέλος, ένα σημαντικό κομμάτι που μας απασχόλησε κατά τη περάτωση της μεταπτυχιακής διατριβής που θα μπορούσε να μελετηθεί σε μελλοντική έρευνα, βασιζόμενο στη πρόταση του Κανονισμού για τον Ευρωπαϊκό Χώρο Δεδομένων για την Υγεία, είναι η δημιουργία κατάλληλα διαμορφωμένων πλατφορμών με στόχο την ασφαλή ανταλλαγή δεδομένων υγείας μεταξύ εμπλεκόμενων φορέων ή και άλλων νοσοκομείων, αξιοποιώντας συνδυασμούς τεχνικών ψευδωνυμοποίησης/ανωνυμοποίησης αλλά και προηγμένες κρυπτογραφικές τεχνικές, σε πλήρη διαφάνεια και συμμόρφωση με τους κανονισμούς του GDPR καθώς και τη συγκατάθεση του ασθενούς όταν ενδείκνυται, με στόχο τη αποτελεσματικότερη υγειονομική του περίθαλψη.

Βιβλιογραφία

[1] Ε. Τσαγκρασούλη , «Εισαγωγή Στα Προσωπικά Δεδομένα Στον Τομέα Της Υγείας,» Πανεπιστήμιο Πειραιώς, Πειραιάς, 2020.

[2] European Commision, "Regulation of The European Parliament and of The Council on The European Health Data Space," Strasbourg, 2022.

[3] Ευρωπαική Ένωση, "Χάρτης Θεμελιωδών Δικαιωμάτων Της Ευρωπαικής Ένωσης," 18 12 2000. [Online]. Available: https://www.europarl.europa.eu/charter/pdf/text_el.pdf.

[4] European Commision, "What is Personal Data?," [Online]. Available: https://commission.europa.eu/law/law-topic/data-protection/reform/what-personal-data_en.

[5] Ευρωπαΐκή Ένωση, «Γενικός Κανονισμός Για τη Προστασία Δεδομένων,» 25 Μάιος 2018. [Ηλεκτρονικό]. Available: https://gdprinfo.eu/el.

[6] California State Legislature, "California Consumer Privacy Act," January 2018. [Online]. Available: https://oag.ca.gov/privacy/ccpa.

[7] Χ. Ακριβοπούλου, «Το Δικαίωμα Στην Προστασία Των Προσωπικών Δεδομένων Μέσα Από Το Φακό Του Δικαιώματος Στην Ιδιωτική Ζωή,» *Θεωρία Και Πράξη Του Διοικητικού Δικαίου,* Ιούλιος 2011.

[8] Ευρωπαική Ένωση, «Άρθρο 8 - Ευρωπαΐκή Σύμβαση Δικαιωμάτων του Ανθρώπου - Δικαίωμα σεβασμού της ιδιωτικής και οικογενειακής ζωής,» 4 Νοέμβριος 1950. [Ηλεκτρονικό]. Available: https://www.lawspot.gr/nomikes-plirofories/nomothesia/esda/arthro-8-eyropaiki-symvasi-dikaiomaton-toy-anthropoy-dikaioma.

[9] GDPRinformer, "7 Essential GDPR Data Processing Principles," 20 September 2017. [Online]. Available: https://gdprinformer.com/gdpr-articles/7-essential-gdpr-data-processing-principles. [Accessed 2 July 2023].

[10] Συμβούλιο Της Ευρώπης, «Ευρωπαΐκή Σύμβαση των Δικαιωμάτων του Ανθρώπου (ΕΣΔΑ),» 1950. [Ηλεκτρονικό]. Available: https://eur-lex.europa.eu/EL/legal-content/glossary/european-convention-on-human-rights-echr.html.

[11] Συμβούλιο της Ευρώπης, «Σύμβαση 108 του 1981,» Ευρωπαικό Κοινοβούλιο, 1981.

[12] Ευρωπαική Ένωση, «Επεξεργασία Δεδομένων Προσωπικού Χαρακτήρα - Οδηγία 95/46/ΕΚ,» 24 Οκτώβριος 1995. [Ηλεκτρονικό]. Available: https://eur-lex.europa.eu/legal-content/EL/TXT/HTML/?uri=LEGISSUM:l14012.

[13] Σύνταγμα Της Ελλάδας, «Άρθρο 9Α: (Προστασία Προσωπικών Δεδομένων),» 2001.

[14] Σύνταγμα Δημοκρατίας Της Κύπρου, «Άρθρο 15 Του Συντάγματος Της Κύπριακής Δημοκρατίας».

[15] European Union, "European Health Data Space," European Union, 2022.

[16] Enisa, "Recommendations on shaping technology according to GDPR provisions," Enisa, 2018.

[17] Enisa, "Deploying Pseudonymisation Techniques - The case of the Health Sector," Enisa, 2022.

[18] Enisa, "Pseudonymisation techniques and best practices," Enisa, 2019.

[19] Γ. Χριστοφίδη, «Πρακτική Μεθοδολογία Ανωνυμοποίησης Προσωπικών Δεδομένων,» Πάτρα, 2019.

[20] ARX, "ARX - Data Anonymization Tool | Privacy Criteria," ARX, [Online]. Available: https://arx.deidentifier.org/overview/privacy-criteria/.

[21] H. W. Stenersen, "Anonymization of Health Data - Anonymization Approaches, Data Utility," University of Oslo, Oslo, 2020.

[22] A. Aminifar, Y. Lamo, K. I. Pun and F. Rabbi, "A Practical Methodology for Anonymization of Structured Health Data," in *17th Scandinavian Conference on Health Informatics*, Oslo, 2019.

[23] Κ. Λιμνιώτης, «Ασφάλεια, Ιδιωτικότητα & Εμπιστοσύνη Στα Μεγάλα Δεδομένα | Προσωπικά Και Ανώνυμα Δεδομένα - Ανωνυμοποίηση Προσωπικών Δεδομένων,» Εκεφε Δημόκριτος - Πανεπιστήμιο Πελλοποννήσου.

[24] L. Kniola, «Plausible Adversaries in Re-Identification Risk Assessment,» PhUSE, Maidenhead, UK, 2017.

[25] DataSec, "Anonymization," [Online]. Available: https://www.imperva.com/learn/data-security/anonymization/.

[26] G. T. Duncan, S. A. Keller-McNulty and L. S. Stokes, "Disclosure Risk vs. Data Utility:," NISS, Pittsburgh, 2001.

[27] J. Domingo-Ferrer, D. Sanchez and J. Soria-Comas, "Data-base anonymization: privacy models, data utility, and microaggregation-based inter-model connections," Morgan & Claypool, 2016.

[28] Enisa, "Data Pseudonymization: Advanced Techniques & Use Cases," Enisa, 2021.

[29] J. Frankenfield, "Merkle Tree in Blockchain: What is and How it Works," Investopedia, 2021.

[30] U. Maurer, "Unifying Zero-Knowledge Proofs of Knowledge," in *Progress in Cryptology – AFRICACRYPT 2009*, Africa, 2009.

[31] H. Aamot, C. D. Kohl, D. Richter and P. Knaup-Gregori, "Pseudonymization of Patient Identifiers for Translational Research," BMC Medical Informatics and Decision Making, 2013.

[32] C. De Canniere, A. Biryukov and B. Preneel, "An introduction to Block Cipher Cryptanalysis," *Procceedings of the IEEE,* March 2006.

[33] B. E. Van Gastel, B. Jacobs and J. Popma, "Data Protection Using Polymorphic Pseudonymisation in a Large-Scale Parkinson's Disease Study," IOS Press Open Library, 2021.

[34] M. Gentili, S. Hajian and C. Castillo, "A Case Study of Anonymization of Medical Surveys," in *International Conference on Digital Health*, 2017.

[35] ARX, "ARX - Data Anonymization Tool," ARX, [Online]. Available: https://arx.deidentifier.org/.

[36] M. Bergeat, "A French Anonymization Experiment With Health Data," Paris, 2014.

[37] L. Grant, F. Pop and M. Koci, "The European Health Data Space: Strengthening patients' rights," EIPA, 2023.

[38] European Data Protection Board, "EDPB-EDPS Joint Opinion 03/2022 on the Proposal for a Regulation on the European Health Data Space," European Data Protection Board, 2022.

[39] Centers for Disease Control and Prevention, «Health Insurance Portability and Accountability Act of 1996,» 1996.

[40] J. De Groot, "What is HIPAA Compiliance?," 8 February 2023. [Online]. Available: https://www.digitalguardian.com/blog/what-hipaa-compliance.

[41] M. Okada, "Big data and real-world data-based medicine in the management of hypertension," Hypertension Research, Tokyo, 2020.

[42] Χ. Γαβαλάς, «Τα Real World Data αλλάζουν το σύστημα υγείας,» Underwriter, Ελλάδα, 2022.

[43] F. Liu and D. Panagiotakos, "Real-world Data: A Brief Review Of The Methods, Applications, Challenges And Opportunities," BMC Medical Research Methodology, 2022.

[44] Dragan, "Challenges in Using of Real-World Data in Healthcare," Climedo, 2022.

[45] K. Yasar, "Synthetic Data," TechTarget, 2023.

[46] NHS - England, "Exploring How To Create Mock Patient Data (Synthetic Data) From Real Patient Data," NHS - England, 2022. [Online]. Available: https://transform.england.nhs.uk/ai-lab/explore-all-resources/develop-ai/exploring-how-to-create-mock-patient-data-synthetic-data-from-real-patient-data/.

[47] A. Sims, "Data Synthesis for Healthcare | The Fake Data Spotlight Series," Tonic, 2022.

[48] Informatica, "What Is Data Quality".

[49] DatProf, «Definition Of Test Data,» DatProf, [Ηλεκτρονικό]. Available: https://www.datprof.com/solutions/what-is-test-data.

[50] GenerateData, «Generate Test Data,» [Ηλεκτρονικό]. Available: https://generatedata.com/.

[51] Online Data generator, «Online Test Data Generator,» [Ηλεκτρονικό]. Available: https://www.onlinedatagenerator.com/.

[52] Mockaroo, "Mockaroo | Random Data Generator," Mockaroo, [Online]. Available: https://www.mockaroo.com/.

[53] R. Sobti and G. Geetha, "Cryptographic Hash Functions: A Review," *Computer Science Issues,* vol. 9, no. 2, 2012.

[54] A. S. Gillis, «Rainbow Tables | Definition,» TechTarget.

[55] S. Falconer, "Does Hashing Sensitive Customer Data Protect Privacy?," Skyflow, 2022.

[56] U.S. Department of Commerce, "The Keyed-Hash Message Authentication Code," NIST, Gaithersburg, 2008.

[57] D. Arias, "Adding Salt to Hashing: A Better Way to Store Passwords," Auth0, 2021.

[58] Β. Ντόκας, «Είναι Τα Ανώνυμα Δεδομένα Πραγματικά Ανώνυμα;,» 2019.

[59] J. Li, J. Liu, M. Baig and R. Chi-Wing Wong, "Information Based Data Anonymization For Classification Utility," Elsevier, 2011.

[60] ARX, "ARX - Data Anonymization Tool | Transformation Models," Arx, [Online]. Available: https://arx.deidentifier.org/overview/transformation-models/.

[61] ARX, "ARX - Data Anonymization Tool | Utility Analysis," ARX, [Online]. Available: https://arx.deidentifier.org/anonymization-tool/analysis/.

[62] ARX, "ARX - Data Anonymization Tool | Risk Analysis," ARX, [Online]. Available: https://arx.deidentifier.org/anonymization-tool/risk-analysis/.

[63] OpenAIRE, "Amnesia | Data Anonymization," OpenAIRE, [Online]. Available: https://amnesia.openaire.eu/index.html.

[64] OpenAIRE, «Amnesia | Features,» [Ηλεκτρονικό]. Available: https://amnesia.openaire.eu/features.html.

[65] J. S.-P. Diaz and A. Lopez Garcia, "pyCanon: A Python Library To Check The Level Of Anonymity Of A Dataset," Arvix, Spain, 2022.

[66] IFCA, "pyCanon," [Online]. Available: https://github.com/IFCA/pycanon.

[67] ARX, «ARX - Data Anonymization Tool | Data Quality Models,» ARX, [Ηλεκτρονικό]. Available: https://arx.deidentifier.org/overview/metrics-for-information-loss/.

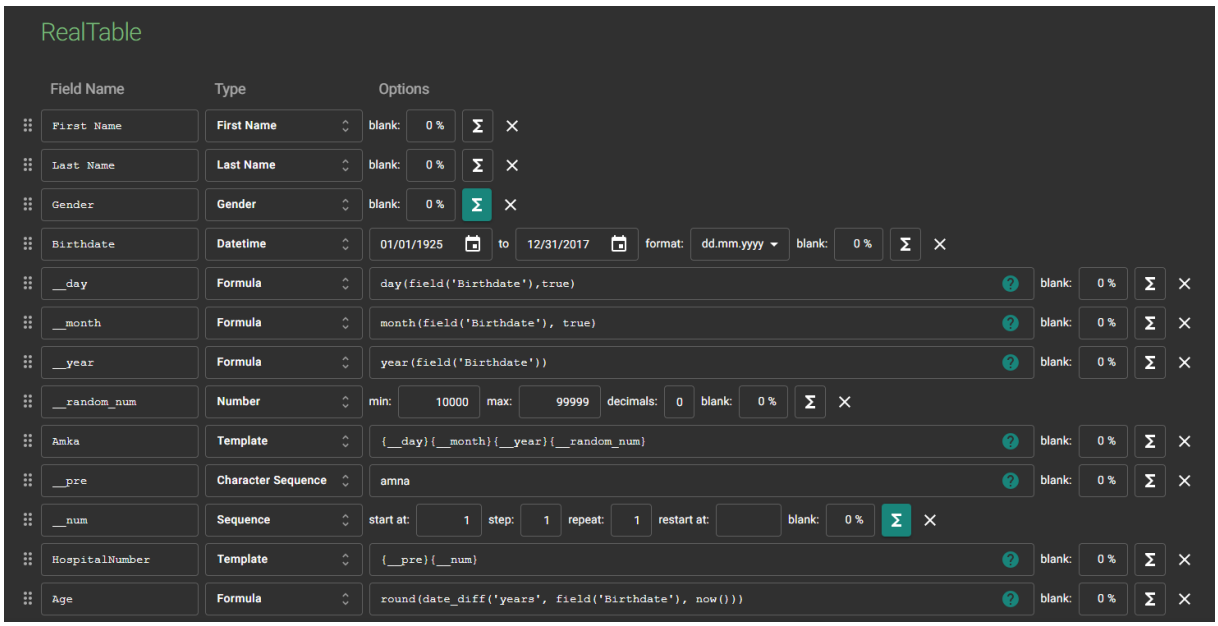
[68] ARX, "ARX - Data Anonymization Tool | Configuration," ARX, [Online]. Available: https://arx.deidentifier.org/anonymization-tool/configuration/.

Παράρτημα Α
Δεδομένα Υγείας

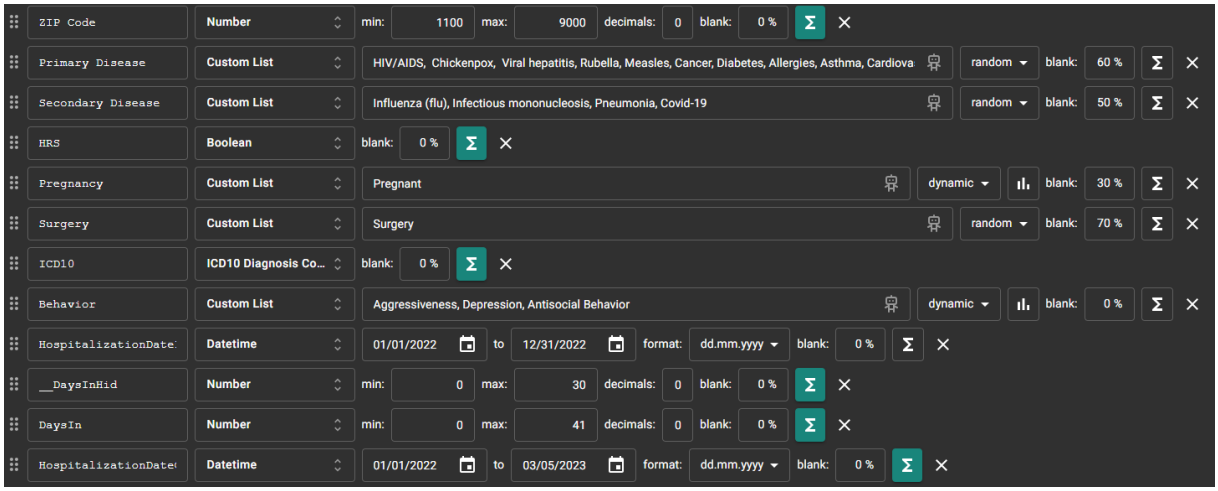
Στο παράρτημα αυτό παρουσιάζεται η δημιουργία των δεδομένων υγείας καθώς και οι συνθήκες που σχετίζονται με τις ασθένειες, τη επιδημία HRS καθώς την επιρροή της στη συμπεριφορά των ασθενών όπως διαμορφώθηκε με το εργαλείο Mockaroo.

Α.1 Δημιουργία Ρεαλιστικών δεδομένων

Παρακάτω (εικόνες Α-1.1 και Α-1.2) παρουσιάζονται τα δεδομένα όπως καταγράφονται σε μία καρτέλα ασθενούς, ενός Νοσοκομείου, κρατώντας δημογραφικές, ιατρικές και διαγνωστικές πληροφορίες.



Εικόνα Α-1.1: Mockaroo – Δημιουργία δεδομένων υγείας (1)



Εικόνα Α-1.2: Mockaroo – Δημιουργία δεδομένων υγείας (2)

Στην εικόνα Α-1.3 παρουσιάζονται οι συνθήκες υπό τις οποίες αλλάζει η συμπεριφορά ποσοστού των ασθενών που διαγνώστηκαν θετικοί στην επιδημία HRS. Χαρακτηριστικά παραδείγματα είναι, ότι διαβητικές γυναίκες σε εγκυμοσύνη που διαγνώστηκαν θετικές στη νόσο δείχνουν να παρουσιάζουν κατάθλιψη μετά τη νόσησή τους. Ενώ θετικοί ασθενείς μεγάλης ηλικιακής ομάδας παρουσιάζουν επιθετική συμπεριφορά.

	Rule	Aggressiveness	Depression	Antisocial Behavior
✕ ↑ ↓	field('Pregnancy') == "Pregnant"	0	1	0
✕ ↑ ↓	field('Age') >= 70 and field('HRS	1	0	0
✕ ↑ ↓	field('Primary Disease') == "Card	0	1	0
✕ ↑ ↓	field('Primary Disease') == "HIV/	0	0	1
✕ ↑ ↓	field('Gender') == "Male" and fie	0	0	1
✕ ↑ ↓	field('Pregnancy') != "Pregnant"	0	0	0
✕ ↑ ↓	field('Age') < 70 or field('HRS')	0	0	0
✕ ↑ ↓	field('Primary Disease') != "Canc	0	0	0

ADD A RULE or ADD RULES FOR ALL VALUES OF...

Εικόνα Α-1.3: Mockaroo – Συνθήκες συμπεριφοράς ασθενούς

Οι ημέρες νοσηλείας των ασθενών επηρεάζονται από συνθήκες που αφορούν την ηλικιακή τους ομάδα, τη πρωταρχική καθώς και τη δευτερεύουσα ασθένεια τους, σε κάποιες περιπτώσεις συνδυαστικά, όπως βλέπουμε στην εικόνα Α-1.4

```
if field("Secondary Disease") == 'Influenza (flu)' and field("Primary Disease") != 'Asthma' and  field("Age") < 35 then random(0, 3)
elsif field("Secondary Disease") == 'Influenza (flu)' and field("Primary Disease") == 'Asthma' and  field("Age") < 35 then random(0, 6)
elsif field("Secondary Disease") == 'Influenza (flu)' and field("Primary Disease") == 'Asthma' and field("Age") >= 35 and field("Age") < 60 then random(2, 10)
elsif field("Secondary Disease") == 'Influenza (flu)' and field("Primary Disease") != 'Asthma' and field("Age") >= 35 and field("Age") < 60 then random(0, 8)
elsif field("Secondary Disease") == 'Influenza (flu)' and field("Primary Disease") == 'Asthma' and field("Age") >= 60 then random(5, 15)
elsif field("Secondary Disease") == 'Influenza (flu)' and field("Primary Disease") != 'Asthma' and field("Age") >= 60 then random(3, 12)
elsif field("Secondary Disease") == 'Infectious mononucleosis' and field("Age") < 35 then random(0, 4)
elsif field("Secondary Disease") == 'Infectious mononucleosis' and field("Age") >= 35 and field("Age") < 55 then random(0, 8)
elsif field("Secondary Disease") == 'Infectious mononucleosis' and field("Age") >= 55 then random(4, 14)
elsif field("Secondary Disease") == 'Pneumonia' and field("Age") < 35 then random(5, 12)
elsif field("Secondary Disease") == 'Pneumonia' and field("Age") >= 35 and field("Age") < 65 then random(4, 20)
elsif field("Secondary Disease") == 'Pneumonia' and field("Age") >= 65 then random(8, 30)
elsif field("Secondary Disease") == 'Covid-19' and field("Primary Disease") == 'Cancer' and field("Age") >= 30 and field("Age") < 65 then random(0, 12)
elsif field("Secondary Disease") == 'Covid-19' and field("Primary Disease") != 'Cancer' and field("Age") >= 30 and field("Age") < 65 then random(0, 7)
elsif field("Secondary Disease") == 'Covid-19' and field("Primary Disease") == 'Cancer' and field("Age") >= 65 then random(4, 25)
elsif field("Secondary Disease") == 'Covid-19' and field("Primary Disease") != 'Cancer' and field("Age") >= 65 then random(0, 18)
elsif field("Secondary Disease") == 'Covid-19' and field("Primary Disease") == 'Cancer' and field("Age") < 30 then random(0, 8)
elsif field("Secondary Disease") == 'Covid-19' and field("Primary Disease") != 'Cancer' and field("Age") < 30 then random(0, 2)
elsif field("Secondary Disease") == 'Covid-19' and field("Primary Disease") == 'Diabetes' and field("Age") >= 30 and field("Age") < 65 then random(0, 10)
elsif field("Secondary Disease") == 'Covid-19' and field("Primary Disease") != 'Diabetes' and field("Age") >= 30 and field("Age") < 65 then random(0, 7)
elsif field("Secondary Disease") == 'Covid-19' and field("Primary Disease") == 'Diabetes' and field("Age") >= 65 then random(4, 19)
elsif field("Secondary Disease") == 'Covid-19' and field("Primary Disease") != 'Diabetes' and field("Age") >= 65 then random(0, 11)
elsif field("Secondary Disease") == 'Covid-19' and field("Primary Disease") == 'Diabetes' and field("Age") < 30 then random(0, 7)
elsif field("Secondary Disease") == 'Covid-19' and field("Primary Disease") != 'Diabetes' and field("Age") < 30 then random(0, 2)
else random(0, 3) end
```

Εικόνα Α-1.4: Mockaroo – Συνθήκες ημερών νοσηλείας

Η πιθανότητα εισαγωγής τους λόγω κάποιας χειρουργικής επέμβασης, αυξάνει τις ημέρες νοσηλείας τους.

```
if field("Surgery") == 'Surgery' then field("__DaysInHid") + random(1, 10)
else field("__DaysInHid") end
```

Εικόνα Α-1.7: Mockaroo – Ημέρες νοσηλείας βάσει χειρουργείου

```
field("HospitalizationDateIn") + days(field("DaysIn"))
```

Εικόνα Α-1.6: Mockaroo – Σύνολο ημερών νοσηλείας

Παράρτημα Β
Ψευδωνυμοποίηση Δεδομένων

Για την ψευδωνυμοποίηση των δεδομένων μας δημιουργήσαμε ένα script το οποίο λαμβάνει συγκεκριμένα στοιχεία από κάθε εγγραφή του αρχικού πίνακα και δημιουργεί ένα νέο πίνακα με τα ψευδώνυμά τους μέσω συγκεκριμένης διαδικασίας.

Β.1 Κώδικας Ψευδωνυμοποίησης

Όπως παρατηρούμε στην εικόνα Β-1.1 παρακάτω, ο κώδικας που δημιουργήθηκε, διαβάζει συγκεκριμένες στήλες από τον αρχικό πίνακα με τα αναγνωριστικά σε ένα εύρος που του έχουμε δώσει, στη συνέχεια δημιουργεί ένα 'salt' ως πρόθεμα που θα προστεθεί στην αρχή της κάθε εγγραφής. Στη συνέχει δημιουργεί ένα ακόμα φύλλο, όπου θα αποθηκεύσει τις ψευδωνυμοποιημένες τιμές. Τέλος ψευδωνυμοποιεί γραμμή-γραμμή την κάθε εγγραφή μαζί με το πρόθεμά της, με τη χρήση συνάρτησης κατακερματισμού SHA-256, αποθηκεύοντας τις στο νέο φύλλο εργασίας και αποθηκεύει το αρχείο.

```powershell
# Load the Excel COM object
$excel = New-Object -ComObject Excel.Application

# Open the Excel workbook
$workbook = $excel.Workbooks.Open("C:\Users\Doulge\Desktop\AYD_Thesis\Script\Pseudonymized_Data.xlsx")

# Select the worksheet and range of cells to read
$worksheet = $workbook.Worksheets.Item("Sheet2")
$range = $worksheet.Range("E2:E100001")

# Create a salt to use for hashing
$salt = "wT8s45_"

# Create a new worksheet to hold the hashed values
$newWorksheet = $workbook.Worksheets.Add()
$newWorksheet.Name = "Hashes"

# Loop through each cell in the range and hash its value with the salt
$sha256 = New-Object -TypeName System.Security.Cryptography.SHA256CryptoServiceProvider
$row = 1
foreach ($cell in $range) {
    $value = $cell.Value2
    $valueWithSalt = $salt + $value
    $hash = [System.BitConverter]::ToString($sha256.ComputeHash([System.Text.Encoding]::UTF8.GetBytes($valueWithSalt)))
    $hash = $hash.Replace("-", "").ToLower()
    $newWorksheet.Cells.Item($row, 1) = $hash
    $row++
}

# Save the workbook and close Excel
$workbook.Save()
$workbook.Close($false)
$excel.Quit()
```

Εικόνα Β-1.1: Κώδικας ψευδωνυμοποίησης

Παράρτημα Γ
Αποτελέσματα Ανωνυμοποίησης

Στο παράρτημα αυτό φαίνονται σε μορφή πίνακα τα γνωρίσματα των αρχικών δεδομένων υγείας καθώς και όλες οι χρήσιμες προσεγγίσεις της ανωνυμοποίησης των δεδομένων υγείας, όπως διαμορφώθηκαν κατά τη διαδικασία. Επισημαίνεται επίσης η αποτελεσματικότερη προσέγγιση που επιλέξαμε κατά τη κρίση μας κατηγοριοποιώντας όλα τα γνωρίσματα ανάλογα.

Γ.1 Γνωρίσματα Δεδομένων Υγείας

Παρακάτω βλέπουμε τα γνωρίσματα του αρχικού πίνακα των δεδομένων υγείας όπως δημιουργήθηκαν κατά τη διαδικασία της συλλογής δεδομένων. Αξιοσημείωτο είναι το γεγονός ότι δεν συμπεριλήφθηκαν όλα τα αρχικά γνωρίσματα στην διαδικασία ανωνυμοποίησης καθώς κάποια εξ' αυτών δεν προσέφεραν κανένα όφελος στην έρευνα, αλλά ούτε επηρεαζόντουσαν από αυτή.

Attribute	**Μορφή**	**Category**

First Name	String	Identifier
Lat Name	String	Identifier
AMKA	Number Sequence	Identifier
Hospital Number	Alphanumeric	Identifier
Gender	String	Quasi-Identifier
DOB	Date	Quasi-Identifier
Age	Integer	Quasi-Identifier
ZIP Code	Random Number Sequence	Quasi-Identifier
Hospitalization Date In / Out	Date	Sensitive
Days In	Integer	Sensitive
HRS	Boolean	Sensitive
Primary Disease	Custom List	Sensitive
Secondary Disease	Custom List	Sensitive
Pregnancy	Custom List	Insensitive
Surgery	Custom List	Sensitive
Patient Behavior	Custom List	Sensitive

Πίνακας Γ-1.2: Κατηγοριοποίηση γνωρισμάτων δεδομένων Υγείας

Γ.1 Προσεγγίσεις Ανωνυμοποίησης

Προσέγγιση 1	**k = 5**	**HRS**	**l = 2**
Μετασχηματισμός	**Απώλεια**	**Κατάργηση**	
[0, 1, 3]	1,42%	0,012%	
	Ρίσκο	**Μέγιστο Ρίσκο**	**Επηρεαζόμενες Εγγραφές**
	0%	0,47%	0,21%
	Κλάσεις Ισοδυναμίας		
	Min	**Max**	**Mean**
	211	421	328
[0, 1, 2]	**Απώλεια**	**Κατάργηση**	
	1,12%	0,012%	
	Ρίσκο	**Μέγιστο Ρίσκο**	**Επηρεαζόμενες Εγγραφές**
	0%	8,33%	0,012%
	Κλάσεις Ισοδυναμίας		
	Min	**Max**	**Mean**
	12	61	33
Προσέγγιση 2	**k = 5**	**HRS**	**l = 2**
		Behavior	**l = 3**
Μετασχηματισμός	**Απώλεια**	**Κατάργηση**	
[0, 1, 3]	15,87%	31,02%	
	Ρίσκο	**Μέγιστο Ρίσκο**	**Επηρεαζόμενες Εγγραφές**
	0%	7,14%	0,067%
	Κλάσεις Ισοδυναμίας		
	Min	**Max**	**Mean**
	241	414	331
Προσέγγιση 3	**k = 5**	**HRS**	**l = 2**
		Behavior	**l = 2**
Μετασχηματισμός	**Απώλεια**	**Κατάργηση**	
[0, 1, 3]	1,42%	0,012%	
	Ρίσκο	**Μέγιστο Ρίσκο**	**Επηρεαζόμενες Εγγραφές**
	0%	0,47%	0,21%
	Κλάσεις Ισοδυναμίας		
	Min	**Max**	**Mean**
	211	421	328

[0, 1, 2]	**Απώλεια**	**Κατάργηση**	
	15,92%	16,19%	
	Ρίσκο	**Μέγιστο Ρίσκο**	**Επηρεαζόμενες Εγγραφές**
	0%	8,33%	0,014%
	Κλάσεις Ισοδυναμίας		
	Min	**Max**	**Mean**
	12	61	33
Προσέγγιση 5	**k = 5**	**HRS**	**l = 2**
		Behavior	**l = 2**
		Primary Disease	**l = 5**
Μετασχηματισμός	**Απώλεια**	**Κατάργηση**	
[0, 1, 3]	1,42%	0,012%	
	Ρίσκο	**Μέγιστο Ρίσκο**	**Επηρεαζόμενες Εγγραφές**
	0%	0,47%	0,21%
	Κλάσεις Ισοδυναμίας		
	Min	**Max**	**Mean**
	211	421	328
[0, 1, 2]	**Απώλεια**	**Κατάργηση**	
	16,35%	16,62%	
	Ρίσκο	**Μέγιστο Ρίσκο**	**Επηρεαζόμενες Εγγραφές**
	0%	8,33%	0,014%
	Κλάσεις Ισοδυναμίας		
	Min	**Max**	**Mean**
	12	61	33
Προσέγγιση 6	**k = 5**	**HRS**	**l = 2**
		Behavior	**l = 3**
		Primary Disease	**l = 5**
Μετασχηματισμός	**Απώλεια**	**Κατάργηση**	
[0, 1, 3]	15,87%	31,002%	
	Ρίσκο	**Μέγιστο Ρίσκο**	**Επηρεαζόμενες Εγγραφές**
	0%	0,41%	0,34%
	Κλάσεις Ισοδυναμίας		
	Min	**Max**	**Mean**

	241	414	331
Προσέγγιση 7	**k = 5**	**HRS**	**l = 2**
		Behavior	**l = 2**
		Primary Disease	**l = 5**
		Secondary Disease	**l = 3**
Μετασχηματισμός	**Απώλεια**	**Κατάργηση**	
[0, 1, 3]	1,42%	0,012%	
	Ρίσκο	**Μέγιστο Ρίσκο**	**Επηρεαζόμενες Εγγραφές**
	0%	0,47%	0,21%
	Κλάσεις Ισοδυναμίας		
	Min	**Max**	**Mean**
	211	421	328
[0, 1, 2]	**Απώλεια**	**Κατάργηση**	
	16,35%	16,62%	
	Ρίσκο	**Μέγιστο Ρίσκο**	**Επηρεαζόμενες Εγγραφές**
	0%	8,33%	0,014%
	Κλάσεις Ισοδυναμίας		
	Min	**Max**	**Mean**
	12	61	33

Πίνακας Γ-1.2: Συνολικός πίνακας προσεγγίσεων ανωνυμοποίησης