ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

UNIVERSITY OF PIRAEUS

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Πρόγραμμα Μεταπτυχιακών Σπουδών

«Προηγμένα Συστήματα Πληροφορικής»

Μεταπτυχιακή Εργασία

Τίτλος:

«Χρήση Big Data στον τομέα της Υγείας»

«The use of BIG DATA in healthcare systems»

Μεταπτυχιακός Φοιτητής: **Δερβένης Κωνσταντίνος**

Πατρωνυμο : **Παύλος**

Αριθμός Μητρώου: **ΜΠΣΠ 16006**

Επιβλέπων: **Αποστόλου Δημήτριος, Αναπληρωτής Καθηγητής**

Οκτώβριος , 2019

Τριμελής Εξεταστική Επιτροπή

(Υπογραφή)	(Υπογραφή)	(Υπογραφή)
Όνομα Επώνυμο Βαθμίδα	Όνομα Επώνυμο Βαθμίδα	Όνομα Επώνυμο Βαθμίδα
Αποστόλου Δ.	Μεταξιώτης Κ.	Κοτζανικολάου Π.
Αναπλ. Καθηγητής	Καθηγητής	Επ. Καθηγητής

Πίνακας Περιεχομένων

Πίνακας Περιεχόμενων Σχημάτων

Ευχαριστίες

Η ολοκλήρωση του παρόντος Μεταπτυχιακού Προγράμματος Σπουδών αποτελεί το αποτέλεσμα όχι ατομικής, αλλά ομαδικής εργασίας. Οφείλω λοιπόν ένα θερμό ευχαριστώ στα μέλη της ομάδας μου. Ευχαριστώ από καρδιάς τον επιβλέποντα Καθηγητή μου κ. Αποστόλου Δημήτριο για τη στήριξη και τις συμβουλές του κατά τη συγγραφή της ανά χείρας διπλωματικής εργασίας. Επιπλέον, ευχαριστώ θερμά οικογένεια και φίλους για τη συνεχή υποστήριξη και κατανόηση τους.

Συντομογραφίες

API – Application Programming Interface

CDSS – Clinical Decision Support System

CPOE – Computerized Physician Order Entry

CT – Computed Tomography

EMR – Electronic Medical Records

FHIR – Fast Healthcare Interoperability Resource

GWAS – Genome-Wide Association Studies

HDFS - Hadoop Distributed File System

HIS – Healthcare Information Systems

ICD – International Classification of Diseases

ICU – Intensive Care Unit

IoT – Internet of Things

JSON - Java Script Object Notation

LOS – Length of Stay

MRI – Magnetic Resonance Imaging

NLP –

PACS - Picture Archiving and Communication System

PHI – Protected Health Information

R & D – Research & Development

RDD – Resilient Distributed Dataset

SQL – Structured Query Language

XML – Extensible Markup Language

Ε & Α – Έρευνα & Ανάπτυξη

ΗΦΥ – Ηλεκτρονικός Φάκελος Υγείας

ΜΕΘ – Μονάδα Εντατικής Θεραπείας

ΤΠΕ – Τεχνολογίες των Πληροφοριών και της Επικοινωνίας

Περίληψη

Όπως υποδηλώνει και το όνομα τους, τα "μεγάλα δεδομένα" αντιπροσωπεύουν μεγάλες ποσότητες δεδομένων σε διάφορους τομείς, τα οποία δεν είναι διαχειρίσιμα με τη χρησιμοποίηση των παραδοσιακών λογισμικών ή διαδικτυακών πλατφορμών, δεδομένου ότι υπερβαίνουν κατά πολύ τις υφιστάμενες δυνατότητες σε όρους αποθήκευσης, επεξεργασίας και αναλυτικής ισχύος. Ειδικότερα στον τομέα της υγείας, ο οποίος παραδοσιακά παράγει τεράστιους όγκους δεδομένων, η αξιοποίηση των Big Data μέσω εξειδικευμένων εφαρμογών υπόσχεται σημαντικές εξελίξεις σε όρους διαχείρισης, λήψης αποφάσεων, παροχής προβλέψεων για εμφάνιση ασθενειών και επιδημιών, καθώς και σημαντική μείωση του κόστους υγειονομικής φροντίδας. Σκοπός της παρούσας εργασίας είναι να διερευνήσει το ρόλο των Big Data στον τομέα της υγείας, από κοινού με τις προκλήσεις και τις ευκαιρίες που αυτά αντιμετωπίζουν, καθώς επίσης τις κύριες εφαρμογές τους. Από μεθοδολογικής πλευράς, πραγματοποιήθηκε μια βιβλιογραφική ανασκόπηση, η οποία βασίστηκε στην αναζήτηση σχετικών άρθρων σε μια σειρά ηλεκτρονικών βάσεων δεδομένων όπως το Google Scholar, το Scopus, το Pub Med και το Research Gate.

Λέξεις Κλειδιά: Μεγάλα Δεδομένα, αναλυτική, Αναλυτική των Μεγάλων Δεδομένων, εφαρμογές, τομέας της υγείας

Abstract

As their name implies, "big data" stand for enormoue amounts of data, produced in various sectors, that cannot be managed through using traditional software or web platforms, as they by far outstrip existing capabilities in terms of storage, processing and analytics. Particularly in the health sector, which traditionally generates huge volumes of data, the utilization of Big Data through specialized applications promises significant developments in terms of management, decision making, disease and epidemic forecasting, and significant reductions in healthcare costs. The purpose of this thesis is to explore the role of Big Data in the healthcafre sector, along with the challenges and opportunities they face, as well as their main applications. From a methodological point of view, a bibliographic review was conducted, based on research for related articles in a number of online databases such as Google Scholar, Scopus, Pub Med, and Research Gate.

Key Words: Big Data, analytics, Big Data Analytics, applications, health care sector

ΕΙΣΑΓΩΓΗ

Από τα μέσα της δεκαετίας του '80, έχει πραγματοποιηθεί, σε παγκόσμιο επίπεδο, μια άνευ προηγουμένου έκρηξη στην ικανότητα παραγωγής, αποθήκευσης και αποστολής δεδομένων, κυρίως σε ψηφιακή μορφή. Μια άλλη εξέλιξη που αποτελεί συνισταμένη αυτής της τάσης, είναι η αυξημένη πρόσβαση σε Τεχνολογίες των Πληροφοριών και της Επικοινωνίας (ΤΠΕ), η οποία καθίσταται εφικτή μέσω των προσωπικών υπολογιστών, έξυπνων τηλεφώνων και άλλων φορητών συσκευών. Οι ως άνω βελτιωμένες δυνατότητες σε όρους αποθήκευσης δεδομένων, από κοινού με την υψηλής ταχύτητας υπολογιστική δύναμη, καθώς και την παράδοση πληροφοριών σε πραγματικό χρόνο μέσω του Διαδικτύου, οδήγησε σε εκθετική αύξηση της μέσης ημερήσιας κατανάλωσης δεδομένων του ατόμου [1].

Υπό το πρίσμα της αξιοσημείωτης αύξησης των παραγόμενων δεδομένων, σημειώθηκε μια ταχεία ψηφιοποίηση σε όλες τις βιομηχανίες. Ο ψηφιακός αυτός μετασχηματισμός έχει συντελεστεί και στον τομέα της υγείας υπό τη μορφή της αύξησης της χρήσης αφενός των Ηλεκτρονικών Ιατρικών Αρχείων (EMR[1]) και των Πληροφοριακών Συστημάτων Υγείας (HIS[2]) και αφετέρου των φορητών και έξυπνων συσκευών. Ως αποτέλεσμα, υπάρχει πλέον διαθέσιμη μια τεράστια ποσότητα και ποικιλία δεδομένων που σχετίζονται με την υγεία και είναι ψηφιακής μορφής, η οποία πέρα από τα κλινικά δεδομένα, περιλαμβάνει βιολογικά δεδομένα (-omics data), δεδομένα που αφορούν κοινωνιο-δημογραφικά στοιχεία ή ακόμα και ασφαλιστικές αξιώσεις. Παρά το γεγονός ότι αυτά τα δεδομένα υψηλής ποιότητας, είναι δυνητικά αυξημένης σημασίας για τη βελτιστοποίηση της παροχής υγειονομικής φροντίδας, εξακολουθούν να γίνονται αντιληπτά περισσότερο ως υποπροϊόν της παροχής υγειονομικής περίθαλψης και λιγότερο ως κεντρική πηγή στοιχείων, κρίσιμων για την επίτευξη ανταγωνιστικών πλεονεκτημάτων στον τομέα της υγείας. Δεδομένου ότι τα ηλεκτρονικά δεδομένα για την υγεία παραμένουν σε μεγάλο βαθμό αναξιοποίητα, είναι επιτακτική η ανάγκη μετατροπής των με επεξεργασμένων δεδομένων σε χρήσιμες και αποτελεσματικές πληροφορίες [2].

Ο τομέας της υγείας ανέκαθεν δημιουργούσε μεγάλο όγκο δεδομένων, λόγω των αυξημένων αναγκών τήρησης αρχείων στο πλαίσιο της περίθαλψης των ασθενών [3].

[1] Electronical Medical Record
[2] Healthcare Information Systems

Πολλά από αυτά τα διαθέσιμα και ιδιαίτερα πολύτιμα δεδομένα είναι σε μη δομημένη ή ημι-δομημένη μορφή. Περαιτέρω, τα περίπλοκα, δυναμικά και ετερογενή χαρακτηριστικά των δεδομένων αυτών [4, 6] καθιστούν δύσκολη την εξαγωγή χρήσιμης πληροφορίας μέσω της χρησιμοποίησης των παραδοσιακών αναλυτικών εργαλείων και τεχνικών. Έτσι, τα Big Data (ή μεγάλα δεδομένα) στον τομέα της υγείας αποτελούν ένα σημαντικό ζήτημα, όχι μόνο λόγω του τεράστιου όγκου τους, αλλά και λόγω της ποικιλότητας των δεδομένων και επομένως της ταχύτητας διαχείρισης τους [4]. Στην πραγματικότητα, ο άνθρωπος έχει πεπερασμένη ικανότητα να επεξεργάζεται αυτά τα δεδομένα και ως εκ τούτου είναι αναγκαία η αποτελεσματική υποστήριξη της λήψης αποφάσεων. Αυτό δημιουργεί την ανάγκη ενσωμάτωσης των Big Data Analytics (Αναλυτική των Μεγάλων Δεδομένων) στον τομέα της υγείας [7].

Τα Big Data Analytics έχουν τη δυνατότητα να αναλύουν μια ευρεία γκάμα σύνθετων δεδομένων και να παράγουν πολύτιμες πληροφορίες που διαφορετικά δεν θα ήταν εφικτό να εξαχθούν. Όταν εφαρμόζονται στα δεδομένα του τομέα της υγείας, έχουν τη δυνατότητα όχι μόνο να εντοπίσουν τα αναδυόμενα μοτίβα, αλλά και να οδηγήσουν στη βελτίωση της ποιότητας της υγειονομικής περίθαλψης, τη μείωση του κόστους και στην έγκαιρη λήψη αποφάσεων [7]. Όπως χαρακτηριστικά αναφέρεται στην έκθεση του Διεθνούς Ινστιτούτου McKinsey [8], αν αξιοποιηθούν αποτελεσματικά τα Big Data, το αμερικανικό υγειονομικό σύστημα αξία θα να εξοικονομεί παραπάνω από 300 δισεκατομμύρια δολάρια ετησίως, εκ των οποίων τα δύο τρίτα θα έχουν τη μορφή μείωσης των δαπανών υγειονομικής περίθαλψης κατά περίπου 8%. Με τη χρήση της τεχνολογίας των Big Data και τη χρησιμοποίηση της αυτοματοποιημένης ανάλυσης των αποτελεσμάτων, είναι δυνατόν να αναδυθούν χρήσιμες πληροφορίες, οι οποίες μέχρι πρότινος παρέμεναν στην αφάνεια [7].

Η πρόοδος στην Υπολογιστική Νέφους (Cloud Computing) και η αυξημένη ανάπτυξη των ηλεκτρονικών ιατρικών αρχείων επιτρέπουν πλέον την εύκολη πρόσβαση σε διαχρονικά δεδομένα ασθενών. Με την πάροδο του χρόνου, η ενοποίηση αυτών των διαχρονικών δεδομένων των ασθενών με δεδομένα από διαφορετικές, δομημένες και μη δομημένες πηγές Big Data προσφέρει τη δυνατότητα πλήρους κατανόησης των ασθενειών. Η ικανότητα των Big Data Analytics να εντοπίζουν την ετερογένεια των ασθενειών επιτρέπει όχι μόνο την έγκαιρη και ακριβή διάγνωση τους, αλλά και την αξιολόγηση των υφιστάμενων θεραπειών. Μέσω

της συσχέτισης δεδομένων από διαφορετικές πηγές και μέσω του εντοπισμού διαφόρων διακριτών μοτίβων, η προγνωστική δύναμη των Big Data Analytics μπορεί επίσης να αξιοποιηθεί για τη μετατροπή συνεχόμενων δεδομένων σε πολύτιμες πληροφορίες σε πραγματικό χρόνο. Αυτή η δυνατότητα που παρέχει η αναλυτική των μεγάλων δεδομένων είναι ιδιαιτέρως σημαντική σε ιατρικές καταστάσεις έκτακτης ανάγκης, καθώς μπορεί να κρίνει τη διαφορά μεταξύ ζωής και θανάτου [7].

Η αποθήκευση και τήρηση των Big Data στον τομέα της υγείας υπόσχεται να βελτιώσει την ποιότητα της υγειονομικής περίθαλψης, μειώνοντας ταυτόχρονα το κόστος. Μεταξύ άλλων έχει τη δυνατότητα να υποστηρίξει διάφορες λειτουργίες που συντελούνται στο πλαίσιο της ιατρικής και υγειονομικής περίθαλψης, συμπεριλαμβανομένης της κλινικής υποστήριξης λήψης αποφάσεων, της επιδημιολογικής επιτήρησης και της διαχείρισης της υγείας του πληθυσμού. Βάσει αναφορών, τα δεδομένα για την υγεία που διατηρούνται μόνο στις ΗΠΑ υπερβαίνουν τα 150 exabytes (10^{18} bytes) το 2011 και έχουν την ικανότητα να υπερβούν ακόμα μεγαλύτερες τάξεις μεγεθών όπως τα zettabytes (10^{21} bytes) και τα yottabytes (10^{24} gigabytes). Γίνεται αντιληπτό ότι λόγω του τεράστιου όγκου των Big Data, η ανάλυση τους αποτελεί πραγματικό ζήτημα. Ως εκ τούτου, οι ειδικοί στον τομέα της υγείας αναζητούν τη συνδρομή των επιστημών τις πληροφορικής, προκειμένου να διερευνήσουν και να βρουν λύσεις για τη μετατροπή των δεδομένων σε πληροφορία και γνώση [2].

Ο ταχέως αναπτυσσόμενος τομέας των Big Data Analytics έχει αρχίσει να διαδραματίζει κεντρικό ρόλο στην εξέλιξη των πρακτικών και της έρευνας στον τομέα της υγείας. Ειδικότερα, η αναλυτική των μεγάλων δεδομένων έχει παράσχει εργαλεία για τη συσσώρευση, τη διαχείριση, την ανάλυση και την αφομοίωση μεγάλων όγκων διαφορετικών, δομημένων και μη δομημένων δεδομένων που παράγονται από τα τρέχοντα συστήματα υγειονομικής περίθαλψης. Ωστόσο, παρά την αξιοσημείωτη προοπτική της σε όρους βελτίωσης της παροχής υγειονομικής φροντίδας και της μεγαλύτερης κατανόησης των ασθενειών, η υιοθέτηση και η ανάπτυξη της έρευνας σε αυτό το χώρο, εξακολουθούν να υστερούν σημαντικά και παρεμποδίζονται από ορισμένα βασικά προβλήματα που ενυπάρχουν στη θεματική των Big Data [9].

Η πολλά υποσχόμενη αξία της τεχνολογίας των Big Data στον τομέα της υγείας έχει δημιουργήσει ένα αυξανόμενο ενδιαφέρον μεταξύ των ερευνητών της ακαδημαϊκής και βιομηχανικής κοινότητας. Παρ' όλα αυτά, είναι δυνατό να εντοπιστούν περιορισμένος αριθμός σχετικών ανασκοπήσεων, ενώ η διεθνής βιβλιογραφία παραμένει σε μεγάλο βαθμό κατακερματισμένη [7]. Σκοπός της παρούσας εργασίας είναι η διερεύνηση της έννοιας των Big Data και της προοπτικής που αυτά παρουσιάζουν στον τομέα της υγείας. Επιμέρους στόχοι είναι η περιγραφή των πλεονεκτημάτων και των προκλήσεων που παρουσιάζει η αξιοποίηση τους.

Η παρούσα εργασία δομείται σε τρία κεφάλαια. Στο Πρώτο Κεφάλαιο παρουσιάζεται η έννοια την Big Data γενικά, καθώς επίσης και το πως αυτά ορίζονται στον τομέα της υγείας, από κοινού με τα κύρια χαρακτηριστικά τους. Επιπλέον, παρουσιάζονται τα μεγαλύτερα πλεονεκτήματα, αλλά και προκλήσεις που συνεπάγεται η χρήση τους. Στο Δεύτερο Κεφάλαιο αναλύεται η έννοια των Big Data Analytics, προσδιορίζονται οι κύριες κατηγορίες τους, ενώ διερευνάται η επίδραση της εφαρμογή τους στην υγειονομική περίθαλψη. Τέλος, το Τρίτο Κεφάλαιο κάνει μια αναδρομή στις κύριες εφαρμογές των Big Data στον τομέα της υγείας.

ΚΕΦΑΛΑΙΟ Ι. Τα Big Data στον τομέα της Υγείας

1.1 Ορισμός & κύρια χαρακτηριστικά των Big Data

Τα δεδομένα αποτελούν ένα πολύτιμο πόρο, ο οποίος προσλαμβάνει διάφορες μορφές. Η έννοια των Big Data ή των «μεγάλων δεδομένων» δεν αποτελεί μια νέα σύλληψη, ενώ ο τρόπος με τον οποίο αυτά ορίζονται μεταβάλλεται συνεχώς. Τα Big Data δεν έχουν ένα καθολικά αποδεκτό ορισμό, ενώ μπορεί να γίνουν αντιληπτά με διαφορετικούς τρόπους, ανάλογα με το πλαίσιο, στο οποίο εντάσσονται [9-10]. Οι διάφορες απόπειρες ορισμού των μεγάλων δεδομένων τα χαρακτηρίζουν ουσιαστικά ως ένα σύνολο δεδομένων, των οποίων το μέγεθος, η ταχύτητα, ο τύπος και/ ή η πολυπλοκότητα απαιτούν να αναζητηθεί, να υιοθετηθεί και να αναπτυχθεί νέος εξοπλισμός, καθώς και νέο λογισμικό για την επιτυχή αποθήκευση, ανάλυση και οπτικοποίηση των δεδομένων [9].

Όπως υποδηλώνει και το όνομα τους, τα Big Data αντιπροσωπεύουν μεγάλες ποσότητες δεδομένων σε διάφορους τομείς, τα οποία δεν είναι διαχειρίσιμα με τη χρησιμοποίηση των παραδοσιακών λογισμικών ή διαδικτυακών πλατφορμών, δεδομένου ότι υπερβαίνουν κατά πολύ τις υφιστάμενες δυνατότητες σε όρους αποθήκευσης, επεξεργασίας και αναλυτικής ισχύος [9-10]. Τα Big Data είναι δεδομένα των οποίων η κλίμακα, η πολυμορφία και η πολυπλοκότητα απαιτούν νέα αρχιτεκτονική, τεχνικές, αλγόριθμους και αναλυτικά στοιχεία για τη διαχείριση και την εξαγωγή αξίας και χρήσιμων συμπερασμάτων από αυτήν. [1] Συχνά τα μεγάλα δεδομένα προκαλούν ανάμικτα συναισθήματα, που κυμαίνονται από δέος έως και φόβο, αλλά στην πραγματικότητα αποτελούν μια έκρηξη στον τομέα της πληροφόρησης. Συμβάλουν σημαντικά στην πραγματοποίηση διαφόρων αναλύσεων, οι οποίες μπορούν να επηρεάσουν την οικονομική ανάπτυξη, δημιουργώντας ευκαιρίες, βελτιώνοντας την αποτελεσματικότητα του ενδιαφερόμενου οργανισμού έναντι άλλων οργανισμών [11]. Δεδομένου ότι το μέγεθος των δεδομένων αυξάνεται πάνω από ένα κρίσιμο σημείο, τα ποσοτικά ζητήματα που αφορούν τα δεδομένα πλέον μετατρέπονται σε ποιοτικά ζητήματα στη συλλογή, επεξεργασία, αποθήκευση, ανάλυση και οπτικοποίηση τους [1].

Η έννοια των Big Data αναδύθηκε στα τέλη της δεκαετίας του ‘90, όταν οι Michael Cox και David Ellsworth πρότειναν ότι η οπτικοποίηση αποτελεί πρόβλημα

των Big Data [7, 11]. Ωστόσο, ένας από τους πρώτους ορισμούς των Big Data προτάθηκε από τον Francis X. Diebold. Το 2000, ο ερευνητής αναφέρονταν στα Big Data ως *«έκρηξη της ποσότητας (και μερικές φορές, της ποιότητας) των διαθέσιμων και δυνητικά σχετιζόμενων δεδομένων».* [7] Παρά την ύπαρξη πληθώρας ορισμών για τα Big Data, ο δημοφιλέστερος και ενδεχομένως ο ευρύτερα αποδεκτός ορισμός τους, δόθηκε από τον Douglas Laney. Ειδικότερα, ο Laney παρατήρησε ότι τα (μεγάλα) δεδομένα αυξάνονταν σε τρεις διαφορετικές διαστάσεις: τον όγκο (όγκος των δεδομένων), την ταχύτητα (ταχύτητα δημιουργίας δεδομένων, ολοκλήρωση, κοινή χρήση και επεξεργασία) και την ποικιλία (ετερογένεια τύπων δεδομένων και των πηγών) [7-11].

Έτσι, ο Doug Laney ανέπτυξε και πρότεινε τον ορισμό των Μεγάλων Δεδομένων ως τρία Vs, ήτοι:

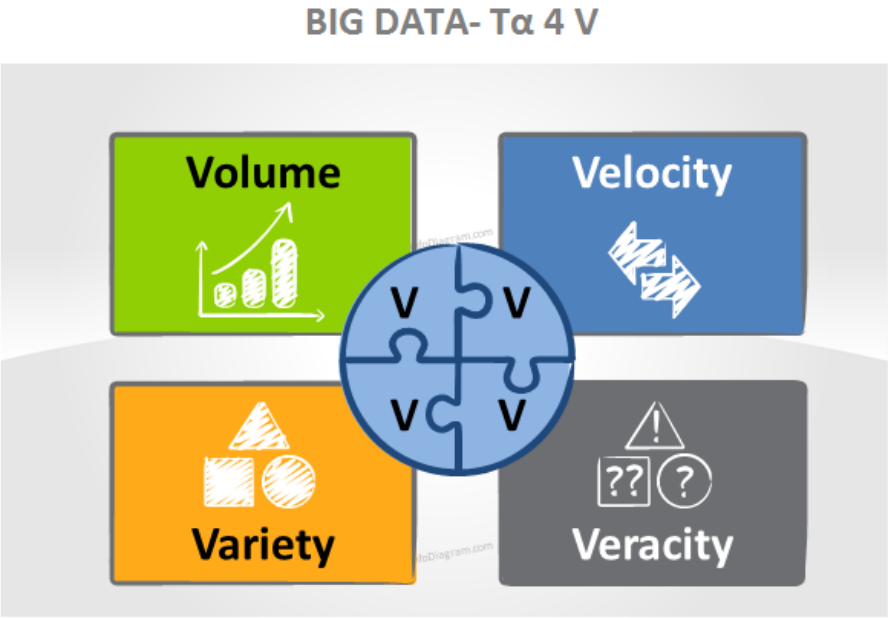
- **Όγκος (Volume):** Αναφέρεται την ποσότητα των μαζικών δεδομένων ή την ποσότητα των παραγόμενων και αποθηκευμένων δεδομένων [7]. Στον όγκο των δεδομένων συνεισφέρουν διάφοροι παράγοντες. Μπορεί να είναι δεδομένα συναλλαγών, τα οποία χρησιμοποιούνται καθ' όλη τη διάρκεια των ετών, ή η ροή των δεδομένων μέσω των μέσω κοινωνικής δικτύωσης (Social Media). Ο όγκος των δεδομένων που παράγονται σε έναν οργανισμό αυξάνεται καθημερινά με έναν απρόβλεπτο ρυθμό, ο οποίος μπορεί να είναι σε petabytes και zeta bytes και διαφοροποιείται ανάλογα με τις παραγωγικές δραστηριότητες και τον τύπο του οργανισμού [11].
- **Ταχύτητα (Velocity)**: Είναι αναμενόμενο, ότι οι αδιάκοπες ροές δεδομένων και η συσσώρευση τους με πρωτοφανείς ρυθμούς θέτει μια σειρά νέων προκλήσεων. Η ταχύτητα υποδηλώνει την ταχύτητα παραγωγής και διαχείρισης των δεδομένων. Τα μεγάλα δεδομένα είναι συνήθως διαθέσιμα σε πραγματικό χρόνο. Επί παραδείγματι, αφορούν σε δραστηριότητες όπως η τακτική παρακολούθηση των ημερήσιων μετρήσεων της γλυκόζης ενός διαβητικού ασθενούς ή της αρτηριακής πίεσης [7]. Η ταχύτητα με την οποία ένας οργανισμός λαμβάνει, παράγει, επεξεργάζεται και αναλύει τα δεδομένα που παράγει για να λάβει αποφάσεις, υπό φυσιολογικές συνθήκες συνεχίζει να επιταχύνεται, ενώ επηρεάζει τη δημιουργία και την παράδοση των δεδομένων από το ένα σημείο στο άλλο. Επιπλέον η ταχύτητα συχνά είναι ευαίσθητη στο χρόνο [11].

- **Variety (Ποικιλία)**: Τα διαθέσιμα δεδομένα πέρα από το χαοτικό όγκο τους, παρουσιάζουν και μια αυξημένη ποικιλομορφία, η οποία έρχεται να θέσει περαιτέρω προκλήσεις στην παραγωγή, επεξεργασία και διαχείριση τους. Ανάλογα με τη μορφή τους, τα δεδομένα είναι δυνατόν να κατηγοριοποιηθούν σε δομημένα (structured), ημι-δομημένα (semi-structured) και μη δομημένα (unstructured). Τα δομημένα, είναι τα δεδομένα εκείνα, για τα οποία η ανεξάρτητη αποθήκευση, ανάλυση και αξιοποίηση με τη χρήση κάποιου υπολογιστικού συστήματος, αποτελούν εύκολη υπόθεση. Τα ημι-δομημένα δεδομένα διαθέτουν ναι μεν ένα είδος δομής, που όμως παρεκκλίνει από τις επίσημες δομές των υφιστάμενων μοντέλων για τη διαχείριση δεδομένων, όπως αυτές περιγράφονται στις σχεσιακές βάσεις δεδομένων. Μολαταύτα τα ημι-δομημένα δεδομένα φέρουν κάποια ετικέτα (tag), ώστε να είναι δυνατή η ιεράρχηση τους, αλλά και η διάκριση των επιμέρους στοιχείων τους από σημασιολογικής πλευράς [12]. Τέλος, τα μη δομημένα δεδομένα δεν είναι δυνατόν να ενταχθούν σε καμία από τις δυο παραπάνω κατηγορίες, εφόσον στερούνται κάποιου προκαθορισμένου περιγραφικού μοντέλου ή στερούνται οργάνωσης βάσει μιας προκαθορισμένης δομής. Ορισμένες μορφές δομημένων δεδομένων είναι τα αριθμητικά δεδομένα, οι παραδοσιακές βάσεις δεδομένων, οι πληροφορίες από επιχειρήσεις, ενώ μορφές μη δομημένων δεδομένων είναι τα αρχεία ήχου (audio), τα video και οι εικόνες [7, 11].

Τα παραπάνω χαρακτηριστικά έχουν υιοθετηθεί ευρέως για να ορίσουν τα Big Data, ενώ σε ορισμένους άλλους ορισμούς προστέθηκε επίσης ένα τέταρτο "V" **Veracity** που αναφέρεται στην εγκυρότητα (για την αντιμετώπιση της ποιότητας των δεδομένων και κατά συνέπεια της ποιότητας των στοιχείων που μπορούν να αντληθούν από αυτά τα δεδομένα) [1,10]. Τα παραδοσιακά, επιφορτισμένα με τη διαχείριση των δεδομένων, συστήματα, αποδέχονται ότι αυτά είναι έγκυρα, ακριβή και ορθά. Ωστόσο, είναι εύλογο ότι μια τέτοια παραδοχή δεν ανταποκρίνεται απόλυτα στην πραγματικότητα, δεδομένου ότι λάθη είναι δυνατό να εντοπιστούν σε οποιαδήποτε μορφή ηλεκτρονικής υπηρεσίας [11].

Το 2012, ο Gartner επικαιροποίησε τον ορισμό των Big Data ως εξής: *«Τα μεγάλα δεδομένα είναι στοιχεία πληροφορίας με υψηλή περιεκτικότητα σε όγκο, με μεγάλη ταχύτητα και/ ή ποικιλία που απαιτούν νέες μορφές επεξεργασίας για να επιτρέψουν τη λήψη αποφάσεων, την ανακάλυψη γνώσεων και τη βελτιστοποίηση των διαδικασιών».* Επίσης, έχει παρατηρηθεί ότι ο όρος μπορεί να σημαίνει διαφορετικά

πράγματα για διαφορετικές ομάδες ανθρώπων [1]. Μια έκθεση που υποβλήθηκε στο Αμερικανικό Κογκρέσο τον Αύγουστο του 2012 ορίζει τα Big Data ως *«μεγάλους όγκους δεδομένων υψηλής ταχύτητας, σύνθετων και μεταβλητών που απαιτούν προηγμένες τεχνικές και τεχνολογίες για την καταγραφή, αποθήκευση, διανομή, διαχείριση και ανάλυση των πληροφοριών»* [4, 13].



Εικόνα 1. Τα 4 Vs των Big Data

Στην έρευνα τους η Manyika et al. [8] θεώρησαν ότι ένα ακόμα σημαντικό στοιχείο των Big Data είναι η **Αξία (Value)**. Η αξία είναι η μέθοδος εξαγωγής πολύτιμων πληροφοριών από τεράστια σύνολα δεδομένων και συνήθως αναφέρεται ως Big Data Analytics (Αναλυτική Μεγάλων Δεδομένων). Η τιμή των δεδομένων είναι χρήσιμη για τη λήψη ορθών αποφάσεων, ενώ συνήθως η ποσοτικοποίηση των αποτελεσμάτων που προκύπτουν πραγματοποιείται τόσο μέσω δεικτών, όσο και μέσω στατιστικών αναλύσεων [14].

Ένα άλλο στοιχείο που ήρθε να προστεθεί στη συνέχεια είναι **Μεταβλητότητα (Variability)** αναφέρεται στις διακυμάνσεις των δεδομένων καθ' όλη τη διάρκεια του χειρισμού και του κύκλου ζωής τους. Η ανάπτυξη του εύρους και της μεταβλητότητας αυξάνει επίσης την έλξη δεδομένων και τη δυνατότητα παροχής πολύτιμων πληροφοριών, απρόβλεπτων και κρυφών [14-15].

1.2 Ορισμός & κύρια χαρακτηριστικά Big Data στον τομέα της υγείας

Ο όγκος των δεδομένων της υγείας αναμένεται να αυξηθεί δραματικά στο μέλλον και συνεπώς καθίσταται σαφές, ότι πλέον η ανάγκη για εξεύρεση νέων και έξυπνων τρόπων διαχείρισης των δεδομένων είναι επιτακτικότερη από ποτέ. Η υγεία συγκαταλέγεται μεταξύ των τομέων εκείνων, όπου παράγονται και καταναλώνονται τεράστιοι όγκοι δεδομένων και πληροφοριών το δευτερόλεπτο και κατά συνέπεια κανείς δεν θα μπορούσε να αμφισβητήσει τη βαρύνουσα σημασία της βέλτιστης αξιοποίησης τους, προκειμένου να εξαχθούν πολύτιμα συμπεράσματα. Μια άλλη εξέλιξη αφορά στο γεγονός ότι τα μοντέλα αποζημίωσης των παρόχων υγειονομικών υπηρεσιών μεταβάλλονται. Η ουσιαστική χρήση των υπηρεσιών υγείας και η αμοιβή βάσει απόδοσης (pay for performance) εμφανίζονται ως νέοι κρίσιμοι παράγοντες στο σημερινό περιβάλλον της υγειονομικής περίθαλψης. Παρόλο που το κέρδος δεν είναι και δεν πρέπει να αποτελεί πρωταρχικό τους κίνητρο, είναι ζωτικής σημασίας για τους οργανισμούς του τομέα της υγείας να αποκτήσουν τα διαθέσιμα εργαλεία, υποδομές και τεχνικές για την αποτελεσματική μόχλευση των Big Data. Σε διαφορετική περίπτωση διακινδυνεύουν να χάσουν τεράστια ποσά από διαφυγόντα έσοδα και κέρδη [3].

Επιπρόσθετα, η ραγδαία αύξηση των δεδομένων υγείας έχει επιφέρει μια μετατόπιση της εστίασης από την θεραπεία στην πρόληψη της ασθένειας. Οι επιστήμονες έχουν επικεντρωθεί στη βελτίωση της αξιοπιστίας και της αποτελεσματικότητας των συστημάτων υγειονομικής περίθαλψης για την ελαχιστοποίηση του κόστους θεραπείας, καθώς επίσης στην παροχή καλύτερων φαρμάκων στους ασθενείς. Τα νοσοκομεία και τα εθνικά συστήματα υγείας αποτελούν μια πλούσια δεξαμενή Big Data, που περιλαμβάνουν φακέλους με το ιστορικό των ασθενών, αποτελέσματα από κλινικές δοκιμές, αλλά και αποτελέσματα από απεικονιστικές δοκιμασίες [16]. Οι υπάρχουσες αναλυτικές τεχνικές μπορούν να εφαρμοστούν στην τεράστια ποσότητα των υφιστάμενων (αλλά προς το παρόν όχι επαρκώς αναλυμένων) πάσης φύσεως ιατρικών δεδομένων, ώστε να επιτευχθεί μια βαθύτερη κατανόηση των αποτελεσμάτων, τα οποία στη συνέχεια θα μεταφερθούν και θα εφαρμοστούν στο σημείο παροχής της περίθαλψης. Ιδανικά, η γνώση αναφορικά με τον πληθυσμό, αλλά και με το ίδιο το άτομο θα υποστηρίζουν και θα παράσχουν πολύτιμες πληροφορίες στο θεράποντα ιατρό και στον ασθενή κατά τη

διάρκεια της διαδικασίας λήψης αποφάσεων και θα διευκολύνουν την επιλογή της κατ' εξοχήν κατάλληλης θεραπευτικής επιλογής για τον συγκεκριμένο ασθενή [4].

Η υγεία αποτελεί ένα έξοχο παράδειγμα του τρόπου με τον οποίο, τα τρία V από τον ορισμό των Big Data, ήτοι η ταχύτητα, η ποικιλία και ο όγκος, είναι μια εγγενής πτυχή των δεδομένων που αυτή παράγει. Τα δεδομένα αυτά διαχέονται μεταξύ πολλαπλών συστημάτων υγειονομικής περίθαλψης, ασφαλιστικών ταμείων, ερευνητών, κυβερνητικών οργανισμών κλπ. Επιπλέον, καθένα από αυτά τα αποθετήρια δεδομένων είναι υπερφορτωμένο και εγγενώς ανίκανο να παράσχει μια πλατφόρμα που χαρακτηρίζεται από διαφάνεια σε παγκόσμιο επίπεδο. Επιπλέον, μια σημαντική προσθήκη στα παραπάνω στοιχεία, είναι αυτή της έννοιας της εγκυρότητας, εφόσον η αξιοπιστία των δεδομένων της υγείας είναι επίσης σημαντική για την ουσιαστική αξιοποίηση τους στην ανάπτυξη της μεταγραφικής έρευνας [4-7]. Ειδικότερα στον τομέα της υγείας, ο ορισμός των Big Data ενσωματώνει τα ακόλουθα στοιχεία:

- **Volume (Όγκος):** Με την πάροδο του χρόνου, τα δεδομένα που σχετίζονται με την υγεία θα συνεχίσουν να παράγονται και να συσσωρεύονται συνεχώς, με αποτέλεσμα τη συσσώρευση ενός απίστευτου όγκου δεδομένων. Ο ήδη αποθαρρυντικός όγκος των δεδομένων των υφιστάμενων υγειονομικών συστημάτων περιλαμβάνει τα προσωπικά ιατρικά αρχεία, τις ακτινολογικές απεικονίσεις, τα αποτελέσματα διαφόρων κλινικών δοκιμών, τις γονιδιωματικές (genomics) ακολουθίες ανθρώπινων γενετικών και πληθυσμιακών δεδομένων κλπ. Νεότερες μορφές δεδομένων μεγάλης κλίμακας, όπως τρισδιάστατες απεικονίσεις, έρευνες του γονιδιώματος και βιομετρικές μετρήσεις από αισθητήρες τροφοδοτούν επίσης αυτή την εκθετική ανάπτυξη των δεδομένων [3].

Σήμερα τα δεδομένα υγείας μετρούνται σε terabytes (10^{12} bytes), petabytes (10^{15} bytes) ή Exabyte’s (1 exabyte =10^{18} = 1 δισ. gigabytes) [13, 17]. Στο μέλλον, το τεράστιο σύνολο των αρχείων κλινικών δεδομένων αναμένεται να αυξηθεί σε zettabytes (10^{21} bytes) ή yottabytes (1024 byte). Τέτοιες τεράστιες ποσότητες δεδομένων δημιουργούν ζητήματα ως προς τον χώρο αποθήκευσης και τη μαζική ανάλυση τους. Ο όγκος αναφέρεται στην ποσότητα των Big Data στον τομέα της υγείας, η οποία εκτιμάται ότι θα αυξηθεί δραματικά σε 35 zettabytes μέχρι το 2020 [7].

Τα μεγάλα δεδομένα στην υγειονομική περίθαλψη είναι σημαντικό ζήτημα, όχι μόνο λόγω του όγκου, αλλά και λόγω της ποικιλότητας των τύπων των δεδομένων και της ταχύτητας με την οποία θα πρέπει να διαχειριστούν αυτά τα δεδομένα. Η αυξημένη χρήση της τηλεϊατρικής θα δοκιμάσει περαιτέρω τη χωρητικότητα αποθήκευσης των δεδομένων των ασθενών, ενώ η καινοτόμος χρήση του Google Glass από τους ιατρούς θα έρθει επίσης να προσθέσει στις κοινωνικές και συμπεριφορικές πτυχές των Big Data. Τα ηλεκτρονικά ιατρικά αρχεία (EMR) περιέχουν μια πληθώρα δεδομένων, όπως τα δημογραφικά στοιχεία των ασθενών και κλινικά και γονιδιωματικά δεδομένα, ενώ ως γνωστόν υποβοηθούν σημαντικά την καλή ροή της υγειονομικής περίθαλψης [1].

Οι μελλοντικές εφαρμογές δεδομένων σε πραγματικό χρόνο, όπως η έγκαιρη ανίχνευση των λοιμώξεων, η έγκαιρη ταυτοποίησή τους και η χορήγηση των ενδεδειγμένων θεραπειών θα μπορούσαν να μειώσουν τη νοσηρότητα και τη θνησιμότητα των ασθενών και μάλιστα να αποτρέψουν την εμφάνιση ενδονοσοκομειακών λοιμώξεων. Ήδη, τα δεδομένα συνεχούς ροής σε πραγματικό χρόνο χρησιμοποιούνται για την παρακολούθηση νεογέννητων στη ΜΕΘ, προειδοποιώντας τους θεράποντες ιατρούς για πιθανές μολύνσεις που απειλούν τη ζωή του νοσηλευόμενου νεογνού. Η δυνατότητα πραγματοποίησης αναλύσεων σε δεδομένα τεράστιου όγκου, σε όλες τις ειδικότητες και σε πραγματικό χρόνο, αναμφίβολα θα φέρει επανάσταση στην υγεία [18].

- **Ποικιλία (Variety):** Η ποικιλία αναφέρεται στους διάφορους τύπους Big Data που συλλέγονται στην υγεία, συμπεριλαμβανομένων των ετερογενών χαρακτηριστικών τους [19]. Η ετερογένεια προκύπτει την άντληση δεδομένων από διαφορετικές και αντικρουόμενες βάσεις δεδομένων ή συνδυασμούς δεδομένων που προέρχονται από αυτόνομες πηγές [13]. Ιστορικά, το σημείο όπου παρέχεται η υγειονομική φροντίδα παράγει μη δομημένα δεδομένα: ιατρικούς φακέλους, χειρόγραφες σημειώσεις νοσηλευτών και γιατρών, αρχεία εισαγωγών και νοσοκομειακά εξιτήρια, συνταγές σε χαρτί, ακτινογραφίες, μαγνητικές τομογραφίες, αξονικές τομογραφίες και αποτελέσματα από άλλες απεικονιστικές δοκιμασίες [3].

Τα δεδομένα στην υγειονομική περίθαλψη μπορεί να ταξινομηθούν ως δομημένα, ημι-δομημένα ή μη δομημένα [3, 7, 13]. Τα δομημένα δεδομένα είναι δεδομένα που μπορούν εύκολα να αποθηκευτούν, να αναζητηθούν, να ανακληθούν,

να αναλυθούν και να υποβληθούν σε χειρισμούς από κάποιο υπολογιστικό πρόγραμμα [3]. Όπως προαναφέρθηκε, τα δομημένα δεδομένα περιλαμβάνουν τα εργαστηριακά δεδομένα, τα κλινικά δεδομένα αισθητήρων και τα δεδομένα από τις σχεσιακές βάσεις δεδομένων [20], τα ημι-οργανωμένα δεδομένα περιλαμβάνουν δεδομένα που είναι αποθηκευμένα σε μορφή διεπαφής XML[3] (Extensible Markup Language). Τα μη δομημένα δεδομένα είναι δεδομένα ελεύθερου κειμένου που συνήθως δεν διαθέτουν ένα συγκεκριμένο σχεδιασμό και συνήθως περιλαμβάνουν ηλεκτρονικά ιατρικά αρχεία [13], χειρόγραφες σημειώσεις, γραφήματα, ακτινογραφίες, περιλήψεις των εξιτηρίων των ασθενών, μετρήσεις φυσιολογικών σημείων, δεδομένα της υγεία που προέρχονται από μέσα κοινωνικής δικτύωσης και τα έξυπνα τηλέφωνα [21]. Το 90% των Big Data της υγείας έχει τη μορφή μη δομημένων δεδομένων [13].

Ήδη, νέες ροές δεδομένων - δομημένες και μη δομημένες - διαπερνούν το χώρο της υγείας, προερχόμενες από μηχανήματα γυμναστικής, γενετικές και γονιδιωματικές μελέτες, διαδικτυακές έρευνες και άλλες πηγές. Ωστόσο, τα δεδομένα, τα οποία μπορούν να διαβαστούν, να αποθηκευτούν και να οργανωθούν, έτσι ώστε να μπορούν να υποβληθούν σε επεξεργασία από υπολογιστικά συστήματα και να αναλυθούν για χρήσιμες πληροφορίες, είναι περιορισμένα. Ειδικότερα, οι εφαρμογές που σχετίζονται με την υγεία θα πρέπει να διαθέτουν πιο αποτελεσματικούς τρόπους συνδυασμού και μετατροπής των ετερογενών δεδομένων, συμπεριλαμβανομένης της τροποποίησης από δομημένα σε μη δομημένα δεδομένα [3].

- **Velocity (Ταχύτητα):** Η ταχύτητα αφορά την ταχύτητα παραγωγής των δεδομένων (δηλ. δεδομένα ασθενών σε πραγματικό χρόνο, καθώς και συλλογή των δεδομένων) [7, 13]. Ως εκ τούτου, η ταχύτητα περιλαμβάνει τόσο την ταχύτητα της παραγωγής δεδομένων, όσο και την ταχύτητα του χειρισμού τους για την κάλυψη της ζήτησης. Η ραγδαία αύξηση των δεδομένων αναφέρεται ως η τρίτη ιδιότητα των Big Data [22]. Τα δεδομένα που δημιουργούνται μπορούν να είναι είτε δεδομένα ανά δεσμίδες (batch), είτε πραγματικού χρόνου (real time). Ένα παράδειγμα της ταχύτητας των δεδομένων είναι η γήρανση του πληθυσμού, η οποία οδηγεί συνεχώς σε μια

[3] Πρόκειται για γλώσσα σήμανσης

αύξηση του αριθμού ασθενών, γεγονός που με τη σειρά του αυξάνει το ρυθμό αύξησης των δεδομένων κατά 55–60% ετησίως [13].

- **Εγκυρότητα (Veracity):** Η εγκυρότητα αναφέρεται στους παράγοντες, οι οποίοι επηρεάζουν την ακρίβεια και αξιοπιστία των δεδομένων, όπως οι ανακρίβειες, οι παραλείψεις, οι αμφισημίες, η εξαπάτηση, η απάτη, η αλληλοεπικάλυψη, αλλά και οι λανθασμένες πληροφορίες [7, 23]. Τα Big Data χαρακτηρίζονται από χαμηλή εγκυρότητα, δεν μπορεί να είναι ποτέ 100% ακριβή, ενώ η εγκυρότητα τους είναι δύσκολο να διασταυρωθεί. Δεδομένου ότι, η πλειοψηφία των δεδομένων προέρχεται από άγνωστες και μη επιβεβαιωμένες πηγές, είναι απαραίτητο να τεθεί ένα πρότυπο, ώστε να διασφαλιστεί η εγκυρότητα των εμπλεκόμενων δεδομένων [13]. Η εγκυρότητα και η ποιότητα των δεδομένων είναι ύψιστης σημασίας στην υγεία, δεδομένου ότι η λήψη αποφάσεων ζωής και θανάτου θα πρέπει να βασίζονται σε αξιόπιστες και ακριβείς πληροφορίες [3, 7].
- **Αξία (Value):** Αντιπροσωπεύει τον υπολογισμό του κόστους από τον υπεύθυνο λήψης αποφάσεων, εφόσον του παρέχει τη δυνατότητα να αναλάβει ουσιαστική δράση βασίζοντας τις αποφάσεις του σε πληροφορίες, προερχόμενες από δεδομένα [13]. Προκειμένου να εξαχθεί αξία από τα μεγάλα δεδομένα στον τομέα της υγείας, είναι σκόπιμο να υποστηριχθούν αποδοτικές πλατφόρμες επεξεργασίας δεδομένων, πιο έξυπνες τεχνολογίες συλλογής δεδομένων, υπολογιστικές αναλύσεις, τεχνικές αποθήκευσης και οπτικοποίησης, ώστε να προκύψουν καινοτόμες γνώσεις και αποτελεσματικές στρατηγικές υποστήριξης λήψης αποφάσεων για διάφορα αναδυόμενα ζητήματα του τομέα [24].

Τα χαρακτηριστικά των ιατρικών δεδομένων που συμβάλλουν στην πολυπλοκότητά τους περιλαμβάνουν την ποικιλομορφία των διαφόρων ασθενειών και τις συννοσηρότητές τους, την ετερογένεια των θεραπειών και των αποτελεσμάτων, τις διαφορές στις κλινικές ροές εργασίας, τα διαφορετικά πρότυπα ιατρικής και νοσηλευτικής πρακτικής, τους πληθυσμούς των ασθενών, τις διαθέσιμες τεχνολογίες και τους μηχανισμούς παραπομπής [7]. Ωστόσο, η αναλυτική των μεγάλων δεδομένων έρχεται να υπερκεράσει αυτά τα ζητήματα, εφόσον χρησιμοποιείται για την πραγματοποίηση αναλύσεων και τη λήψη αποφάσεων σε μεγάλη κλίμακα και σε χαμηλό κόστος. Από την άλλη πλευρά, το χάσμα μεταξύ του κόστους της περίθαλψης

και των αποτελεσμάτων της υγείας είναι ένα από τα σημαντικότερα ζητήματα και πολλές ανεπτυγμένες χώρες καταβάλλουν σημαντικές προσπάθειες για την κάλυψη του. Βάσει εκτιμήσεων το κενό αυτό αποτελεί την κοινή συνιστώσα της κακής διαχείρισης των αποτελεσμάτων της έρευνας, της περιορισμένης αξιοποίησης των διαθέσιμων στοιχείων, της περιορισμένης εμπειρίας, που από κοινού οδήγησαν σε χαμένες ευκαιρίες, σπατάλη πόρων και πιθανή βλάβη στους ασθενείς [1].

1.3 Πηγές Big Data στον τομέα της υγείας

Τα δεδομένα στην υγειονομική περίθαλψη είναι αποδιοργανωμένα και κατακερματισμένα, προέρχονται από διαφορετικές πηγές και έχουν διαφορετικές δομές και μορφές [1, 7]. Τα κλινικά δεδομένα, όπως τα ζωτικά σημεία, το ιατρικό ιστορικό του ασθενούς, τα φάρμακα, οι ανοσοποιήσεις και οι απεικονιστικές δοκιμασίες, μπορούν να αντληθούν από ηλεκτρονικά αρχεία υγείας, συστήματα εισαγωγής ιατρικών εντολών (CPOE[4]), συστήματα υποστήριξης κλινικών αποφάσεων (CDSS[5]), αρχεία χορήγησης φαρμακευτικής θεραπείας, εργαστηριακά και φαρμακευτικά αρχεία [3, 25] μελέτες κοόρτης, κυβερνητικές έρευνες και κλινικές δοκιμές. Από την άλλη πλευρά, τα διοικητικά δεδομένα, περιέχουν δημογραφικά δεδομένα των ασθενών και πληροφορίες από τις επισκέψεις που αυτοί πραγματοποιούν στο γιατρό τους, την ημερομηνία εισαγωγής του ασθενούς, την ημερομηνία που αυτός έλαβε εξιτήριο, τη διάγνωση κατά την Διεθνή Στατιστική Ταξινόμηση Νόσων και Συναφών Προβλημάτων Υγεία (ICD[6]), την κατάσταση της υγείας του ασθενούς κατά το εξιτήριο καθώς επίσης δεδομένα οικονομικών απαιτήσεων που συμπεριλαμβάνουν μεταξύ άλλων τις χρεώσεις επίσκεψης, το ποιος θα πρέπει να καταβάλει το αντίτιμο της υπηρεσίας υγείας που έλαβε ο ασθενής, καθώς και την αποζημίωση του παρόχου υγειονομικής περίθαλψης [7, 25].

Οι πηγές των Big Data στην υγειονομική περίθαλψη μπορούν να ταξινομηθούν ευρέως ως εσωτερικές (π.χ. ηλεκτρονικά αρχεία υγείας, συστήματα υποστήριξης κλινικών αποφάσεων, CPOE κ.λπ.) και εξωτερικές (κυβερνητικές πηγές, εργαστήρια, φαρμακεία, ασφαλιστικές εταιρείες κλπ.). Τα δεδομένα αυτά διακρατούνται σε πολλαπλές μορφές (επίπεδα αρχεία, σχεσιακοί πίνακες, ASCII/ κείμενα κ.λπ.),

[4] Computerized Physician Order Entry
[5] Clinical Decision Support System
[6] International Classification of Diseases

φυλάσσονται σε πολλαπλές τοποθεσίες (τόσο με την γεωγραφική έννοια της τοποθεσίας, όσο και με την έννοια ότι φιλοξενούνται στις εγκαταστάσεις διαφορετικών παρόχων υγειονομικής περίθαλψης), ενώ τηρούνται και σε διαφορετικές εφαρμογές (εφαρμογές επεξεργασίας συναλλαγών, βάσεις δεδομένων κλπ.) [1]. Από τα παραπάνω προκύπτει ότι τα Big Data στην υγεία περιλαμβάνουν δεδομένα σχετικά με τη φυσιολογία, τη συμπεριφορά, τη μοριακή, την κλινική πρακτική, την περιβαλλοντική έκθεση, την ιατρική απεικόνιση, τη διαχείριση των ασθενειών, το ιστορικό συνταγογραφούμενων φαρμάκων, τη διατροφή ή ακόμα και παραμέτρους άσκησης [26-27].

Στη διεθνή βιβλιογραφία δεν υπάρχει μια ευρύτερα αποδεκτή ταξινόμηση των πηγών των Big Data στην υγειονομική περίθαλψη. Μια άποψη ταυτοποιεί δύο κύριες πηγές Big Data, ήτοι τη γονιδιωματική έρευνα (δεδομένα που προκύπτουν από την έρευνα αναφορικά με το γονότυπο, τη γονιδιακή έκφραση και τα δεδομένα αλληλουχίας) και τον αγοραστή – πάροχο υπηρεσιών υγείας (δεδομένα που προέρχονται από ηλεκτρονικά αρχεία υγείας, ασφαλιστικά αρχεία, συνταγές φαρμάκων, ανατροφοδότηση ασθενών κτλ.) [7]. Από την άλλη πλευρά, σε μελέτη του ο Swan [28] ταξινόμησε τις μεγάλες ροές των Big Data στην υγεία σε:

1. Παραδοσιακά δεδομένα υγείας, που προέρχονται από διοικητικές βάσεις δεδομένων (ασφαλιστικές απαιτήσεις και φαρμακευτικά προϊόντα), κλινικές βάσεις δεδομένων, τα δεδομένα των ηλεκτρονικών φακέλων υγείας και τα δεδομένα των πληροφοριακών συστημάτων των κατά τόπους εργαστηρίων. Τα δεδομένα αυτά συνεισφέρουν σημαντικά στην καλύτερη κατανόηση των αποτελεσμάτων της νόσου και στη βελτιστοποίηση της παροχής υγειονομικής περίθαλψης.
2. Δεδομένα που προέρχονται από ολόκληρο το φάσμα των βιολογικών δεδομένων[7] (συμπεριλαμβανομένων των δεδομένων γονιδιωματικής[8], πρωτεομικής[9], και μεταβολομικής[10]). Τα δεδομένα αυτά είναι ύψιστης σημασίας για την κατανόηση των μηχανισμών των ασθενειών και την επιτάχυνση της εξατομίκευσης των παρεχόμενων θεραπειών.
3. Άλλες πηγές δεδομένων είναι τα λεγόμενα βιομετρικά δεδομένα (δεδομένα από συσκευές που έχουν τοποθετηθεί στον ασθενή ή δεδομένα που

[7] Τα λεγόμενα –omics data λόγω της κοινής κατάληξης των ονομασιών τους.
[8] Genomic
[9] Proteomic
[10] Metabolomics

καταγράφονται από αισθητήρες) ή δεδομένα από μέσα κοινωνικής δικτύωσης. Αυτά τα δεδομένα παρέχουν πολύτιμες πληροφορίες για τον τρόπο ζωής και τη συμπεριφορά του ατόμου.

Σύμφωνα με τους Belle et al. [9], το δεδομένα στον τομέα της υγείας είναι διάσπαρτα μεταξύ διαφορετικών υγειονομικών συστημάτων, ασφαλιστικές εταιρείες, ερευνητές και κυβερνητικές οντότητες. Στη μελέτη τους οι Huang et al. [25] ότι τα Big Data στην ιατρική ακριβείας (precision medicine) προέρχονται από 4 κύριους εμπλεκόμενους φορείς (stakeholders):

i. Κυβέρνηση και μεγάλες επιχειρήσεις,
ii. Μικρότερους φορείς όπως ακαδημαϊκές ομάδες και νεοφυείς επιχειρήσεις (start-ups) στους τομείς της τεχνολογίας, βιοτεχνολογίας και των ιατροτεχνολογικών συσκευών,
iii. Αγοραστές και παρόχους υπηρεσιών υγείας, και
iv. Μη κερδοσκοπικούς οργανισμούς, ομάδες υπεράσπισης ασθενών [7].

1.4 Προοπτικές Big Data στον τομέα της υγείας

1.4.1 Πλεονεκτήματα που παρουσιάζουν τα Big Data

Μεταξύ των κύριων τομέων της υγειονομικής περίθαλψης, στους οποίους έχουν αποδειχθεί τα οφέλη των Big Data συγκαταλέγονται η πρόληψη της ασθένειας, ο εντοπισμός των τροποποιήσιμων παραγόντων κινδύνου για τη νόσο, καθώς επίσης ο σχεδιασμός παρεμβάσεων για την αλλαγή συμπεριφοράς του ατόμου [1]. Όπως ισχυρίζονται οι Rumsfeld et al. [27] η ανάλυση των Big Data στον τομέα της υγείας μπορεί να συμβάλει σημαντικά στους ακόλουθους τομείς:

1. Ανάπτυξη των προγνωστικών μοντέλων αναφορικά με τους κινδύνους και τη χρήση των πόρων,
2. Διαχείριση πληθυσμού,
3. Επιτήρηση της ασφάλειας των φαρμάκων και του τεχνολογικού εξοπλισμού,
4. Ετερογένεια των ασθενειών και των θεραπειών,
5. Ιατρική ακριβείας και συστήματα υποστήριξης κλινικών αποφάσεων,
6. Ποιότητα φροντίδας και μέτρηση της απόδοσης (performance measurement),
7. Δημόσια υγεία, και

8. Εφαρμογές έρευνας.

Η ψηφιοποίηση, η συσχέτιση και η αποτελεσματική αξιοποίηση των Big Data συνεπάγεται μια σειρά από σημαντικά πλεονεκτήματα για όλο το φάσμα των οργανισμών υγείας, που ποικίλλουν από ιδιωτικά ιατρεία έως μεγάλα νοσοκομειακά δίκτυα [29]. Τα πιθανά οφέλη περιλαμβάνουν την ανίχνευση ασθενειών σε προγενέστερα στάδια, όταν μπορούν να αντιμετωπιστούν κατά τρόπο απλούστερο και αποτελεσματικότερο, τη διαχείριση της υγείας τόσο του ατόμου μεμονωμένα, όσο και του πληθυσμό, καθώς και την έγκαιρη και αποτελεσματική ανίχνευση της απάτης στην υγειονομική περίθαλψη. Πολλά ερωτήματα ενδέχεται να αντιμετωπιστούν και με την ανάλυση αυτών των δεδομένων, που αποτελεί ένα διακριτό επιστημονικό πεδίο: τα Big Data Analytics (Αναλυτική Μεγάλων Δεδομένων). Ορισμένες εξελίξεις ή αποτελέσματα είναι δυνατόν να προβλεφθούν ή και να υπολογιστούν στη βάση τεραστίου όγκου ιστορικών δεδομένων, όπως η διάρκεια νοσηλείας (LOS[11]), οι ασθενείς που θα επιλέξουν να υποβληθούν σε κάποια εκλεκτική χειρουργική επέμβαση, οι ασθενείς που πιθανότατα δεν θα ήθελαν να υποβληθούν σε χειρουργική επέμβαση, οι επιπλοκές, οι ασθενείς που διατρέχουν κίνδυνο ανάπτυξης ιατρικών επιπλοκών, οι ασθενείς με κίνδυνο να παρουσιάσουν σήψη, οι ενδονοσοκομειακές ασθένειες, η εξέλιξη της νόσου και οι αιτιολογικοί παράγοντες της, καθώς επίσης και τα συνοδά νοσήματα [3-4, 8].

Βάσει εκτιμήσεων του Διεθνούς Ινστιτούτου McKinsey η αξιοποίηση της ανάλυσης των Big Data Analytics μπορεί να επιτύχει μια εξοικονόμηση της τάξεως των $300 δισ. ετησίως στο Αμερικανικό σύστημα υγείας, τα 2/3 εκ των οποίων αφορούν μειώσεις στις δαπάνες για την υγεία. Οι κλινικές εφαρμογές και ο τομέας της Έρευνας & Ανάπτυξης (E&A ή R & D) αποτελούν δυο από τους κύριους τομείς, όπου δυνητικά μπορεί να εξοικονομηθούν πόροι, εφόσον σημειώνουν απώλειες λόγω κατασπατάλησης πόρων, που αγγίζουν τα 165 δισ. δολάρια και 108 δισ. δολάρια, αντίστοιχα [8, 30]. Περαιτέρω, το εν λόγω Ινστιτούτο υποστηρίζει ότι τα Big Data μπορούν να συμβάλλουν σημαντικά στην εξάλειψη της κατασπατάλησης των πόρων και της αναποτελεσματικότητας στους ακόλουθους τομείς:

[11] Length of Stay

1. Κλινικές Εφαρμογές

Τα Big Data καθιστούν εφικτή την πραγματοποίηση συγκριτικής ανάλυσης της αποτελεσματικότητας των διαφόρων διαδικασιών και κλινικών πρακτικών, προκειμένου να καταδειχθούν οι καταλληλότεροι από κλινικής απόψεως, καθώς επίσης οι αποδοτικοί από οικονομικής απόψεως τρόποι διάγνωσης και θεραπείας των ασθενών.

2. Έρευνα & Ανάπτυξη

Τα Big Data μπορούν να συμβάλλουν σημαντικά στον τομέα της Ε&Α με τους ακόλουθους τρόπους:

- Καθιστούν εφικτή την ανάπτυξη προγνωστικών μοντέλων για τη διευκόλυνση των φαρμακευτικών και των λοιπών εταιρειών που δραστηριοποιούνται στο χώρο της υγείας για την ταχύτερη ανάπτυξη, παραγωγή και κυκλοφορία στην αγορά νέων φαρμάκων και ιατροτεχνολογικών προϊόντων.
- Αποτελούν τη βάση της ανάπτυξης στατιστικών εργαλείων και αλγόριθμων για την ενίσχυση των κλινικών δοκιμών και της πρόσληψης ασθενών για την καλύτερη αντιστοίχιση των θεραπειών σε μεμονωμένους ασθενείς, μειώνοντας τρόπον τινά τις αποτυχίες των κλινικών δοκιμών και επιταχύνοντας την κυκλοφορία των νέων θεραπειών στην αγορά.
- Συμβάλλουν στην ανάλυση των κλινικών δοκιμών και των αρχείων των ασθενών για τον έγκαιρο εντοπισμό των πιθανών παρενεργειών και των δυσμενών επιπτώσεων πριν το φαρμακευτικό προϊόν φθάσει στην αγορά [8, 30].

3. Δημόσια Υγεία

Τα Big Data μπορούν να σηματοδοτήσουν μια σειρά σημαντικών εξελίξεων στον τομέα της Δημόσιας Υγείας με τους εξής τρόπους:

- Ευνοούν την ταχύτερη ανάπτυξη εμβολίων, τα οποία έχουν αναπτυχθεί με ακριβή στόχευση π.χ. την επιλογή των ετήσιων στελεχών της γρίπης.
- Μετατρέπουν τεράστιες ποσότητες δεδομένων σε πληροφορίες που μπορούν να χρησιμοποιηθούν για τον προσδιορισμό των αναγκών, την παροχή

υπηρεσιών και την πρόβλεψη και πρόληψη κρίσεων, ιδιαίτερα προς όφελος των πληθυσμών αναφοράς [8, 30].

4. **Τεκμηριωμένη ιατρική (Evidence-based medicine)**

Τα Big Data επιτρέπουν το συνδυασμό και ανάλυση μιας ποικιλίας δομημένων και μη δομημένων δεδομένων, των οικονομικών και επιχειρησιακών δεδομένων, των κλινικών δεδομένων και των δεδομένων γονιδιωματικής για το συνταίριασμα των θεραπειών με τα θεραπευτικά αποτελέσματα, την πρόβλεψη των ασθενών που αντιμετωπίζουν κίνδυνο για ασθένεια ή επανεισαγωγή και την προσφορά της βέλτιστης υγειονομικής φροντίδας [3-4].

5. **Ανάλυση Βιολογικών Δεδομένων**

Η εξέταση της αλληλουχίας των γονιδίων με ένα πιο αποτελεσματικό και οικονομικά αποδοτικό τρόπο, καθώς και η ανάλυση του γονιδιώματος αποτελεί αναπόσπαστο τμήμα της συνήθους διαδικασίας λήψης ιατρικών αποφάσεων, καθώς και ένα αυξανόμενης σημασίας τμήμα του ηλεκτρονικού φακέλου υγείας του ασθενούς [25].

6. **Ανάλυση Απάτης**

Η απάτη πλέον αντιπροσωπεύει ένα ζήτημα που έχει προσλάβει σημαντικές διαστάσεις στους τομείς της υγειονομικής περίθαλψης και της ασφάλισης, εφόσον υπάρχουν ασθενείς που προβαίνουν σε ψευδείς αξιώσεις ελπίζοντας να λάβουν κάποια αποζημίωση από την ασφαλιστική τους εταιρεία. Τα Big Data παρουσιάζουν αυξημένη χρησιμότητα για την επίλυση του προβλήματος αυτού, δεδομένου ότι μπορούν να χρησιμοποιήσουν ένα μεγάλο αριθμό πληροφοριών για να βρουν διαφορές στις γραπτές αξιώσεις και τις απαιτήσεις, να εντοπίσουν εκείνες που ενέχουν δόλο και στη συνέχεια να προχωρήσουν στην περαιτέρω διερεύνηση τους. Μέσω των αναρίθμητων δυνατοτήτων που προσφέρουν, τα μεγάλα δεδομένα μπορούν να συγκρίνουν ένα τεράστιο όγκο εγγραφών για να εντοπίσουν τα λάθη ή αναντιστοιχίες πιο γρήγορα από ένα ανθρώπινο μάτι [31]. Κάτι τέτοιο θα διευκολύνει σημαντικά τις ασφαλιστικές εταιρείες να αποτρέψουν και να περιορίσουν τις απώλειες τους [32].

7. **Παρακολούθηση συσκευών, ακόμα και από απόσταση**

Πλέον είναι δυνατή η καταγραφή και η ανάλυση σε πραγματικό χρόνο τεράστιων όγκων δεδομένων, τα οποία παράγονται και διακινούνται με τεράστιες ταχύτητες και

προέρχονται από συσκευές είτε εγκατεστημένες στο νοσοκομείο, είτε στο σπίτι του ασθενούς. Σκοπός αυτής της παρακολούθησης είναι η επιτήρηση της ασφάλειας των ιατροτεχνολογικών αυτών συσκευών, καθώς επίσης η πρόβλεψη κάποιου δυσμενούς γεγονότος [3-4]. Πράγματι, τα Big Data πραγματικού χρόνου παρουσιάζουν σημαντικά πλεονεκτήματα. Επί παραδείγματι, τυχόν λάθη ή σφάλματα σε έναν οργανισμό μπορούν να αναγνωριστούν άμεσα και το επιχειρησιακό πρόβλημα μπορεί να ξεπεραστεί. Αυτό θα εξοικονομήσει χρόνο, κόστος και θα αυξήσει την παραγωγικότητα του οργανισμού. Οι υπηρεσίες μπορούν επίσης να βελτιωθούν, καθώς τα δεδομένα πραγματικού χρόνου παρέχουν τις πλέον πρόσφατες πληροφορίες σχετικά με το θέμα. Για παράδειγμα, τα Big Data πραγματικού χρόνου θα είναι σε θέση να παρέχουν τις πλήρεις πληροφορίες για τους ασθενείς και ταυτόχρονα να θέτουν σε εφαρμογή την ενδεδειγμένη ιατρική παρέμβαση χωρίς καμία καθυστέρηση [32-33].

8. Ανάλυση προφίλ ασθενών

Η εφαρμογή προηγμένων αναλυτικών τεχνικών στα προφίλ των ασθενών (π.χ. ταξινόμηση και ανάπτυξη προγνωστικών μοντέλων) επιτρέπει τον προσδιορισμό των ατόμων που θα επωφεληθούν από μια παρέμβαση προληπτικής υφής ή τις αλλαγές στον τρόπο ζωής, παραδείγματος χάριν είναι δυνατό να προσδιοριστούν εκείνοι οι ασθενείς που κινδυνεύουν να αναπτύξουν μια συγκεκριμένη ασθένεια (π.χ. διαβήτης) και οι οποίοι θα ωφεληθούν από την παροχή προληπτικής φροντίδας υγείας [3-4]. Μέσω της διαχείρισης δεδομένων, της τήρησης ηλεκτρονικών ιατρικών αρχείων και της ανάλυσης δεδομένων τα Big Data μπορούν επίσης να βοηθήσουν στην εύρεση και ταυτοποίηση του σωστού πληθυσμού ή της ομάδας - στόχου. Δεδομένου ότι αυτά συμπεριλαμβάνουν ποικίλες πληθυσμιακές ομάδες, παρέχεται η δυνατότητα για τον προσδιορισμό μιας συγκεκριμένης ομάδας, για την οποία θα πρέπει να πραγματοποιηθεί εκτίμηση κινδύνου και προσυμπτωματικός έλεγχος (screening). Επιπλέον, η ύπαρξη των Big Data επιτρέπει επίσης την ανάπτυξη ή την τροποποίηση ενός προγράμματος ή μιας παρέμβασης για την αντιμετώπιση του προβλήματος υγείας [34] από κοινού με την άμεση έναρξη των κλινικών δοκιμών. Τα μεγάλα δεδομένα θα δώσουν μια σαφέστερη εικόνα για τον τύπο του πληθυσμού καθώς και για το ιατρικό πρόβλημα που αυτοί αντιμετωπίζουν. Το μοτίβο της κατανομής της νόσου ή σχετικές με αυτήν πληροφορίες θα επιτρέψουν την ταχεία ανάπτυξη ενός προγράμματος παρέμβασης, καθώς και τη στόχευση της πληγείσας πληθυσμιακής ομάδας το συντομότερο δυνατό [32].

Τέλος, τα μεγάλα δεδομένα θα μπορούσαν να μειώσουν τη λεγόμενη νεωτεριστική προκατάληψη (recency bias). Η νεωτεριστική προκατάληψη υφίσταται όταν τα πρόσφατα γεγονότα θεωρούνται σημαντικότερα και εγκυρότερα σε σχέση με τα προηγούμενα γεγονότα. Ωστόσο, θα πρέπει να σημειωθεί ότι ιδιαίτερα στον τομέα της υγείας, η νεωτεριστική προκατάληψη μπορεί να οδηγήσει σε λανθασμένες αποφάσεις. Επιπλέον, στην υγειονομική περίθαλψη, τα Big Data χρησιμοποιούνται επίσης στην προγνωστική ανάλυση, η οποία ως κύριους στόχους θέτει τον εντοπισμό και αντιμετώπιση ενός ιατρικού ζητήματος, πριν αυτό καταστεί ένα ανεξέλεγκτο. Έτσι, με τη χρήση των προερχόμενων, από τα επεξεργασμένα Big Data, πληροφοριών οι επαγγελματίες υγείας μπορούν να μειώσουν τον κίνδυνο και να ξεπεράσουν το πρόβλημα εν τη γενέσει του [32].

Από την παραπάνω ανάλυση προκύπτει ότι, για τη βελτίωση της ποιότητας της υγειονομικής περίθαλψης και των αποτελεσμάτων των ασθενών, είναι επιτακτική η ανάγκη για αύξηση τόσο της διαθεσιμότητας των δεδομένων, όσο και των δυνατοτήτων ανάλυσης τους, παράγοντες που από κοινού θεωρούνται ως οι κεντρομόλες δυνάμεις για την έλευση των μεγάλων δεδομένων στον τομέα της υγείας [1].

1.4.2 Προκλήσεις που αντιμετωπίζουν τα Big Data

Ακόμη και με τα τεράστια δυνητικά οφέλη τους, ο κλάδος της υγειονομικής περίθαλψης βρίσκεται σε πρώιμη φάση όσον αφορά υιοθέτηση των Big Data [7]. Η υγειονομική περίθαλψη, η βιοϊατρική έρευνα και η υγεία του πληθυσμού παράγουν μαζικά, σύνθετα, διασκορπισμένα και συχνά δυναμικά σύνολα δεδομένων, το μέγεθος και η πολυπλοκότητα των οποίων συνεπάγονται σημαντικές προκλήσεις για τους οργανισμούς που δραστηριοποιούνται στον τομέα της υγείας. Η εγκυρότητά και αξιοπιστία των Big Data στην υγεία μπορούν να επηρεαστούν από διάφορους παράγοντες, όπως σφάλματα μέτρησης, ελλιπή δεδομένα ή σφάλματα κατά την κωδικοποίηση των πληροφοριών που εμπεριέχονται σε κειμενικές αναφορές. Επομένως, ιδιαίτερη βαρύτητα αποδίδεται στην έγκριτη επιστημονική γνώση τόσο στο κομμάτι της ανάλυσης των δεδομένων, όσο και στο κομμάτι της ερμηνείας των αποτελεσμάτων [1, 34]. Ένα άλλο χαρακτηριστικό που διαφοροποιεί τα Big Data στον τομέα της υγείας, είναι ότι αυτά περιλαμβάνουν διαφορετικά χαρακτηριστικά των ασθενών, τα οποία πολλές φορές θα πρέπει να σταθμιστούν ανά περίπτωση.

Τέτοια χαρακτηριστικά συμπεριλαμβάνουν τη βαρύτητα της ασθένειας και την εξέλιξη της στο χρόνο, που μπορεί να αποτελεί μια πρόσθετη διάσταση, καθώς επίσης τις πληροφορίες για τη θεραπεία, που συμπεριλαμβάνουν τόσο το χρονοδιάγραμμα, όσο και τις αλλαγές που σημειώθηκαν στο θεραπευτικό σχήμα που ακολουθήθηκε [1, 35].

Ένα από τα μεγαλύτερα εμπόδια στην ευρύτερη διάδοση της αξιοποίησης των Big Data στην υγειονομική περίθαλψη, είναι το γεγονός ότι τα ιατρικά δεδομένα προέρχονται από πολλές διαφορετικές πηγές, οι οποίες τελούν υπό τη διαχείριση διαφόρων κρατών, νοσοκομείων ή και διοικητικών τμημάτων. Η ενοποίηση αυτών των πηγών δεδομένων απαιτεί την ανάπτυξη μιας νέας υποδομής, όπου όλοι οι πάροχοι των δεδομένων θα συνεργάζονται μεταξύ τους. Ο τομέας της υγείας εισέρχεται στον τομέα των Big Data με βραδείς ρυθμούς, κυρίως λόγω του αυξημένου κόστους της προσθήκης αναλυτικών λειτουργιών σε υπάρχοντα ηλεκτρονικά αρχεία υγείας, των ζητημάτων που εγείρονται ως προς το ιδιωτικό απόρρητο των δεδομένων, της χαμηλής ποιότητας των δεδομένων και της έλλειψης προθυμίας ανταλλαγής δεδομένων μεταξύ των διαφορετικών φορέων [36]. Παράλληλα έχουν αναδυθεί μια σειρά από μεθοδολογικά ζητήματα, όπως η ασυνέπεια και η ανακρίβεια των δεδομένων, οι περιορισμοί που παρουσιάζουν οι μελέτες παρατήρησης, το ζήτημα της επαλήθευσης, καθώς επίσης και τα νομικά ζητήματα που εγείρονται [1, 18].

Οι κύριες προκλήσεις για τα Big Data αφορούν στη διαθεσιμότητα, την ευκολία χρήσης, τη δυνατότητα κλιμάκωσης, τη δυνατότητα επέκτασης, την ικανότητα χειρισμού διαφορετικών επιπέδων ευαισθησίας, την προστασία της ιδιωτικής ζωής και ασφάλειας και τη διασφάλιση της ποιότητας [1,18]. Προκειμένου να είναι επιτυχημένα τα Big Data Analytics στην υγεία, θα πρέπει τα δεδομένα να είναι πακετοποιημένα, να μην υπάρχει κενό μεταξύ της συλλογής και επεξεργασίας τους, καθώς επίσης θα πρέπει να είναι φιλικά προς το χρήστη και διαφανή. Η Αναλυτική μεγάλων δεδομένων πραγματικού χρόνου αποτελεί βασική προϋπόθεση στην υγειονομική περίθαλψη. Η διαθεσιμότητα πολλών αναλυτικών αλγορίθμων, μοντέλων και μεθόδων ανάλυσης σε ένα αναπτυσσόμενο μενού τύπου pull - down αποτελεί επίσης απαραίτητη προϋπόθεση για την υιοθέτηση των Big Data σε μεγάλη κλίμακα. Επιπρόσθετα, θα πρέπει να εξεταστούν τα σημαντικά θέματα που άπτονται της διαχείρισης της ιδιοκτησίας, της διακυβέρνησης και της προτυποποίησης των

δεδομένων, εφόσον τα δεδομένα στην υγειονομική περίθαλψη σπανίως τυποποιούνται, συχνά κατακερματίζονται ή παράγονται σε παλαιότερα συστήματα πληροφορικής με ασυμβίβαστες μεταξύ τους μορφές [18]. Αυτές οι μεγάλες προκλήσεις θα πρέπει να αντιμετωπιστούν [1]. Ωστόσο, η αντίσταση των οργανισμών υγείας στον επανασχεδιασμό των διαδικασιών και στην υιοθέτηση της τεχνολογίας που επηρεάζει το σύστημα υγειονομικής περίθαλψης [37], από κοινού με την ανάγκη για τεράστιες αρχικές επενδύσεις καθιστά πιο δύσκολη τη χρήση της τεχνολογίας των Big Data [7].

Οι κύριοι τομείς στους οποίους εντοπίζονται οι αναδυόμενες προκλήσεις για τα Big Data είναι οι ακόλουθοι:

1. **Αποθήκευση:** Η αποθήκευση μεγάλου όγκου δεδομένων είναι μια από τις κύριες προκλήσεις για την τήρηση των Big Data, ωστόσο πολλοί οργανισμοί διαθέτουν τη δυνατότητα αποθήκευσης δεδομένων στις δικές τους εγκαταστάσεις. Κάτι τέτοιο συνεπάγεται σημαντικά πλεονεκτήματα συμπεριλαμβανομένου του ελέγχου της ασφάλειας, της εύκολης πρόσβασης και του χρόνου λειτουργικότητας (up-time). Ωστόσο, ένα δίκτυο εξυπηρετητών εγκατεστημένο στην έδρα ενός οργανισμού μπορεί να είναι αφενός δαπανηρό να αναπτυχθεί και αφετέρου δύσκολο να διατηρηθεί. Για το λόγο αυτό, οι περισσότεροι οργανισμοί υγειονομικής περίθαλψης στρέφονται στην εναλλακτική επιλογή της υπολογιστικής νέφους, η οποία αποτελεί μια φθηνότερη και αρκετά αξιόπιστη λύση [10, 32].
2. **Καθαρισμός:** Ο καθαρισμός των δεδομένων είναι απαραίτητος, προκειμένου να εξασφαλίζεται η ακρίβεια, η αξιοπιστία, η συνέπεια, η σχετικότητα και η καθαρότητα μετά την απόκτηση τους. Αυτή η διαδικασία καθαρισμού μπορεί να πραγματοποιείται χειροκίνητα ή να είναι αυτοματοποιημένη χρησιμοποιώντας λογικούς κανόνες για την εξασφάλιση υψηλών επιπέδων ακρίβειας και ακεραιότητας. Τα πιο εξελιγμένα και ακριβή εργαλεία χρησιμοποιούν τεχνικές εκμάθησης μηχανών (machine learning) με στόχο να μειώσουν το χρόνο και το κόστος, καθώς επίσης να παρεμποδίσουν αναξιόπιστα δεδομένα να παρεισφρήσουν και να θέσουν σε κίνδυνο έργα ανάλυσης και αξιοποίησης των Big Data [10].
3. **Ενιαία Κωδικοποίηση:** Οι ασθενείς παράγουν ένα τεράστιο όγκο δεδομένων, ο οποίος λόγω της πολυπλοκότητας και της ανομοιογένειας του, συχνά δεν

είναι διαχειρίσιμος υπό τη μορφή του παραδοσιακού ηλεκτρονικού φακέλου. Επιπλέον, είναι πολύ δύσκολη η διαχείριση των μεγάλων δεδομένων, ειδικά όταν αυτά περιέρχονται στους παρόχους υγειονομικής περίθαλψης χωρίς την απαιτούμενη οργάνωση. Συνεπώς, προκύπτει η ανάγκη μιας κωδικοποίησης όλων των πληροφοριών που σχετίζονται με την κλινική πράξη και που χρησιμοποιούνται μεταξύ άλλων για την τιμολόγηση των ασφαλιστικών απαιτήσεων, την αποζημίωση των παρόχων υγειονομικών υπηρεσιών κτλ. Ως εκ τούτου, αναπτύχθηκαν συστήματα κωδικοποίησης όπως το σύστημα της Διεθνούς Στατιστικής Ταξινόμησης των Ασθενειών (ICD-10[12]) και η κωδικοποίηση της ορολογίας των ιατρικών διαδικασιών που έχει αναπτυχθεί στις ΗΠΑ, η λεγόμενη Current Procedural Terminology (CPT), προκειμένου να αποτυπώσουν βασικές κλινικές έννοιες. Ωστόσο, αυτά τα συστήματα κωδικοποίησης δεν στερούνται των δικών τους εγγενών περιορισμών [7, 10, 32].

4. **Ακρίβεια & Αξιοπιστία:** Ορισμένες μελέτες αναφέρουν ότι οι αναφορές των δεδομένων των ασθενών σε ΗΦΥ και ηλεκτρονικά ιατρικά αρχεία εν γένει, δεν είναι απολύτως ακριβή [38], πιθανώς λόγω της περιορισμένης χρησιμότητας των τηρούμενων ηλεκτρονικών αρχείων, των περίπλοκων ροών εργασίας και της περιορισμένης κατανόησης της σημασίας συλλογής των Big Data, που να χαρακτηρίζονται από ακρίβεια. Όλοι αυτοί οι παράγοντες μπορούν να δημιουργήσουν ζητήματα αναφορικά με την ποιότητα των Big Data καθ' όλη τη διάρκεια του κύκλου ζωής τους. Η τήρηση ηλεκτρονικών ιατρικών αρχείων και ΗΦΥ αποσκοπούν στη βελτίωση της ποιότητας και της διάχυσης των δεδομένων στις κλινικές εργασιακές ροές, αν και επί του παρόντος οι σχετικές εκθέσεις δείχνουν αποκλίσεις σε αυτά τα πλαίσια. Η ποιότητα της καταγραφής των δεδομένων των ασθενών θα μπορούσε ενδεχομένως να βελτιωθεί με τη συμπλήρωση ερωτηματολογίων αυτοαναφοράς από ασθενείς για τα συμπτώματά τους [10].
5. **Προεπεξεργασία Εικόνων:** Σειρά μελετών έχει καταδείξει ότι υπάρχουν διάφοροι φυσικοί παράγοντες που μπορούν να οδηγήσουν σε αλλοιωμένη ποιότητα δεδομένων και παρερμηνείες από υπάρχοντα ιατρικά αρχεία [9]. Η ποιότητα των ιατρικών εικόνων συχνά πάσχει από τεχνικά εμπόδια που

[12] International Statistical Classification of Diseases 10th Edition

περιλαμβάνουν πολλαπλούς τύπους θορύβου και αντικειμένων. Επιπλέον, ο ακατάλληλος χειρισμός των απεικονιστικών μεθόδων μπορεί επίσης να προκαλέσει την αλλοίωση των ιατρικών εικόνων, για παράδειγμα, μπορεί να οδηγήσει στην απεικόνιση ανατομικών δομών, όπως οι φλέβες, που δεν συσχετίζονται με την ασθένεια που διερευνάται ή που μπορεί να οδηγούν σε εσφαλμένα συμπεράσματα. Η μείωση του θορύβου, ο καθαρισμός από παρεμβαλλόμενά αντικείμενα, η προσαρμογή της αντίθεσης των αποκτώμενων εικόνων και η προσαρμογή της ποιότητας τους, είναι μερικά από τα μέτρα που μπορούν να υλοποιηθούν προς όφελος αυτού του σκοπού [10].

6. **Ασφάλεια & Ιδιωτικό Απόρρητο:** Η προστασία του ιδιωτικού απόρρητου και η εμπιστευτικότητα των στοιχείων των ασθενών είναι υψίστης σημασίας στον τομέα της υγείας. Ωστόσο, η ανταλλαγή δεδομένων μεταξύ διαφόρων ενδιαφερόμενων μερών έρχεται να δημιουργήσει περαιτέρω ανησυχίες για την προστασία των προσωπικών δεδομένων των ασθενών [7, 19]. Σύμφωνα με τους Mittelstadt et al. [39] η πληροφορημένη συναίνεση και η προστασία του απορρήτου των δεδομένων των ασθενών αποτελούν καίρια ζητήματα. Δεδομένου ότι τα Big Data περιέχουν αφενός προσωπικές πληροφορίες του ασθενούς και αφετέρου το ιστορικό υγείας του, είναι σημαντικό να προστατεύεται η βάση δεδομένων από πιθανή παραβίαση, την κυβερνο-κλοπή και το ηλεκτρονικό ψάρεμα[13], όπου τα δεδομένα που έχουν κλαπεί μπορούν να πωληθούν για ένα τεράστιο ποσό. Πέραν του τομέα της υγείας, Big Data από άλλους εμπορικούς οργανισμούς, όπως εταιρείες τηλεπικοινωνιών, τράπεζες ή χρηματοπιστωτικά ιδρύματα είναι επίσης ευάλωτα και ενδέχεται να υποκλαπούν χωρίς τη γνώση των πελατών. Πριν την ευρύτερη διάδοση των Big Data, είναι απαραίτητο να διασφαλιστεί ένα υψηλό επίπεδο προστασίας της διοίκησης, ιδιωτικότητας και της ασφάλειας τους. Η προσβασιμότητα στα δεδομένα της υγειονομικής περίθαλψης πρέπει να επανεξετάζεται και να παρακολουθείται συνεχώς [32]. Κατόπιν διαπίστωσης μιας σειράς από τρωτά σημεία, αναπτύχθηκε ένας κατάλογος τεχνικών εγγυήσεων για τις Προστατευόμενες Πληροφορίες Υγείας (PHI)[14]. Αυτοί οι κανονισμοί, βοηθούν σημαντικά τους οργανισμούς κατά τη χρησιμοποίηση πρωτοκόλλων

[13] Απόδοση στην ελληνική του όρου Phishing
[14] Protected Health Information

αποθήκευσης και μετάδοσης, καθώς επίσης πρωτοκόλλων ταυτοποίησης[15] και ελέγχων ως προς την πρόσβαση και την ακεραιότητα των δεδομένων. Τα κοινά μέτρα ασφαλείας, όπως η χρήση σύγχρονου λογισμικού προστασίας από ιούς, τα τείχη προστασίας[16], η κρυπτογράφηση των ευαίσθητων δεδομένων και η ταυτοποίηση πολλών παραγόντων[17], μπορούν να προλάβουν πολλά από αυτά τα προβλήματα [10].

7. **Meta-data/ Μετα-δεδομένα:** Προκειμένου ένας οργανισμός να αναπτύξει ένα επιτυχημένο σχέδιο διαχείρισης δεδομένων, θα πρέπει να διατηρεί πλήρη, ακριβή και ενημερωμένα μεταδεδομένα σχετικά με όλα τα αποθηκευμένα δεδομένα. Τα μεταδεδομένα αποτελούνται από πληροφορίες όπως η ώρα της δημιουργίας, ο σκοπός και το πρόσωπο που είναι υπεύθυνο για τα δεδομένα, η προηγούμενη χρήση (από ποιον, γιατί, πώς και πότε) για τους ερευνητές και τους αναλυτές δεδομένων. Αυτό επιτρέπει στους αναλυτές να αναπαράγουν προηγούμενα ερευνητικά ερωτήματα και να συμβάλλουν σε μεταγενέστερες επιστημονικές μελέτες, αλλά και στην ακριβή συγκριτική αξιολόγηση[18]. Η τήρηση μεταδεδομένων αυξάνει τη χρησιμότητα των δεδομένων και αποτρέπει τη δημιουργία "κάδων απορριμμάτων δεδομένων" χαμηλής ή ελάχιστης χρήσης [10].
8. **Υποβολή ερωτημάτων προς τη βάση δεδομένων[19]:** Τα μεταδεδομένα διευκολύνουν τους οργανισμούς να υποβάλουν τα ερωτήματα τους στη βάση δεδομένων και να λαμβάνουν ορισμένες απαντήσεις. Ωστόσο, ελλείψει κατάλληλης διαλειτουργικότητας μεταξύ των συνόλων δεδομένων, τα εργαλεία αναζήτησης ενδέχεται να μην έχουν πρόσβαση στο σύνολο ενός αποθετηρίου δεδομένων. Επίσης, τα διαφορετικά στοιχεία ενός συνόλου δεδομένων θα πρέπει να είναι καλά διασυνδεδεμένα και εύκολα προσβάσιμα, διαφορετικά δεν μπορεί να δημιουργηθεί μια ολοκληρωμένη εικόνα της υγείας ενός μεμονωμένου ασθενούς. Η εφαρμογή των συστημάτων ιατρικής κωδικοποίησης όπως τα ICD-10, το SNOMED-CT ή το LOINC είναι επιβεβλημένη, ώστε να μπορεί χρησιμοποιηθεί η δομημένη γλώσσα

[15] Identification protocols
[16] Firewalls
[17] Multi-factor authentication
[18] Benchmarking
[19] Querying

ερωτημάτων (SQL) για την υποβολή ερωτημάτων σε μεγάλα σύνολα δεδομένων και σε σχεσιακές βάσεις δεδομένων [10].

9. **Οπτικοποίηση:** Μια ξεκάθαρη και ελκυστική απεικόνιση των δεδομένων με γραφήματα και ιστογράμματα για την απεικόνιση αντιθέτων μεγεθών και την ορθή επισήμανση των πληροφοριών μπορεί να διευκολύνει την απορρόφηση των πληροφοριών και την κατάλληλη αξιοποίηση τους. Άλλοι μέθοδοι οπτικοποίησης περιλαμβάνουν διαγράμματα, διαγράμματα πίτας και διαγράμματα σκέδασης[20] [10].
10. **Κοινοχρησία δεδομένων[21]:** Οι ασθενείς μπορεί να λαμβάνουν φροντίδα σε πολλές διαφορετικές περιοχές. Στην περίπτωση αυτή, η ανταλλαγή δεδομένων μεταξύ των διαφορετικών οργανισμών υγειονομικής περίθαλψης είναι απαραίτητη. Κατά τη διάρκεια αυτής της κοινής χρήσης, εάν τα δεδομένα δεν είναι διαλειτουργικά, τότε η ανταλλαγή δεδομένων μεταξύ διαφορετικών οργανώσεων θα μπορούσε να αντιμετωπίσει σημαντικούς περιορισμούς. Κάτι τέτοιο μπορεί να οφείλεται σε τεχνικά και οργανωτικά εμπόδια [7, 10], ενώ ενδέχεται να αφήσει τους ιατρούς χωρίς βασικές πληροφορίες για τη λήψη αποφάσεων σχετικά με τις επακόλουθες ενέργειες και τις στρατηγικές θεραπείας για τους ασθενείς. Λύσεις όπως το πρότυπο ανταλλαγής δεδομένων υγείας, FHIR[22], οι δημόσιες Διεπαφές Προγραμματισμού (API)[23], η CommonWell (μη κερδοσκοπική εμπορική ένωση) και η Carequality (κοινό πλαίσιο διαλειτουργικότητας), καθιστούν εφικτή τη διαλειτουργικότητα, ενώ διευκολύνουν την εύκολη και ασφαλή ανταλλαγή δεδομένων. Το σημαντικότερο εμπόδιο για την κοινοχρησία δεδομένων είναι η αντιμετώπιση των δεδομένων ως προϊόν που μπορεί να προσφέρει ένα σημαντικό ανταγωνιστικό πλεονέκτημα. Επομένως, μερικές φορές τόσο οι πάροχοι, όσο και οι πωλητές παρεμβαίνουν σκόπιμα στη ροή πληροφοριών για να εμποδίσουν τη ροή πληροφοριών μεταξύ διαφορετικών συστημάτων ηλεκτρονικών αρχείων υγείας [40].

Οι πάροχοι υγειονομικής περίθαλψης θα πρέπει αφενός να ξεπεράσουν όλες αυτές τις προκλήσεις και αφετέρου να αναπτύξουν ένα μεγάλο οικοσύστημα ανταλλαγής

[20] Scatterplot
[21] Data sharing
[22] Fast Healthcare Interoperability Resource
[23] Application Programming Interface

δεδομένων που παρέχει αξιόπιστες, έγκαιρες και ουσιαστικές πληροφορίες συνδέοντας όλα τα σημεία του συνεχούς της φροντίδας. Και για να ξεπεραστούν αυτές οι προκλήσεις, χρειάζονται χρόνος, δέσμευση, χρηματοδότηση και επικοινωνία [10]. Οι στρατηγικές για την αντιμετώπιδη των προαναφερθέντων ζητημάτων περιλαμβάνουν:

- **Διαχείριση των δεδομένων:** Λόγω της εσφαλμένης διαχείρισης, οι οργανισμοί υγειονομικής περίθαλψης επιβαρύνονται με τεράστια κόστη για επενδύσεις σε ΤΠΕ. Η κατάλληλη και αποτελεσματική διαχείριση των δεδομένων μπορεί να συντελέσει στη δημιουργία πρόσθετης επιχειρηματικής αξίας [7, 15].
- **Ανάπτυξη μιας κουλτούρας ανταλλαγής δεδομένων**: Η ανταλλαγή πληροφοριών και η συγκέντρωση των δεδομένων μπορούν να αντιμετωπίσουν το ζήτημα της διαλειτουργικότητας και να επιτρέψουν την αποτελεσματική αξιοποίηση των δυνατοτήτων των αναλυτικής των Big Data Analytics, ενισχύοντας την προγνωστική τους αξία [7, 15, 41].
- **Χρήση μέτρων ασφαλείας:** Η ισχυρή κρυπτογράφηση των δεδομένων, η επικύρωση της πηγής των δεδομένων, ο έλεγχος πρόσβασης, από κοινού με την ταυτοποίηση και την αποταυτοποίηση[24], είναι μερικά από τα μέτρα διασφάλισης των δεδομένων και διατήρησης της εμπιστευτικότητας τους [7].
- **Εκπαίδευση προσωπικού στη χρήση των Big Data Analytics:** Προκειμένου να εξαχθούν ουσιαστικές γνώσεις και πολύτιμες πληροφορίες από τα Big Data, οι επαγγελματίες υγείας θα πρέπει να εκπαιδεύονται, ώστε να αποκτήσουν δεξιότητες ανάλυσης των Big Data. Αυτό είναι κρίσιμο για τον τομέα της υγείας, διότι η εσφαλμένη ερμηνεία των αναφορών που δημιουργούνται από την ανάλυση των Big Data μπορεί να οδηγήσει σε απρόβλεπτες συνέπειες [7, 15].
- **Αξιοποίηση της υπολογιστικής νέφους:** Η πρόκληση της αποθήκευσης των τεράστιων όγκων δεδομένων μπορεί να αντιμετωπιστεί με τη χρήση του cloud computing. Κάτι τέτοιο θα επέτρεπε στα μικρά και μεσαίου μεγέθους νοσοκομεία και οργανισμούς παροχής υγενιονομικής φροντίδας να εξαλείφουν τα θέματα κόστους και αποθήκευσης δεδομένων [7,15].

[24] Deidentification

ΚΕΦΑΛΑΙΟ ΙΙ. Τα Big Data Analytics στον τομέα της Υγείας

2.1 Ορισμός των Big Data Analytics στον τομέα της Υγείας

Σύμφωνα με ένα ρητό του Peter Sondergaard «*Η πληροφορία είναι το καύσιμο του 21ου αιώνα και η ανάλυση τους είναι ο κινητήρας*." Τα δεδομένα που συλλέγονται σε διάφορα αποθετήρια από διάφορους οργανισμούς, καθώς επίσης τα δεδομένα που παράγονται από τα μεμονωμένα άτομα μπορούν να κάνουν τη διαφορά μόνο αν αναλυθούν και χρησιμοποιηθούν σωστά. Με άλλα λόγια, χωρίς την κατάλληλη ανάλυση, τα δεδομένα θα αποτελούν απλά ένα πόρο για τον οργανισμό, ο οποίος όμως δεν αξιοποιείται. Επιπλέον, θα πρέπει να σημειωθεί ότι ο όρος Big Data δεν παραπέμπει μόνο στον όγκο των δεδομένων, αλλά και στην ισχύ τους. Τα σύνολα δεδομένων είναι μεγάλα και σύνθετα, θέτοντας τεράστιες προκλήσεις για τις υφιστάμενες τεχνικές, που επιχειρούν να τα αναλύσουν και να καταγράψουν αποτελέσματα και πιθανούς συσχετισμούς [11].

Ένας όρος που τυγχάνει ευρείας χρήσης σήμερα είναι τα Big Data Analytics ή Αναλυτική Μεγάλων Δεδομένων. Πρόκειται για μια διαδικασία μετασχηματισμού ακατέργαστων δεδομένων σε πληροφορία. Ο όρος αυτός επί της ουσίας περιγράφει την ανάγκη να εξαχθούν πολύτιμα συμπεράσματα και προβλέψεις από τα δεδομένα, αποσκοπώντας αφενός στην υποβοήθηση της λήψης αποφάσεων και αφετέρου στη βελτίωση των διαδικασιών των επιμέρους οργανισμών [14]. Προκειμένου να υποστηρίξουν τη λήψη αποφάσεων στον ταχέως αναπτυσσόμενο επιχειρηματικό τομέα, τα Big Data Analytics επιχειρούν να αποκαλύψουν κρυφά μοτίβα, να εντοπίσουν άγνωστες συσχετίσεις, να κατανοήσουν τις τάσεις της αγοράς, τις προτιμήσεις των πελατών και να αντλήσουν άλλες χρήσιμες επιχειρηματικές πληροφορίες [11].

Η Αναλυτική δεδομένων στην Υγεία ή Health Analytics θα μπορούσε να οριστεί ως *«η συστηματική χρήση ιατρικών δεδομένων και σχετικών πληροφοριών διαχείρισης μέσω της εφαρμογής μεθόδων και εργαλείων ανάλυσης, καθώς και ποσοτικών και ποιοτικών στατιστικών στοιχείων, ανάλυσης περιβάλλοντος και προβλέψεων για την εξαγωγή συμπερασμάτων, τα οποία μπορούν να αποτελέσουν τη βάση για την ανάληψη*

δράσης, καθώς και την ανάπτυξη μιας στρατηγικής και επιχειρησιακής διαχείρισης που βασίζεται στην πληροφόρηση, για μια καλύτερη υγειονομική περίθαλψη»[43].

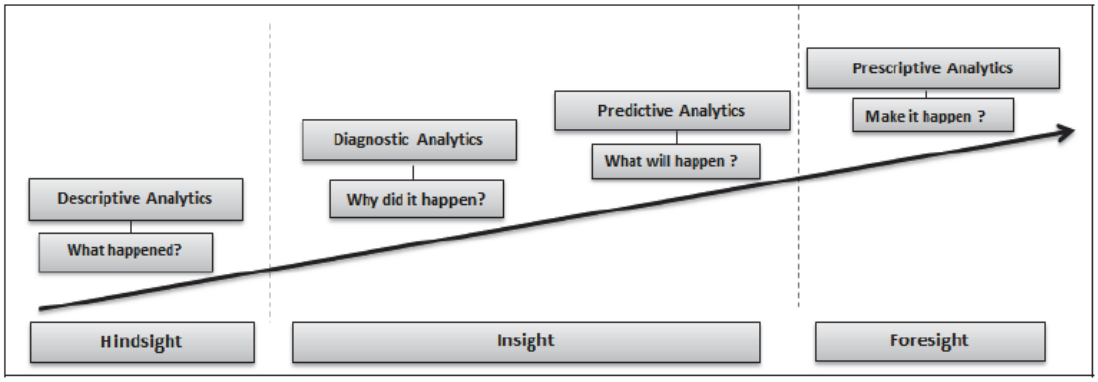
Επιπλέον, η Αναλυτική δεδομένων στην Υγεία ή Health Analytics είναι ένας επιχειρησιακός όρος που περιλαμβάνει ένα ευρύ φάσμα εφαρμογών επιχειρηματικής ευφυΐας και ανάλυσης Big Data. Αυτή η νεοεισηχθείσα έννοια βασίζεται στη διαθεσιμότητα και προσβασιμότητα δεδομένων και πληροφοριών που συγκεντρώνονται μέσω της καλής ενσωμάτωσης και διαλειτουργικότητας ενός ευρέος φάσματος συστημάτων πληροφοριών για την υγεία, όπως τα ηλεκτρονικά ιατρικά αρχεία, την αρχειοθέτηση εικόνων και το σύστημα επικοινωνίας, τα πληροφοριακά συστήματα των εργαστηρίων και τα back-end συστήματα αποθήκευσης δεδομένων υγειονομικής περίθαλψης [43]. Κινούμενοι προς την ίδια κατεύθυνση, οι ερευνητές Frost & Sullivan [42] ήταν εκείνοι που πρότειναν την έννοια της Προηγμένης Αναλυτικής στην Υγεία (Advanced Health Analytics).

Πρόσφατα, ο αριθμός των πηγών δεδομένων στον τομέα της υγείας έχει αυξηθεί ραγδαία ως αποτέλεσμα της ευρείας χρήσης τεχνολογιών αισθητήρων για έξυπνα κινητά τηλέφωνα και φορητά συστήματα. Επομένως, καθίσταται δύσκολη η διεξαγωγή ανάλυσης των δεδομένων αυτών με βάση τις παραδοσιακές αναλυτικές μεθόδους, που κρίνονται ακατάλληλες για χειρισμό του μεγάλου όγκου των ετερογενών ιατρικών δεδομένων [44]. Υπό αυτό το πρίσμα, οι οργανισμοί υγειονομικής περίθαλψης διέρχονται της διαδικασίας πλήρους αξιοποίησης των τεράστιων ποσοτήτων δεδομένων και πληροφοριών που διακρατούν μέσω της αξιοποίησης των Big Data Analytics [43, 45].

Τα Big Data Analytics περιλαμβάνουν την ενσωμάτωση ετερογενών δεδομένων, τον έλεγχο της ποιότητας των δεδομένων, την ανάλυση, τη μοντελοποίηση, και την ερμηνεία και την εξακρίβωση [46-47]. Ιδιαίτερα, η χρησιμοποίηση τους στην ιατρική και την υγεία εν γένει επιτρέπουν την ανάλυση των μεγάλων συνόλων δεδομένων από χιλιάδες ασθενείς, την ταυτοποίηση συστάδων και τη συσχέτιση μεταξύ συνόλων δεδομένων, καθώς και την ανάπτυξη προγνωστικών μοντέλων με τη χρήση τεχνικών εξόρυξης δεδομένων[25] [48]. Η Αναλυτική Μεγάλων Δεδομένων στην ιατρική και την υγειονομική περίθαλψη ενσωματώνει την ανάλυση διαφόρων επιστημονικών πεδίων όπως η βιοπληροφορική, η ιατρική απεικόνιση, η πληροφορική των αισθητήρων, η

[25] Data mining techniques

ιατρική πληροφορική και η υγειονομική πληροφορική. Η νέα γνώση που θα προκύψει από τις τεχνικές ανάλυσης μεγάλων δεδομένων αναμένεται να παρέχει τεράστια οφέλη στους ασθενείς, τους κλινικούς ιατρούς και τους υπεύθυνους χάραξης πολιτικής στον τομέα της υγείας [46].



Εικόνα 2. Ταξινόμηση των Big Data Analytics στον τομέα της Υγείας

Κατά τη διάρκεια των τελευταίων δύο δεκαετιών, η Αναλυτική δεδομένων στην Υγεία έχει αναδυθεί ως ένα σημαντικό επιστημονικό πεδίο, τόσο για τους ερευνητές όσο και για τους επαγγελματίες υγείας, αντικατοπτρίζοντας το μέγεθος της επιρροής της διοίκησης με βάση την πληροφόρηση στην επίλυση των προβλημάτων και στη λήψη αποφάσεων [49]. Τα συστήματα πληροφόρησης για την υγεία υιοθετούνται ταχέως παγκοσμίως, γεγονός που θα αυξήσει σημαντικά την ποσότητα και θα βελτιώσει την ποιότητα των διαθέσιμων δεδομένων για την υγεία. Ταυτόχρονα, έχει πραγματοποιηθεί επαναστατική πρόοδος στις αναλυτικές μεθόδους για την ανάλυση τεράστιων ποσοτήτων δεδομένων και την απόκτηση νέων γνώσεων. Συνεπώς, αναδύονται άνευ προηγούμενου ευκαιρίες για τη χρήση τέτοιων μεθόδων για τη βελτίωση της ποιότητας και τη μείωση του κόστους της υγειονομικής περίθαλψης [43, 50].

2.2 Ταξινόμηση Big Data Analytics στην Υγεία

Τα Big Data Analytics στην υγεία μπορούν να ταξινομηθούν σε Descriptive (Περιγραφική Αναλυτική), Diagnostic (Διαγνωστική Αναλυτική), Predictive (Προγνωστική Αναλυτική) και Prescriptive (Καθοδηγητική Αναλυτική) [11, 14, 43].

Εικόνα 3. Περιγραφική, Διαγνωστική, Προγνωστική & Καθοδηγητική Αναλυτική στον τομέα της Υγείας

2.2.1. Περιγραφική Αναλυτική (Descriptive Analytics)

Η Περιγραφική Αναλυτική συνίσταται στην περιγραφή της υφιστάμενης κατάστασης και συμβάλλει στη σκιαγράφηση της εικόνας σχετικά με τις προηγούμενες επιδόσεις, στη βάση των ιστορικών δεδομένων και μέσω της χρησιμοποίησης της επιχειρηματικής πληροφόρησης και της εξόρυξης δεδομένων. Η χρήση αυτού του αναλυτικού στοιχείου αποτελεί τη βάση, επί της οποίας θα δομηθεί η ενδεδειγμένη προσέγγιση για την επίτευξη του στόχου [11, 14]. Για την εκτέλεση αυτού του επιπέδου ανάλυσης χρησιμοποιούνται διάφορες τεχνικές [44]. Η Περιγραφική Αναλυτική είναι επίσης γνωστή και ως μάθηση χωρίς επίβλεψη[26]. Ειδικότερα, συνοψίζει:

- Τι συνέβει στη διοίκηση των υπηρεσιών υγείας;
- Ποιος είναι ο αντίκτυπος μιας παραμέτρου στο σύστημα; [14].

Η Περιγραφική Αναλυτική είναι το απλούστερο επίπεδο κατανόησης και χρήσης. Απλώς περιγράφει τα δεδομένα χωρίς περαιτέρω αναλύσεις, διερευνήσεις ή συσχετισμούς μεταξύ μεταβλητών ή στοιχείων πληροφοριών, ενώ είναι απόλυτα ελεγχόμενη. Οι περιγραφικές αναλύσεις λειτουργούν κατηγοριοποιώντας, χαρακτηρίζοντας, συγκεντρώνοντας και ταξινομώντας τα δεδομένα, τα οποία πρέπει να μετατραπούν σε πολύτιμες πληροφορίες για να βοηθήσουν τους επαγγελματίες του τομέα υγείας να κατανοήσουν και να αναλύσουν τις αποφάσεις, τις επιδόσεις και τα αποτελέσματα. Η παρουσίαση των δεδομένων αναπαριστώνται συνήθως με απλά γραφήματα και πίνακες που παρουσιάζουν π.χ. τα ποσοστά πληρότητας του νοσοκομείου, τα εξιτήρια, τη μέση διάρκεια νοσηλείας και άλλους συναφείς δείκτες.

[26] Unsupervised learning

Η οπτικοποίηση των δεδομένων χρησιμοποιείται για να βοηθήσει στο να απαντηθούν συγκεκριμένες ερωτήσεις ή να προσδιοριστούν τα πρότυπα φροντίδας, παρέχοντας έτσι μια ευρύτερη εικόνα για την τεκμηριωμένη κλινική πρακτική[27]. Επιτρέπουν τη διαχείριση δεδομένων σε πραγματικό χρόνο ή περίπου σε πραγματικό, ενώ καταγράφουν τα οπτικά δεδομένα όλων των ασθενών ή ηλεκτρονικά ιατρικά αρχεία. Αυτή η δυνατότητα επιτρέπει τον εντοπισμό μοτίβων σε ασθενείς, τα οποία δεν είχαν γίνει αντιληπτά στο παφρελθόν και που σχετίζονται με τις νοσοκομειακές επανεισαγωγές [3, 44-45].

2.2.2 Διαγνωστική Αναλυτική (Diagnostic Analytics)

Η χρήση των ιστορικών δεδομένων προβλέπει την κύρια αιτία του προβλήματος και τη διάγνωση. Στόχος της Διαγνωστικής Αναλυτικής είναι να εξηγήσει γιατί συνέβησαν ορισμένα γεγονότα και ποιοι είναι οι παράγοντες που τα προκάλεσαν. Για παράδειγμα, η Διαγνωστική Αναλυτική επιχειρεί να κατανοήσει τους λόγους πίσω από τις συχνές επανεισαγωγές μερικών ασθενών, χρησιμοποιώντας διάφορες μεθόδους, όπως η συσταδοποίηση[28] και τα δέντρα απόφασης [11, 14]. Χρειάζεται εκτεταμένη διερεύνηση και κατευθυνόμενη ανάλυση των υφιστάμενων δεδομένων χρησιμοποιώντας εργαλεία όπως τεχνικές απεικόνισης, προκειμένου να ανακαλυφθούν οι ρίζες ενός προβλήματος και να βοηθηθούν οι χρήστες στο να να συνειδητοποιήσουν τη φύση και τον αντίκτυπο των προβλημάτων. Αυτό μπορεί να περιλαμβάνει την κατανόηση της επίδρασης των παραγόντων και των διαδικασιών που αποτελούν εισροές του συστήματος, στην απόδοση. Για παράδειγμα, ο αυξημένος χρόνος αναμονής για την παροχή ορισμένων υπηρεσιών υγειονομικής περίθαλψης θα μπορούσε να αποδοθεί σε πολλούς σημαντικούς παράγοντες, όπως παράγοντες σχετιζόμενους με τον ασθενή, τους παρόχους ή τον οργανισμό [43, 51].

2.2.3 Προγνωστική Αναλυτική (Predictive Analytics)

Η Προγνωστική Αναλυτική δουλεύει με ένα πιο περίπλοκο τρόπο από την Περιγραφική Αναλυτική, εφόσον επικεντρώνεται στη χρήση της πληροφορίας παρά απλώς στα δεδομένα. Εξετάζει υπάρχουσες ερμηνείες και δείκτες του παρελθόντος

[27] Evidence based clinical practice
[28] Clustering

για να προβλέψει τις μελλοντικές επιδόσεις [43]. Αντικατοπτρίζει την ικανότητα πρόβλεψης μελλοντικών γεγονότων, ενώ συμβάλλει στον εντοπισμό των τάσεων και τον προσδιορισμό πιθανών αβέβαιων αποτελεσμάτων. Επί παραδείγματι μπορεί να κληθεί να προβλέψει εάν ένας ασθενής θα εμφανίσει επιπλοκές ή όχι. Τα προγνωστικά μοντέλα κατασκευάζονται συχνά χρησιμοποιώντας τεχνικές εκμάθησης μηχανών[29] [44].

Τα Predictive Analytics χρησιμοποιούν τα τεράστια σύνολα δεδομένων, προκειμένου να βελτιώσουν την εμπειρία των πελατών, βελτιώνοντας τα αποτελέσματα συγκριτικά με τις συμβατικές επιχειρηματικές στρατηγικές. Χρησιμοποιούνται για την ανάλυση μεγάλων όγκων δεδομένων, καθώς και μη δομημένων δεδομένων, τα οποία παράγουν τα αποτελέσματα, ώστε να προβλέψουν τις μελλοντικές εξελίξεις. Η πρόβλεψη των μελλοντικών εξελίξεων, με βάση τα διαθέσιμα σύνολα δεδομένων, αποτελεί μια δύσκολη αποστολή από καταβολής της επιστήμης της πληροφορικής. Τα προγράμματα επιχειρηματικής ευφυΐας αυτού του είδους βοηθούν στον υπολογισμό των ροών δεδομένων σε μεγαλύτερη έκταση, συμπεριλαμβανομένου του περιεχομένου των μέσων κοινωνικής δικτύωσης, των αγοραστικών εμπειριών, των καθημερινών δραστηριοτήτων των χρηστών και των ερευνών [11].

Αναλύουν τόσο δεδομένα σε πραγματικό χρόνο, όσο και ιστορικά δεδομένα, ενώ είναι γνωστά και ως εποπτευόμενη μάθηση[30]. Μπορούν μόνο να προβλέψουν τι μπορεί να συμβεί στο μέλλον, διότι όλα τα προβλεπόμενα σενάρια είναι πιθανά, αλλά δεν μπορεί να προβλέψει το ίδιο το μέλλον. Ειδικότερα, επιχειρούν να δώσουν απαντήσεις στα ακόλουθα ερωτήματα:

- Τι θα συμβεί; Ποιες είναι οι μελλοντικές τάσεις;
- Ποια απόφαση θα ληφθεί βάσει του προηγούμενου ιστορικού; [14].

Επί παραδείγματι, ένας φαρμακοποιός μπορεί να χρειαστεί να γνωρίζει σε τι ποσότητες θα πρέπει να διατηρήσει το απόθεμα ενός φαρμακευτικού σκευάσματος εν αναμονή της εμφάνισης μιας επιδημικής έξαρσης. Ορισμένες εξελίξεις ή κλινικά αποτελέσματα ασθενών θα μπορούσαν να προβλεφθούν και να αξιολογηθούν με βάση τις τεράστιες ποσότητες δεδομένων που συλλέχθηκαν το προηγούμενο

[29] Machine learning
[30] Supervised learning

διάστημα, όπως η διάρκεια νοσηλείας του ασθενούς, οι ασθενείς που μπορεί να επιλέξουν να υποβληθούν σε χειρουργική επέμβαση, οι ασθενείς που πιθανόν δεν θα επωφεληθούν από μια τέτοια χειρουργική επέμβαση ή θα παρουσιάσουν επιπλοκές ή ακόμα και η θνησιμότητα [43].

2.2.4 Καθοδηγητική Αναλυτική (Prescriptive Analytics)

Η Καθοδηγητική Αναλυτική αναλαμβάνει δράση στην περίπτωση που θα πρέπει να ληφθούν αποφάσεις σχετικά με ένα ευρύ φάσμα εφικτών εναλλακτικών λύσεων, επιτρέποντας στους ιθύνοντες ενός οργανισμού όχι μόνο να εξετάζουν τις συνέπειες και τα αναμενόμενα αποτελέσματα των αποφάσεών, αλλά να διαγνώσουν τις αναδυόμενες ευκαιρίες ή προβλήματα και προτείνοντας τους την καλύτερη πορεία δράσης, ώστε να αξιοποιήσουν εγκαίρως την ανάλυση που τους παρέχεται [43]. Αυτή η αναλυτική μέθοδος συνθέτει αυτόματα τα Big Data και παρέχει καθοδήγηση για ένα μεγάλο αριθμό διαφορετικών πιθανών αποτελεσμάτων πριν από την πραγματική λήψη των αποφάσεων. Ο υπεύθυνος λήψης αποφάσεων μπορεί να λάβει υπόψει αυτές τις πληροφορίες και να πράξει αναλόγως. Η Καθοδηγητική Αναλυτική παρέχει καθοδήγηση ως προς:

- Τι θα πρέπει να κάνουμε;
- Ποιο είναι το καλύτερο αποτέλεσμα και πως μπορούμε να το πετύχουμε; [14].

Η επιτυχία της Καθοδηγητικής Αναλυτικής εξαρτάται κυρίως από την υιοθέτηση πέντε βασικών παραμέτρων: τη χρήση υβριδικών δεδομένων, συμπεριλαμβανομένων των δομημένων και μη δομημένων δεδομένων, την ενσωμάτωση πρόγνωσης και καθοδήγησης λαμβάνοντας υπόψη όλες τις πιθανές παρενέργειες, τη χρήση προσαρμοστικών αλγόριθμων που μπορούν εύκολα να προσαρμοστούν σε κάθε κατάσταση, καθώς επίσης και τη χρήση ισχυρών και αξιόπιστων μηχανισμών ανάδρασης [43].

2.3 Επίδραση των Big Data Analytics στον τομέα της Υγείας

Το Big Data Analytics έχουν τη δυνατότητα να μεταμορφώσουν ριζικά τα υφιστάμενα επιχειρηματικά και κλινικά μοντέλα για μια έξυπνη και αποτελεσματική παροχή φροντίδας [23]. Επιτρέπουν την ενσωμάτωση των απο-ταυτοποιημένων

πληροφοριών για την υγεία, ώστε να επιτρέπονται οι δευτερεύουσες χρήσεις των δεδομένων [52]. Επίσης, μέσω της αναγνώρισης των μοτίβων και της αποκρυπτογράφησης των συσχετισμών, μπορούν να διευκολύνουν την αυτόνομη λήψη αποφάσεων [53]. Στην κλινική πρακτική, τα Big Data Analytics μπορεί να βοηθήσουν στην έγκαιρη ανίχνευση της νόσου, στην ακριβή πρόβλεψη της πορείας της και στον εντοπισμό της απόκλισης από την υγιή κατάσταση, την παρουσίαση επιπλοκών, καθώς επίσης και στην ανίχνευση της απάτης. Παρέχοντας αυτές τις πληροφορίες, βοηθούν τους οργανισμούς υγειονομικής περίθαλψης να εξατομικεύουν τις προβλέψεις, να παράσχουν στοχοθετημένη θεραπεία συνυπολογίζοντας τη σχέση κόστους-αποτελεσματικότητας της περίθαλψης, να μειώσουν τη σπατάλη πόρων, καθώς και να ενθαρρύνουν τα μεμονωμένα άτομα να διατηρήσουν την υγεία τους μέσω της παροχής σχετικών συστάσεων [3, 26, 50]. Τα Big Data Analytics παρέχουν την ευκαιρία ανίχνευσης γεγονότων σχετικά χαμηλής συχνότητας, τα οποία ωστόσο μπορεί να έχουν σημαντικό κλινικό αντίκτυπο. Πέραν τούτου, η ομογενοποίηση των κλινικών δεδομένων και η αποτελεσματική χρήση τους υποστηρίζουν ένα ευρύ φάσμα εφαρμογών, όπως η παρακολούθηση των ασθενειών, τα συστήματα υποστήριξης κλινικών αποφάσεων, την ατομική διαχείριση της υγειονομικής περίθαλψης, τη βελτίωση της αποτελεσματικότητας και της ποιότητας της υγειονομικής περίθαλψης, και τη μείωση του κόστους της [3, 7, 54].

Η μελέτη των Sukumar et al. [23] αποκαλύπτει ότι η ενσωμάτωση των Big Data Analytics στον τομέα της υγείας μπορεί να δώσει απάντηση σε 8 καίρια ερωτήματα που ταλανίζουν την υγεία:

1. Πώς μπορεί να αυξηθεί το κόστος για διάφορες πτυχές της υγειονομικής περίθαλψης στο μέλλον;
2. Πως επηρεάζουν ορισμένες αλλαγές σε όρους πολιτικής το κόστος και την συμπεριφορά;
3. Πως διαφοροποιούνται τα κόστη της υγειονομικής περίθαλψης ανά γεωγραφική περιοχή;
4. Μπορούν να εντοπιστούν ψευδείς απαιτήσεις;
5. Ποιες επιλογές θεραπείας φαίνονται πιο αποτελεσματικές για διάφορες ασθένειες;
6. Γιατί φαίνεται ότι κάποιοι πάροχοι έχουν καλύτερα αποτελέσματα για την υγεία;
7. Γιατί οι ασθενείς επιλέγουν έναν πάροχο έναντι άλλου;

8. Υπάρχουν ενδείξεις κάποιας επιδημίας;

Οι τομείς στους οποίους, η αναλυτική των δεδομένων αναμένεται να αποφέρει τα μεγαλύτερα αποτελέσματα περιλαμβάνουν: τον εντοπισμό των ασθενών, οι οποίοι είναι οι συχνότεροι χρήστες των υπηρεσιών υγείας ή αντιμετωπίζουν μεγαλύτερο κίνδυνο εμφάνισης αρνητικών αποτελεσμάτων, την παροχή στους ασθενείς των απαραίτητων πληροφοριών για τη λήψη τεκμηριωμένων αποφάσεων και συνεπώς την καλύτερη διαχείριση της υγείας τους, την ευκολότερη υιοθέτηση και παρακολούθηση πιο υγιεινών συμπεριφορών, τον προσδιορισμό θεραπειών, προγραμμάτων και διαδικασιών που δεν παρέχουν αποδεδειγμένα οφέλη ή κοστίζουν υπερβολικά, τη μείωση των επανεισαγωγών μέσω του προσδιορισμού των παραγόντων του περιβάλλοντος ή του τρόπου ζωής που αυξάνουν τον κίνδυνο ή προκαλούν ανεπιθύμητα συμβάντα και της ανάλογης προσαρμογής των θεραπευτικών σχεδίων, την προώθηση εφαρμογών παρακολούθησης της πορείας της υγείας των ασθενών στο σπίτι, τη διαχείριση της υγείας του πληθυσμού με την ανίχνευση των επιμέρους αναγκών κατά τη διάρκεια εμφάνισης επιδημιών ή καταστροφών, και τέλος τη συγκέντρωση κλινικών, χρηματοοικονομικών και επιχειρησιακών δεδομένων για την παραγωγική και σε πραγματικό χρόνο ανάλυση της αξιοποίησης των πόρων χρόνο [3].

ΚΕΦΑΛΑΙΟ ΙΙΙ. Εφαρμογές Big Data στον τομέα της υγείας

Όπως προαναφέρεται η μεγάλη πρόκληση με τα Big Data είναι ο τρόπος χειρισμού αυτού του μεγάλου όγκου πληροφοριών. Προκειμένου να διατεθούν στην επιστημονική κοινότητα, τα δεδομένα πρέπει να αποθηκεύονται σε μορφή αρχείου που να είναι εύκολα προσβάσιμο και να είναι εύκολο να διαβαστεί, ώστε να πραγματοποιηθεί η αποτελεσματική ανάλυση τους. Στο πλαίσιο του τομέα της υγείας, μια άλλη σημαντική πρόκληση είναι η εφαρμογή εργαλείων πληροφορικής, πρωτοκόλλων και εξοπλισμού υψηλού επιπέδου στο κλινικό περιβάλλον. Και για την επίτευξη αυτού του στόχου, θα πρέπει να συνεργαστούν επιστήμονες προερχόμενοι από διαφορετικά περιβάλλοντα, όπως η βιολογία, η πληροφορική, η στατιστική και τα μαθηματικά [3, 10, 44].

Η ετερογένεια των δεδομένων είναι μια άλλη πρόκληση στην ανάλυση των μεγάλων δεδομένων. Το τεράστιο μέγεθος και ο εξαιρετικά ετερογενής χαρακτήρας των Big Data στην υγεία, καθιστά δυσκολότερη την εξαγωγή συμπερασμάτων όταν χρησιμοποιούνται συμβατικές τεχνολογίες ανάλυσης. Οι πιο συνηθισμένες πλατφόρμες για τη λειτουργία του λογισμικού που καθιστά εφικτή την ανάλυση των Big Data είναι οι συστάδες υψηλής υπολογιστικής ισχύος μέσω υποδομών υπολογιστικού πλέγματος[31]. Η υπολογιστική νέφους είναι ένα τέτοιο σύστημα, εφόσον διαθέτει εικονικές τεχνολογίες αποθήκευσης και παρέχει αξιόπιστες υπηρεσίες. Προσφέρει υψηλή αξιοπιστία, κλιμάκωση[32] και αυτονομία μαζί με τη δυναμική ανακάλυψη πόρων και τη συνθεσιμότητα[33]. Τέτοιες πλατφόρμες μπορούν να λειτουργήσουν ως δέκτες δεδομένων από τους πανταχού παρόντες αισθητήρες, ως υπολογιστές που αναλύουν και ερμηνεύουν τα δεδομένα, καθώς επίσης μπορούν να παρέχουν στον χρήστη τα δεδομένα οπτικοποιημένα, προκειμένου να είναι περισσοτερο κατανοητά. Επιπλέον, στο λεγόμενο Διαδίκτυο των Πραγμάτων[34] η επεξεργασία και ανάλυση των Big Data μπορούν να πραγματοποιηθούν πλησιέστερα στην πηγή των δεδομένων, χρησιμοποιώντας τις υπηρεσίες κινητών υπολογιστών και

[31] Grid computing
[32] Scalability
[33] Composability
[34] Internet of Things

υπολογιστικής ομίχλης[35]. Απαιτούνται προηγμένοι αλγόριθμοι για την εφαρμογή προσεγγίσεων εκμάθησης μηχανών και τεχνητής νοημοσύνης για την ανάλυση των Big Data συστάδες εξυπηρετητών. Για γράψιμο τέτοιων αλγορίθμων ή λογισμικού θα μπορούσε να χρησιμοποιηθεί μια γλώσσα προγραμματισμού κατάλληλη για την επεξεργασία των Big Data (π.χ. Python, R ή άλλες γλώσσες). Ως εκ τούτου, απαιτείται καλή γνώση της βιολογίας και της πληροφορικής για τη διαχείριση των Big Data από τη βιοϊατρική έρευνα. Μεταξύ των διαφόρων πλατφορμών που χρησιμοποιούνται για την επεξεργασία των Big Data, το Hadoop και Apache Spark είναι οι πλέον διαδεδομένες [10, 44].

3.1 Μη Σχεσιακές Βάσεις Δεδομένων

Οι λεγόμενες σχεσιακές βάσεις δεδομένων, οι οποίες δομούνται βάσει της δομημένης γλώσσας ερωτημάτων SQL, υπήρξε επί σειράς ετών ο δημοφιλέστερος τρόπος για τη διαχείριση δεδομένων από πλευράς οργαμισμών και επαγγελματιών υγείας. Η ανάδυση των Big Data κατέστησε επιτακτική την ανάγκη επεξεργασίας των ετερογενών και τεράστιου όγκου, δεδομένων σε μεγάλη κλίμακα, προκειμένου να είναι εφικτό να εξαχθούν ενιαία συμπεράσματα. Ωστόσο, τα βασισμένα στη γλώσσα SQL συστήματα δεν μπορούν να αντιμετωπίσουν το ζήτημα της διαχείρισης των Big Data κατά τρόπο αυτόνομο [55-56]. Τη λύση ήρθαν να δώσουν οι μη σχεσιακές βάσεις δεδομένων[36], (NoSQL Databases), οι οποίες μεταξύ άλλων επιτρέπουν τη δυναμική διαχείριση των δεδομένων και παρέχουν αυξημένες δυνατότητες ως προς την ευελιξία και την κλιμάκωση, σε σύγκριση με τις SQL βάσεις δεδομένων. Αυτά τα εγγενή χαρακτηριστικά, τις αναδεικνύουν ως την ιδανική λύση δεδομένων, τα οποία χαρακτηρίζονται από το μεγάλο τους μέγεθος, την μεταξύ τους ανομοιογένεια, τη συχνή ανανέωση όχι μόνο των δεδομένων καθεαυτών, αλλά και τη μεταβολή των τύπων και των πεδίων τους. Η χρήση σχεσιακών συστημάτων διαχείρισης βάσεων δεδομένων για την επεξεργασία τέτοιων δεδομένων απαιτεί εξαιρετικά πολύ χρόνο, ενώ σε ορισμένες περιπτώσεις αποτελεί μια διαδικασία ανέφικτη [55-57].

Η πιο διαδεδομένη NoSQL πλατφόρμα είναι η Apache Hadoop, η οποία ήταν και πρωτοπόρος του τομέα, ενώ στη συνέχεια αναπτύχθηκαν άλλες βάσεις με δυνατότητα διαχείρισης τεράστιου όγκου δεδομένων, όπως η Cassandra Apache, η CouchBase

[35] Fog Computing
[36] NoSQL Databases

και η MongoDB. Η χρήση δε, της τελευταίας στον τομέα της υγείας έχει επισκιάσει τις υπόλοιπες, οι οποίες ωστόσο εφαρμόζονται σε αρκετές περιπτώσεις [1].

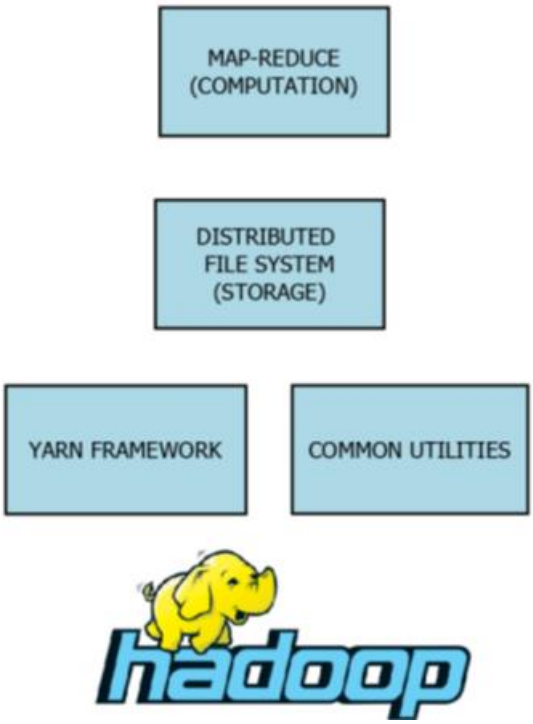
3.1.1 Apache Hadoop

Η φόρτωση μεγάλων ποσοτήτων (μεγάλων) δεδομένων στη μνήμη ακόμη και των πιο ισχυρών υπολογιστικών συστάδων δεν αποτελεί αποτελεσματικό τρόπο διαχείρισης των Big Data. Επομένως, η καλύτερη λογική προσέγγιση για την ανάλυση τεράστιων όγκων σύνθετων Big Data είναι η κατανομή και η επεξεργασία τους παράλληλα σε πολλαπλούς υπολογιστικούς κόμβους. Ωστόσο, το μέγεθος των δεδομένων είναι συνήθως τόσο μεγάλο, ώστε χιλιάδες υπολογιστικές μηχανές πρέπει να κατανέμουν μεταξύ τους και να ολοκληρώσουν την επεξεργασία σε εύλογο χρονικό διάστημα. Όταν κανείς εργάζεται με εκατοντάδες ή χιλιάδες κόμβους, πρέπει να χειριστεί ζητήματα όπως ο τρόπος επίτευξης της παραλληλίας, ήτοι της παράλληλης επεξεργασίας, η κατανομή των δεδομένων και η αντιμετώπιση των πιθανών αστοχιών. Μια από τις πιο δημοφιλείς εφαρμογές ανοιχτού κώδικα που προσφέρονται για το σκοπό αυτό είναι το Apache Hadoop [10, 58].

Ειδικότερα, το το Apache Hadoop έχει τη δυνατότητα επεξεργασίας εξαιρετικά μεγάλων ποσοτήτων δεδομένων, κυρίως με την κατανομή της επεξεργασίας μεγάλων συνόλων δεδομένων σε πολυάριθμους εξυπηρετητές (κόμβους), ο καθένας από τους οποίους επιλύει διαφορετικά τμήματα του συνολικού προβλήματος και στη συνέχεια τα ενσωματώνει για την εξαγωγή του τελικού αποτελέσματος. Οι υποστηριζόμενες από το Hadoop εφαρμογές δύνανται να αναλύσουν δεδομένα τεράστιου όγκου, τα οποία μπορεί να είναι δομημένα ή και αδόμητα [1].

Το Hadoop εφαρμόζει τον αλγόριθμο MapReduce, το οποίο είναι ένα μοντέλο προγραμματισμού για την επεξεργασία και παραγωγή μεγάλων συνόλων δεδομένων. Το MapReduce αποτελείται από δυο συναρτήσεις, τη Map και τη Reduce. Η Map είναι συνάρτηση απεικόνισης και αποστολή της είναι το φιλτράρισμα και η διάταξη των δεδομένων στους επιμέρους κόμβους. Η Reduce αποτελεί συνάρτηση μείωσης, που επί της ουσίας είναι επιφορτισμένη με την πραγματοποίηση των υπολογισμών καταμέτρησης, που λαμβάνουν χώρα στους επιμέρους κόμβους. Επιπρόσθετα, παραλληλίζει τις αστοχίες των υπολογιστικών χειρισμών και προγραμματίζει την επικοινωνία μεταξύ των υπολογιστικών συστάδων μεγάλης κλίμακας. Το Σύστημα

Κατανομής Αρχείων Hadoop (HDFS[37]), που αποτελεί το τμήμα αποθήκευσης των κατανεμημένων δεδομένων, είναι το στοιχείο του συστήματος που παρέχει μια κλιμακούμενη και αποδοτική αποθήκευση δεδομένων σε διάφορους κόμβους που αποτελούν μέρος μιας συστάδας [10, 58].



Εικόνα 4. Η Δομή της πλατφόρμας Hadoop

Το Hadoop έχει άλλα εργαλεία που ενισχύουν τα τμήματα αποθήκευσης και επεξεργασίας και επομένως πολλές μεγάλες εταιρείες όπως το Yahoo, το Facebook και άλλοι έσπευσαν να το υιοθετήσουν. Το Hadoop επιτρέπει στους ερευνητές να χρησιμοποιήσουν σύνολα δεδομένων, τα οποία δεν θα θα μπορούσαν να χειριστούν με άλλο τρόπο. Πολλές μεγάλες εργασίες, όπως ο καθορισμός μιας πιθανής συσχέτισης μεταξύ των δεδομένων που αφορούν στην ποιότητα του αέρα και στις νοσοκομειακές εισαγωγές λόγω άσθματος, η ανάπτυξη φαρμάκων με τη χρήση γονιδιωματικών και πρωτεομικών δεδομένων και άλλες τέτοιες πτυχές της υγειονομικής περίθαλψης πραγματοποιούνται με τη χρήση της πλατφόρμας Hadoop [10, 44].

[37] Hadoop Distributed File System

3.1.2 MongoDB

Η εταιρεία 10Gen έχει αναπτύξει και διαθέτει στην αγορά μια εκ των πλέον διαδεδομένων βάσεων δεδομένων, ήτοι την MongoDB, που είναι επίσης ανοικτού κώδικα και βασίζεται σε έγγραφα[38]. Η MongoDB βασίζεται στο μορφότυπο JSON[39] για την αποθήκευση αρχείων. Αποτελείται από ανοικτό format, αναγνώσιμο τόσο από τον άνθρωπο, όσο και από την μηχανή, γεγονός που διευκολύνει την ανταλλαγή δεδομένων σε σύγκριση με τα κλασσικά format όπως οι σειρές και οι πίνακες. Επιπλέον, η κλιμάκωση του μορφότυπου JSON είναι ευκολότερη, δεδομένου ότι δεν είναι απαραίτητη η υποβολή ξεχωριστών ερωτημάτων, λόγω του γεγονότος ότι τα σχετικά δεδομένα μιας δεδομένης εγγραφής περιέχονται σε ένα ενιαίο έγγραφο JSON. Μερικά από τα πλεονεκτήματα της περιλαμβάνουν την αυτόματη κλιμάκωση, την ευελιξία, την ευκολία στη χρήση, την παροχή υψηλών επιδόσεων και τη διαθεσιμότητα. Επιπλέον, παρέχει στο χρήστη τη δυνατότητα να πραγματοποιήσει αναζήτηση σε κείμενα, ενώ είναι δυνατή η σύνδεση του και με το Apache Hadoop [44].

Όσον αφορά το ρόλο της MongoDB στην υγεία, αυτή παρέχει μια σειρά λύσεων σε σημαντικά ζητήματα. Επί παραδείγματι, επιτρέπει στο χειριστή να δημιουργήσει ένα πλήρες προφίλ για τον εκάστοτε ασθενή, το οποίο περιλαμβάνει όλες τις διαγνωστικές εξετάσεις, στις οποίες έχει υποβληθεί, καθώς επίσης και να εξάγει συμπεράσματα για πιθανούς συσχετισμούς μέσω της αξιοποίησης των διαθέσιμων τεχνικών εξόρυξης δεδομένων. Επιπλέον αφενός μπορεί να τροποποιήσει ή και να προσθέσει νέα δεδομένα, επικαιροποιώντας το προφίλ και αφετέρου να τροποποιήσει συγκρίσεις σε βάθος χρόνου. Μια άλλη σημαντική εφαρμογή της MongoDB αφορά τον εντοπισμό και διάγνωση των σπανίων νοσημάτων σε πολύ πρώιμο στάδιο, εφόσον παρέχει τη δυαντότητα συσχετισμού στοιχείων, που ένας κλινικός ιατρός δεν θα μπορούσε να κάνει, ανεξάρτητα από το μέγεθος της εμπειρίας του. Περαιτέρω, καθιστά εφικτό τον περιορισμό της εξάπλωσης μιας επιδημικής έξαρσης, εν τη γενέσει της. Τέλος, είναι σε θέση να προχωρήσει άμεσα στη γνωμάτευση μιας ασθένειας σε πραγματικό χρόνο [44, 56-57].

[38] Document database
[39] Java Script Object Notation

3.1.3 Άλλες NoSQL βάσεων δεδομένων

Μια άλλη NoSQL βάση δεδομένων είναι η CouchBase, η οποία κάνει χρήση ενός ευέλικτου μοντέλου JSON, το οποίο επιτρέπει στο χρήστη τη διαμόρφωση των εφαρμογών του, χωρίς να περιορίζεται από ένα συγκεκριμένο σχήμα βάσης δεδομένων. Επιπλέον, τα πλεονεκτήματα της περιλαμβάνουν την ανάγνωση και εγγραφή με υψηλές ταχύτητες και την ευκολία κλιμάκωσης. Παράληλα, με τη χρήση της CouchBase είναι δυνατή η εξυπηρέτηση πολλών χρηστών ταυτόχρονα, με την ίση κατανομή του φόρτου και των δεδομένων στο σύνολο των διαθέσιμων εξυπηρετητών [59].

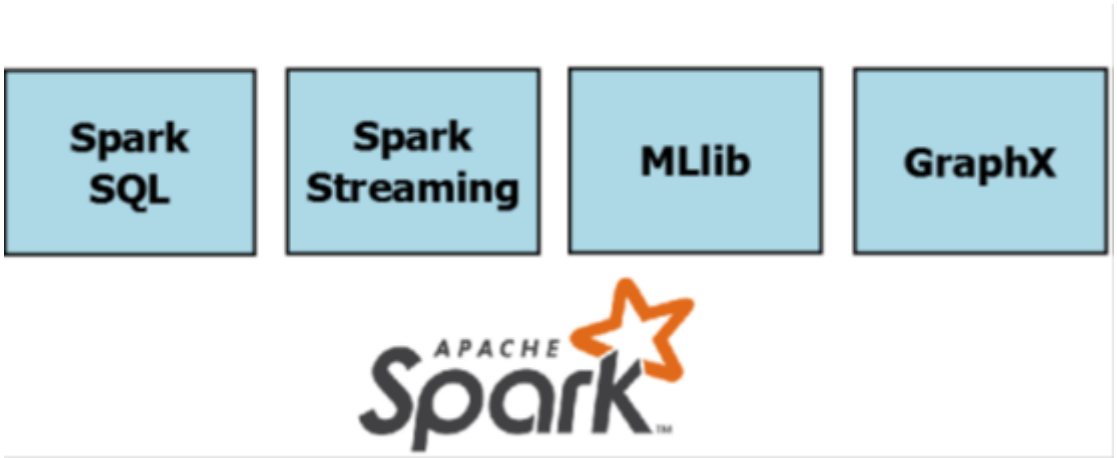
Επιπλέον, η εταιρεία η DataStax έχει αναπτύξει τη βάση δεδομένων Apache Cassandra, η οποία είναι ανοικτού λογισμικού και προωθείται ως το αντίπαλο δέος της βάσης δεδομένων Oracle. Καθιστά εφικτή τη διαχείριση μεγάλου όγκου μη δομημένων, ημι-δομημένων και δομημένων δεδομένων, τα οποία είναι αποθηκευμένα σε μια σειρά διαφορετικών υπολογιστικών κέντρων ή και στο νέφος. Τέλος, είναι δυνατό να συνδυαστεί με το Hadoop [60].

3.2 Apache Spark

Το Apache Spark αποτελεί άλλη μια εναλλακτική λύση για το Hadoop, ενώ αποτελεί λογισμικό ανοικτού κώδικα. Είναι μια ενοποιημένη μηχανή για την επεξεργασία κατανεμημένων δεδομένων που περιλαμβάνει βιβλιοθήκες υψηλότερου επιπέδου για την υποστήριξη χαρακτηριστικών όπως η υποβολή ερωτημάτων[40] σε δεδομένα μέσω της χρησιμοποίησης της γλώσσας SQL (*Spark SQL*), η επεξεργασία δεδομένων ροής[41] (*Spark Streaming*), τη μηχανική μάθηση (*MLlib*) και την επεξεργασία γραφημάτων (*GraphX*) [10, 61]. Αυτές οι βιβλιοθήκες βοηθούν στην αύξηση της παραγωγικότητας των προγραμματιστών, επειδή η διεπαφή προγραμματισμού απαιτεί λιγότερες προσπάθειες κωδικοποίησης και μπορεί να συνδυαστεί με απλό τρόπο για να δημιουργηθούν περισσότεροι τύποι πολύπλοκων υπολογισμών. Με την εφαρμογή του λεγόμενου Ελαστικού Κατανεμημένου Συνόλου Δεδομένων (RDD[42]) υποστηρίζεται η επεξεργασία δεδομένων εντός της μνήμης που

[40] Queries
[41] Streaming data
[42] Resilient Distributed Dataset

μπορεί να κάνει το Spark περίπου 100 φορές ταχύτερο από το Hadoop [10, 62]. Αυτό ισχύει περισσότερο όταν το μέγεθος των δεδομένων είναι μικρότερο από τη διαθέσιμη μνήμη. Αυτό καταδεικνύει ότι η επεξεργασία Big Data με το Apache Spark θα απαιτούσε μεγάλη μνήμη. Δεδομένου ότι το κόστος της μνήμης είναι υψηλότερο, το MapReduce αναμένεται να είναι οικονομικότερο για μεγάλα σύνολα δεδομένων σε σύγκριση με το Apache Spark. Παρομοίως, η εφαρμογή Apache Storm αναπτύχθηκε για να παρέχει ένα πλαίσιο πραγματικού χρόνου για την επεξεργασία δεδομένων ροής. Αυτή η πλατφόρμα υποστηρίζει τις περισσότερες γλώσσες προγραμματισμού. Επιπλέον, προσφέρει καλή οριζόντια δυνατότητα κλιμάκωσης και ενσωματωμένη δυνατότητα ανοχής σφαλμάτων, που καθιστούν δυνατή την ανάλυση των Big Data [10,44].



Εικόνα 5. Τα δομικά στοιχεία του Apache Spark

3.3 Εργαλεία Ανάλυσης Εικόνας

Οι τεχνικές απεικόνισης στην υγεία περιλαμβάνουν ιατρικές εικόνες υψηλής ευκρίνειας (δεδομένα ασθενών) μεγάλων μεγεθών. Οι επαγγελματίες του τομέα της υγείας όπως οι ακτινολόγοι και οι ιατροί κάνουν εξαιρετική δουλειά στην ανάλυση ιατρικών δεδομένων με τη μορφή αυτών των αρχείων για στοχευμένες ανωμαλίες. Ωστόσο, είναι επίσης σημαντικό να αναγνωριστεί η έλλειψη εξειδικευμένων επαγγελματιών για τη διάγνωση πολλών ασθενειών. Προκειμένου να αντισταθμιστεί αυτή η έλλειψη επαγγελματιών, έχουν αναπτυχθεί αποδοτικά συστήματα όπως το

Σύστημα Διαχείρισης Ιατρικών εικόνων (PACS[43]) για την αποθήκευση και εύκολη πρόσβαση σε δεδομένα ιατρικών εικόνων και αναφορών [10].

Ωστόσο, η ανταλλαγή δεδομένων με ένα PACS βασίζεται στη χρήση δομημένων δεδομένων για την ανάκτηση των ιατρικών εικόνων και συνεπώς δεν λαμβάνει υπόψη ορισμένες από τις αδόμητες πληροφορίες που περιέχονται σε ορισμένες από τις βιοϊατρικές εικόνες. Επιπλέον, είναι πιθανό να μην παρατηρήσει κανείς πρόσθετες πληροφορίες που υπάρχουν σε αυτές τις εικόνες ή παρόμοια δεδομένα και αφορούν την κατάσταση υγείας του ασθενούς. Έτσι, ένας επαγγελματίας που επικεντρώνεται στη διάγνωση μιας συγκεκριμένης, άσχετης ασθένειας μπορεί να αγνοήσει αυτές τις πρόσθετες πληροφορίες, που καταδεικνύουν την ύπαρξη μιας άλλης ασθένειας. Για την αποφυγή τέτοιων καταστάσεων, οι εφαρμογές ανάλυσης της εικόνας επηρεάζουν αποσκοπούν στην ενεργό εξαγωγή βιολογικών δεικτών από βιοϊατρικές εικόνες. Αυτή η προσέγγιση χρησιμοποιεί τεχνικές μηχανικής μάθησης και αναγνώρισης μοτίβων για να αντλήσει στοιχεία από τεράστιους όγκους δεδομένων κλινικών εικόνων για τη διάνγωσης, θεραπεία και παρακολούθηση των ασθενών. Επικεντρώνεται δε, στην ενίσχυση της διαγνωστικής χρησιμότητας της ιατρικής απεικόνισης για τη λήψη κλινικών αποφάσεων [10].

Προς αυτή την κατεύθυνση έχουν αναπτυχθεί διάφορα εργαλεία λογισμικού. Για παράδειγμα, το Visualization Toolkit είναι ένα ελεύθερα διαθέσιμο λογισμικό που επιτρέπει ισχυρή επεξεργασία και ανάλυση 3D εικόνων από ιατρικές εξετάσεις, ενώ η SPM μπορεί να επεξεργαστεί και να αναλύσει 5 διαφορετικούς τύπους εγκεφαλικών εικόνων. Άλλα λογισμικά όπως το GIMIAS, το Elastix και το MITK υποστηρίζουν όλους τους τύπους εικόνων. Τέτοιες αναλύσεις δεδομένων βασισμένες μπορούν να εξάγουν σημαντικότερα συμπεράσματα και αξία από τα απεικονιστικά δεδομένα, προκειμένου να ενισχύσουν και να υποστηρίξουν προγράμματα ιατρικής ακρίβειας, εργαλεία υποστήριξης κλινικών αποφάσεων και άλλα μέσα υγειονομικής περίθαλψης. Επί παραδείγματι αυτά τα εργαλεία μπορούν να χρησιμοποιηθούν για την παρακολούθηση νέων στοχευμένων θεραπειών για τον καρκίνο [10].

[43] Picture Archiving and Communication System

3.4 Εργαλεία Ανάλυσης Big data από Βιολογικά Δεδομένα

Τα big data από τις μελέτες βιολογικών δεδομένων είναι ένα νέο είδος πρόκλησης για τους βιοπληροφορικούς. Απαιτούνται ισχυροί αλγόριθμοι για την ανάλυση τέτοιων περίπλοκων δεδομένων από βιολογικά συστήματα. Ο απώτερος στόχος είναι η μετατροπή αυτών των τεράστιων όγκων δεδομένων σε μια πολύτιμη βάση γνώσεων. Η εφαρμογή προσεγγίσεων βιοπληροφορικής για τη μετατροπή των δεδομένων βιοϊατρικής και γονιδιωματικής σε προσεγγίσεις για την προγνωστική και προληπτική υγεία είναι γνωστή ως Μεταφραστική Βιοπληροφορική[44]. Βρίσκεται στην πρώτη γραμμή της βασισμένης σε δεδομένα υγειονομικής περίθαλψης. Διάφορα είδη ποσοτικών δεδομένων στην υγειονομική περίθαλψη, για παράδειγμα εργαστηριακές μετρήσεις, δεδομένα φαρμάκων και γονιδιωματικά προφίλ, μπορούν να συνδυαστούν και να χρησιμοποιηθούν για τον εντοπισμό νέων μεταδεδομένων που μπορούν να βοηθήσουν τις θεραπείες ακρίβειας [63]. Αυτός είναι ο λόγος για τον οποίο αναδυόμενες νέες τεχνολογίες είναι απαραίτητες για να βοηθήσουν στην ανάλυση αυτού του ψηφιακού πλούτου [10]. Ορισμένα εργαλεία ανάλυσης Big Data από βιολογικά δεδομένα είναι τα ακόλουθα:

1. **SparkSeq:** είναι μια αποδοτική πλατφόρμα βασισμένη στο πλαίσιο Apache Spark και στη βιβλιοθήκη Hadoop που χρησιμοποιείται για αναλύσεις γονιδιωματικών δεδομένων για ανάλυση αλληλεπιδραστικών γονιδιωματικών δεδομένων με ακρίβεια νουκλεοτιδίων.
2. **SAMQA:** εντοπίζει σφάλματα και εξασφαλίζει την ποιότητα των γενομικών δεδομένων μεγάλης κλίμακας.
3. **ART:** μπορεί να προσομοιώνει τα προφίλ των σφαλμάτων ανάγνωσης και τα μήκη ανάγνωσης για τα δεδομένα που λαμβάνονται χρησιμοποιώντας πλατφόρμες προσδιορισμού αλληλουχίας υψηλής απόδοσης, συμπεριλαμβανομένων των πλατφορμών SOLiD και Illumina.
4. **DistMap:** είναι ένα άλλο εργαλείο που χρησιμοποιείται για κατανεμημένη χαρτογράφηση μικρού μήκους με βάση τις υπολογιστικές συστάδες του Hadoop, που στοχεύει να καλύψει ένα ευρύτερο φάσμα εφαρμογών αλληλουχίας. Για παράδειγμα, μία από τις εφαρμογές της είναι ο χαρτογράφος BWA, ο οποίος μπορεί να εκτελέσει 500 εκατομμύρια ζεύγη ανάγνωσης σε

[44] Translational Bioinformatics

περίπου 6 ώρες, περίπου 13 φορές ταχύτερα απ' ό,τι ένας συμβατικός χαρτογράφος με ένα μόνο κόμβο.

5. **SeqWare**: είναι ένας μηχανισμός υποβολής ερωτημάτων βασισμένος στο σύστημα βάσης δεδομένων Apache HBase που επιτρέπει την πρόσβαση για μεγάλης κλίμακας δεδομένων σε ολόκληρο το εύρος του γονιδιώματος με την ενσωμάτωση φυλλομετρητών γονιδιώματος και εργαλείων.
6. **CloudBurst:** είναι ένα υπολογιστικό μοντέλο παραλληλίας που χρησιμοποιείται σε πειράματα χαρτογράφησης του γονιδιώματος για τη βελτίωση της κλιμάλωσης της ανάγνωσης μεγάλων δεδομένων αλληλουχίας.
7. **BlueSNP:** είναι ένα πακέτο R βασισμένο στο Hadoop για την ανάλυση μελετών συσχέτισης ολόκληρου του γονιδιώματος (GWAS[45]), με κύριο στόχο τη στατιστική ανάλυση, ώστε να επιτευχθούν σημαντικές συσχετίσεις μεταξύ των συνόλων δεδομένων γονότυπου-φαινοτύπου.
8. **Myrna:** που βασίζεται στην υπολογιστική νέφους, παρέχει πληροφορίες σχετικά με τις διαφορές επιπέδων έκφρασης των γονιδίων, και περιλαμβάνει μεταξύ άλλων την κανονικοποίησης των δεδομένων και τη στατιστική μοντελοποίηση [10].

3.5 Εμπορικές Πλατφόρμες Data Analytics της Υγείας

Προκειμένου να αντιμετωπιστούν οι μεγάλες προκλήσεις στον τομέα των δεδομένων και να πραγματοποιηθούν πιο αξιόπιστες αναλύσεις, διάφορες εταιρείες έχουν εφαρμόσει τεχνητή νοημοσύνη για την ανάλυση των δημοσιευμένων αποτελεσμάτων, των κειμένων δεδομένων και των δεδομένων εικόνας για την επίτευξη σημαντικών αποτελεσμάτων. Η IBM Corporation είναι ένας από τους μεγαλύτερους και εμπειρότερους παίκτες στον τομέα της παροχής υπηρεσιών εμπορικής ανάλυσης στην υγεία. Η Watson Health της IBM είναι μια πλατφόρμα τεχνητής νοημοσύνης για την ανταλλαγή και την ανάλυση δεδομένων υγείας μεταξύ νοσοκομείων, παρόχων και ερευνητών. Παρομοίως, η Flatiron Health παρέχει υπηρεσίες τεχνολογίας προσανατολισμένες σε αναλύσεις περίθαλψης ειδικά εστιασμένες στην έρευνα για τον καρκίνο. Άλλες μεγάλες εταιρείες όπως η Oracle Corporation και η Google Inc. εστιάζουν επίσης στην ανάπτυξη πλατφόρμων

[45] Genome-Wide Association Studies

αποθήκευσης υπολογιστών που βασίζονται σε υπλογιστική νέφους και πλατφόρμες κατανεμημένης υπολογιστικής ισχύος. Είναι ενδιαφέρον ότι τα τελευταία χρόνια, πολλές εταιρείες και νεοσύστατες επιχειρήσεις έχουν επίσης αναπτύξει δραστηριότητες που αποσκοπούν να παρέχουν αναλύσεις και λύσεις βασισμένες στην υγειονομική περίθαλψη [10].

3.5.1 AYASDI

Η εταιρεία Ayasdi αποτελεί ένα σημαντικό προμηθευτή που επικεντρώνεται στις μεθοδολογίες που βασίζονται στην εκμάθηση των μηχανών για την παροχή μιας πλατφόρμας μηχανικής νοημοσύνης μαζί με ένα πλαίσιο εφαρμογής με δοκιμασμένη επιχειρησιακή κλιμάκωση. Παρέχει διάφορες εφαρμογές για αναλύσεις στην υγεία, για παράδειγμα, παρέχει εφαρμογή για την κατανόηση και τη διαχείριση των κλινικών διαφοροποιήσεων και για την αντίστοιχη μετατροπή του κόστους κλινικής περίθαλψης. Είναι επίσης σε θέση να αναλύσει και να διαχειριστεί τον τρόπο με τον οποίο οργανώνονται τα νοσοκομεία, τις συνομιλίες μεταξύ των γιατρών, τις αποφάσεις που λαμβάνονται από τους γιατρούς με γνώμονα τον κίνδυνο και τη φροντίδα που παρέχουν στους ασθενείς. Παρέχει επίσης μια εφαρμογή αξιολόγησης και διαχείρισης της υγείας του πληθυσμού, μια προληπτική στρατηγική που υπερβαίνει τις παραδοσιακές μεθοδολογίες ανάλυσης κινδύνου [10, 64].

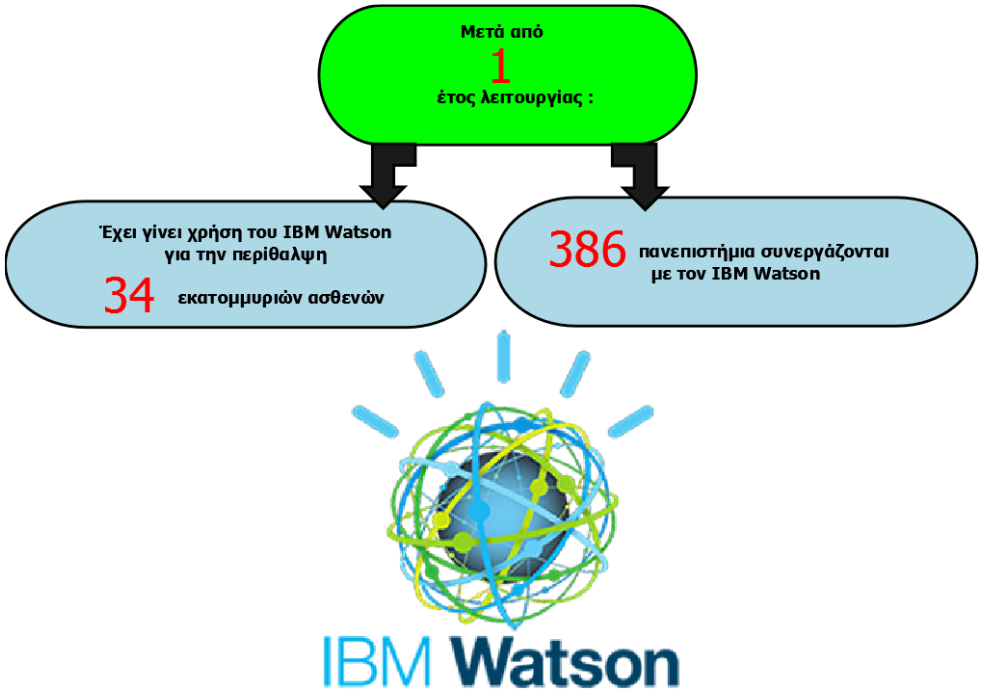
3.5.2 Linguamatics

Πρόκειται για ένα αλγόριθμο βασισμένο στην επεξεργασία φυσικής γλώσσας (NLP[46]) που βασίζεται σε ένα αλγόριθμο εξόρυξης γνώσης από διαδραστικό κείμενο (I2E[47]). Το I2E μπορεί να εξαγάγει και να αναλύσει μια μεγάλη ποικιλία πληροφοριών. Τα αποτελέσματα που λαμβάνονται χρησιμοποιώντας αυτή την τεχνική είναι δεκαπλάσια από άλλα εργαλεία και δεν απαιτούν ειδικές γνώσεις για την ερμηνεία των δεδομένων. Αυτή η προσέγγιση μπορεί να παρέχει πληροφορίες σχετικά με γενετικές σχέσεις και γεγονότα από μη δομημένα δεδομένα. Ως γνωστόν, η εκμάθηση απαιτεί απαιτεί ποιοτικά δεδομένα για την παραγωγή καθαρών και φιλτραρισμένων αποτελεσμάτων. Ωστόσο, όταν το NLP ενσωματώνεται στον ΗΦΥ ή σε άλλα κλινικά αρχεία καθ'εαυτό, διευκολύνει την εξαγωγή καθαρών και δομημένων πληροφοριών που συχνά παραμένουν κρυμμένες σε μη δομημένα δεδομένα εισόδου [10].

[46] Natural Language Processing
[47] Interactive text mining algorithm

3.5.3 IBM Watson

Αυτή είναι μια από τις μοναδικές τεχνολογικές εφαρμογές της εταιρείας IBM που στοχεύει σε αναλύσεις Big Data σε σχεδόν κάθε επαγγελματικό τομέα. Αυτή η πλατφόρμα χρησιμοποιεί εκτεταμένα αλγορίθμους βασισμένους σε εκμάθηση μηχανών και τεχνητή νοημοσύνη, ώστε να εξαγάγει τις μέγιστες πληροφορίες από ελάχιστες εισροές. Ο IBM Watson είναι ένας υπερυπολογιστής, που επιβάλλει το σχήμα ενσωμάτωσης ενός ευρέος φάσματος τομέων της υγείας για την παροχή ουσιαστικών και δομημένων δεδομένων. Σε μια προσπάθεια να αποκαλυφθούν νέα φάρμακα για συγκεκριμένα μοντέλα καρκινικών νοσημάτων, ο IBM Watson και η Pfizer έχουν διαμορφώσει μια παραγωγική συνεργασία για την επιτάχυνση της ανακάλυψης νέων συνδυασμών ανοσο-ογκολογίας [10].



Εικόνα 6. Διάδοση IBM Watson εντός ενός έτους λειτουργίας

Συμπεράσματα

Το Big Data, και δη η ανάλυση αυτών, καθιστούν εφικτή την αξιοποίηση των μυριάδων ανόμοιων, δομημένων και μη δεδομένων πηγών δεδομένων, και αναμφίβολα θα διαδραματίσουν ζωτικό ρόλο στον τρόπο, με τον οποίο παρέχεται η υγειονομική περίθαλψη στο μέλλον. Μπορεί κανείς να δει ήδη ένα φάσμα των δυνατοτήτων της αναλυτικής που χρησιμοποιούνται, βοηθώντας στη λήψη αποφάσεων και στη βελτίωση της απόδοσης του προσωπικού της υγειονομικής περίθαλψης και της εξέλιξης της πορείας της υγείας του ασθενούς. Τα Big Data πέραν των σημαντικών πλεονεκτημάτων που ενέχουν για τον τομέα της υγείας, αντιμετωπίζουν μια σειρά προκλήσεων όπως η προστασία του ιδιωτικού απορρήτου, η ακρίβεια και αξιοπιστία τους κλπ. Περαιτέρω, θα πρέπει να σημειωθεί ότι τα ηλεκτρονικά δεδομένα για την υγεία παραμένουν σε μεγάλο βαθμό αναξιοποίητα, και κατά συνέπεια είναι επιτακτική η ανάγκη μετατροπής των με επεξεργασμένων δεδομένων σε χρήσιμες και αποτελεσματικές πληροφορίες.

Εκτός από την προφανή ανάγκη για περαιτέρω έρευνα στον τομέα της ομογενοποίησης και της εναρμόνισης των συνεχόμενων και διακριτών μορφών ιατρικών δεδομένων, υπάρχει εξίσου σημαντική ανάγκη για την ανάπτυξη νέων τεχνικών επεξεργασίας τους. Οι έρευνες που σχετίζονται με την εξόρυξη βιολογικών δεικτών και «λαθραίων» μοτίβων για την κατανόηση και την πρόβλεψη κρουσμάτων ασθενειών έχουν παρουσιάσει μια αξιοσημείωτη δυναμική για την παροχή πληροφοριών που μπορούν να μετουσιωθούν στην ανάληψη μιας ενέργειας. Ωστόσο, υπάρχουν ευκαιρίες για την ανάπτυξη αλγορίθμων για την αντιμετώπιση του φιλτραρίσματος δεδομένων, της παρεμβολής, του μετασχηματισμού, της εξαγωγής χαρακτηριστικών, της επιλογής χαρακτηριστικών και ούτω καθεξής. Επιπλέον, με την αξιοπιστία και τη βελτίωση των αλγορίθμων μηχανικής μάθησης, υπάρχουν ευκαιρίες για τη βελτίωση και ανάπτυξη ισχυρών CDSS για την κλινική πρόβλεψη, συνταγογράφηση και διάγνωση.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Garapati SL, Garapati S. Application of Big Data Analytics: An Innovation in Health Care. *International Journal of Computational Intelligence Research* 2018; 14(1)15-27.
2. Shafqat S, Kishwer S, Rasool R, Junaid Q, Tehmina A, Hafiz FA. Big Data Analytics enhanced healthcare systems: a review. *J Supercomput* 2018. Διαθέσιμο στο: https://doi.org/10.1007/s11227-017-2222-4 [Ανάκτηση την 12.05.2019].
3. Raghupathi W, Raghupathi V. Big Data Analytics in healthcare: promise and potential. *Health Inf Sci Syst.* 2014;2:3. doi:10.1186/2047-2501-2-3.
4. Ahmad Z, Tripathi MM. Approaches of Big Data in Healthcare: A Critical Review. *International Journal of Advanced Research in Computer Science* 2018; 9(2): 122-127.
5. Fernandes L, O'Connor M, Weaver V. Big Data, bigger outcomes: Healthcare is embracing the Big Data movement, hoping to revolutionize HIM by distilling vast collection of data for specific analysis. *J AHIMA.* 2012; 83(10):38-43.
6. Fernald GH, Capriotti E, Daneshjou R, Karczewski KJ, Altman RB. Bioinformatics challenges for personalized medicine. *Bioinformatics* 2011; 27(13): 1741–1748.
7. Mehta N, Pandit A. Concurrence of Big Data Analytics and healthcare: A systematic review. *International Journal of Medical Informatics* 2018; 114:57–65.
8. Manyika J, Chui M, Brown B, Dobbs R, Roxburgh R., et al. Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute. 2011. Διαθέσιμο στο: https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Big%20data%20The%20next%20frontier%20for%20innovation/MGI_big_data_exec_summary.ashx [Ανάκτηση την 19.08.2019].
9. Belle A, & Thiagarajan R, Soroushmehr R, Navidi F, Beard D, Najarian K. Big Data Analytics in Healthcare. *BioMed Research International* 2015.

Διαθέσιμο στο: https://www.researchgate.net/publication/279198958_Big_Data_Analytics_in_Healthcare [Ανάκτηση την 12.05.2019].

10. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big Data in healthcare: management, analysis and future prospects. *J Big Data* 2019; 6:54. Διαθέσιμο στο: https://doi.org/10.1186/s40537-019-0217-0 [Ανάκτηση την 22.06.2019].

11. Sonnati R. Improving Healthcare Using Big Data Analytics. *International Journal of Scientific & Technology Research* 2017; 6(03):142-146.

12. Buneman P. Semistructured Data. 1997. Διαθέσιμο στο: https://www.csd.uoc.gr/~hy561/Data/Papers/tutorial-semi-pods97.pdf [Ανάκτηση την 14.05.2019].

13. Borckardt JJ, Nash MR, Murphy MD, Moore M, Shaw D, O'Neil P. Clinical practice as natural laboratory for psychotherapy research: a guide to case-based time-series analysis. *The American Psychologist* 2018; 63(2):77-95.

14. Senthilkumar SA, Bharatendara K Rai, Amruta A Meshram, Angappa Gunasekaran, Chandrakumarmangalam S. Big Data in Healthcare Management: A Review of Literature. *American Journal of Theoretical and Applied Business* 2018; 4(2):57-69.

15. Wang L, Alexander CA. Big Data in Medical Applications and Health Care. *Am. Med. J.* 2015; 6:1–8. doi:10.3844/amjsp.2015.1.8

16. Sarwar MU, Hanify MK, Talibz R, Mobeenx A, Aslam M. A Survey of Big Data Analytics in Healthcare. *(IJACSA) International Journal of Advanced Computer Science and Applications* 2017; 8(6):355-359.

17. Berger ML, Doban V. Big Data, advanced analytics and the future of comparative effectiveness research. *Journal of Comparative Effectiveness Research* 2014; 167–176.

18. Institute for Health Technology Transformation. Transforming Health Care through Big Data Strategies for leveraging Big Data in the health care industry. 2013. Διαθέσιμο στο: http://c4fd63cb482ce6861463-bc6183f1c18e748a49b87a25911a0555.r93.cf2.rackcdn.com/iHT2_BigData_2013.pdf [Ανάκτηση την 15.04.2019].

19. Costa FF. Big Data in biomedicine. *Drug Discov. Today* 2014; 19:433–440. Διαθέσιμο στο: http://dx.doi.org/10.1016/j.drudis.2013.10.012 . [Ανάκτηση την 15.04.2019].

20. Bello-Orgaz G, Jung JJ, Camacho D. Social Big Data: Recent achievements and new challenges. *Inf. Fusion* 2015; 28:45-49. doi:10.1016/j. inffus.2015.08.005.

21. Davenport TH, Patil DJ. Data scientist: the sexiest job of the 21st century. *Harvard Business Review* 2012; 90(10):70-76.

22. He KY, Ge D, He MM. Big Data Analytics for genomic medicine. *Int. J. Mol. Sci.* 2017; 18:412.

23. Sukumar S, Natarajan R, Ferrell R. Quality of Big Data in health care. *International Journal of Health Care Quality Assurance*; 2015; 8(6): 621-634.

24. Palanisamy V, Thirunavukarasu R. Implications of big data analytics in developing healthcare frameworks – A review. *Journal of King Saud University – Computer and Information Sciences* 2019; 31:415–425.

25. Huang BE, Mulyasasmita W, Rajagopal G. The path from big data to precision medicine. *Expert Rev. Precis. Med. Drug Dev.* 2016; 1(3):129–143.

26. Auffray C, Balling R, Barroso I, Bencze L, Benson M, Bergeron J, et al. Making sense of big data in health research: towards an EU action plan, *Genome Med.* 2016; 8. Διαθέσιμο στο: https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-016-0323-y [Ανάκτηση την 12.06.2019].

27. Rumsfeld JS, Joynt KE, Maddox TM. Big data analytics to improve cardiovascular care: promise and challenges. *Nat. Rev. Cardiol.* 2016; 13: 350–359.

28. Swan M. The quantified self: fundamental disruption in big data science and biological discovery. *Big Data* 2013; 1(2):85-99. Διαθέσιμο στο: https://www.liebertpub.com/doi/pdf/10.1089/big.2012.0002 [Ανάκτηση την 13.04.2019].

29. McAfee A, Brynjolfsson E, Davenport TH, Patil DJ, Barton D. Big data: the management revolution. *Harvard Business Review* 2012; 90(10):60-68.

30. Fernald GH, Capriotti E, Daneshjou R, Karczewski KJ, Altman RB. Bioinformatics challenges for personalized medicine. *Bioinformatics* 2011; 227(13): 1741–1748.

31. Alexandru AG, Radu MI, Bizon ML. Big Data in Healthcare - Opportunities and Challenges. *Informatica Economică* 2018; 22(2):43-58.

32. Fatt QK, Ramadas A. The Usefulness and Challenges of Big Data in Healthcare. *Journal of Healthcare Communications* 2018; 3(2):21. Διαθέσιμο στο: http://healthcare-communications.imedpub.com/archive.php [Ανάκτηση την 13.07.2019].

33. Basco JA, Senthilkumar NC. Real-time analysis of healthcare using big data analytics. IOP Conference Series: *Materials Science and Engineering* 2017; 263(4):42056. Διαθέσιμο στο: http://adsabs.harvard.edu/abs/2017MS%26E..263d2056B [Ανάκτηση την 18.05.2019].

34. Alexander CA, Wang L. Big data in healthcare: a new frontier in personalized medicine. *Open Access J Trans Med Res.* 2017; 1(1):15-18.

35. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. *Int J Med Inform.* 2008; 77:81–97.

36. Binder H, Blettner M. Big data in medical science--a biostatistical view. *Dtsch Arztebl Int*. 2015; 112:137–142.

37. Mohammed EA, Far BH, Naugler C. Applications of the MapReducE programming framework to clinical big data analysis: current landscape and future trends. *BioData Min.* 2014; 22. Διαθέσιμο στο: https://biodatamining.biomedcentral.com/articles/10.1186/1756-0381-7-22 [Ανάκτηση την 02.08.2019].

38. Valikodath NG, Newman-Casey PA, Lee PP, Musch DC, Niziol LM, Woodward MA. Agreement of Ocular Symptom Reporting Between Patient-Reported Outcomes and Medical Records. *JAMA Ophthalmol.* 2017; 135(3):225-231.

39. Mittelstadt BD, Floridi L. The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. *Sci Eng Ethics* 2015. Διαθέσιμο στο: https://www.researchgate.net/publication/277079050_The_Ethics_of_Big_Data_Current_and_Foreseeable_Issues_in_Biomedical_Contexts [Ανάκτηση την 14.07.2019].

40. Adler-Milstein J, Pfeifer E. Information blocking: is it occurring and what policy strategies can address it? *Milbank Q* 2017;95(1):117–35.

41. Dimitrov D. Medical Internet of Things and Big Data in Healthcare. *Healthcare Informatics Research* 2016; 22 (3):156. Διαθέσιμο στο: http://dx.doi.org/10.4258/hir.2016.22.3.156 [Ανάκτηση την 19.07.2019].

42. Frost & Sulllivan. U.S. Hospital Health Data Analytics Market: Growing EHR Adoption Fuels A New Era in Analytics. 2012. Διαθέσιμο στο: https://store.frost.com/u-s-hospital-health-data-analytics-market.html [Ανάκτηση την 14.09.2019].

43. Khalifa, M. Health Analytics Types, Functions and Levels: A Review of Literature. In Data, Informatics and Technology: An Inspiration for Improved Healthcare A. Hasman et al. (Eds.) *IOS Press* 2018; 137-140. doi:10.3233/978-1-61499-880-8-137.

44. El aboudi N, Benhlima L. Big Data Management for Healthcare Systems: Architecture, Requirements, and Implementation. *Advances in Bioinformatics 2018*. Article ID 4059018. 10 pages. Διαθέσιμο στο: https://doi.org/10.1155/2018/4059018. [Ανάκτηση την 05.07.2019].

45. LaValle S, Lesser E, Shockley R, Hopkins MS, Kruschwitz N. Big data, analytics and the path from insights to value. *MIT Sloan Management Review* 2011; 52(2):21-32. Διαθέσιμο στο: http://foresight.ifmo.ru/ict/shared/files/201309/1_9.pdf [Ανάκτηση την 20.07.2019].

46. Ristevski B, Chen M. Big Data Analytics in Medicine and Healthcare. *J Integr Bioinform*. 2018; 15(3):20170030. doi:10.1515/jib-2017-0030

47. Wu PY, Cheng CW, Kaddi CD, Venugopalan J, Hoffman R, Wang MD. Omic and Electronic Health Record Big Data Analytics for Precision Medicine. *IEEE Trans Biomed Eng* 2017; 64:263‑73.

48. Viceconti M, Hunter P, Hose R. Big data, big knowledge: big data for personalized healthcare. *IEEE J Biomed Health Inform* 2015; 19:1209–1215.

49. Chen H, Chiang RH, Storey VC. Business intelligence and analytics: From big data to big impact. *MIS Quarterly* 2012; 36(4):1165-1188. Διαθέσιμο στο: http://hmchen.shidler.hawaii.edu/Chen_big_data_MISQ_2012.pdf [Ανάκτηση την 20.07.2019].

50. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Affairs* 2014; 33(7):1123-1131.

51. Simpao AF, Ahumada LM, Gálvez JA, Rehman MA. A review of analytics and clinical informatics in health care. *Journal of medical systems* 2014; 38(4):45.

52. Cano I, Tenyi A, Vela E, Miralles F, Roca J. Perspectives on Big Data applications of health information. *Curr. Opin. Syst. Biol*. 2017; 3:36–42.

53. Asante-Korang A, Jacobs J. Big Data and paediatric cardiovascular disease in the era of transparency in healthcare. *Cardiology in the Young* 2016; 26(8): 1597-1602. doi:10.1017/S1047951116001736

54. Schultz T. Turning healthcare challenges into big data opportunities: a use-case review across the pharmaceutical development lifecycle. *Bull. Assoc. Inf. Sci. Technol.* 2013; 35(5):34-40. Διαθέσιμο στο: https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/bult.2013.1720390508 [Ανάκτηση την 29.07.2019].

55. Freire SM, Teodoro D, Wei-Kleiner F, Sundvall E, Karlsson D, Lambrix P. Comparing the Performance of NoSQL Approaches for Managing Archetype-Based Electronic Health Record Data. *PLoS One.* 2016;11(3):e0150069. Published 2016 Mar 9. doi:10.1371/journal.pone.0150069

56. Sreekanth R, Golajapu VM, Srinivas N. Big Data Electronic Health Records Data Management and Analysis on Cloud with MongoDB: A NoSQL Database. *International Journal of Advanced Engineering and Global Technology* 2015; 3(7)943-949.

57. Celesti A, Fazio M, Villari, M. A Study on Join Operations in MongoDB Preserving Collections Data Models for Future Internet Applications. *Future Internet* 2019; 11(83). doi:10.3390/fi11040083

58. Shvachko K, Kuang K, Radia S, Chansler R. The hadoop distributed file system. In: *Proceedings of the 2010 IEEE 26th symposium on mass storage systems and technologies (MSST).* New York: IEEE Computer Society; 2010: 1–10.

59. Sandwell D, Weston L. Designing High-Performance Data Structures for the Couchbase Data Platform. Erwin.inc, 2018. Διαθέσιμο στο: https://info.couchbase.com/rs/302-GJY-034/images/erwin_DM_High%20Perf_NoSQL-QOM-Technical_Whitepaper_Couchbase_Version.pdf [Ανάκτηση την 29.08.2019].

60. DataStax.com. Apache Cassandra, 2019. Διαθέσιμο στο: https://www.datastax.com/products/apache-cassandra [Ανάκτηση την 29.07.2019].

61. Zacharia M, Xin R, Wendell P, Das T, Armbrust M, Dave A, et al. Apache Spark: a unified engine for big data processing. *Commun ACM* 2016; 59(11):56–65.

62. Ahmed H, AliIsmail M, Hyder MF, Sheraz SM, Fouq N. Performance comparison of spark clusters configured conventionally and a cloud service. *Procedia Comput Sci.* 2016; 82:99–106.

63. Li L, Cheng WY, Glicksberg BS, Gottesman O, Tamler R, Chen R, et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci Transl Med.* 2015; 7(311):311ra174. doi: 10.1126/scitranslmed.aaa9364.

64. Ayasdi. Understanding Ayasdi: What we do, how we do it, why we do it. White Paper, 2018. Διαθέσιμο στο: https://s3.amazonaws.com/cdn.ayasdi.com/wp-content/uploads/2018/04/04142230/UnderstandingAyasdi_WP_061617v011.pdf [Ανάκτηση την 01.09.2019].