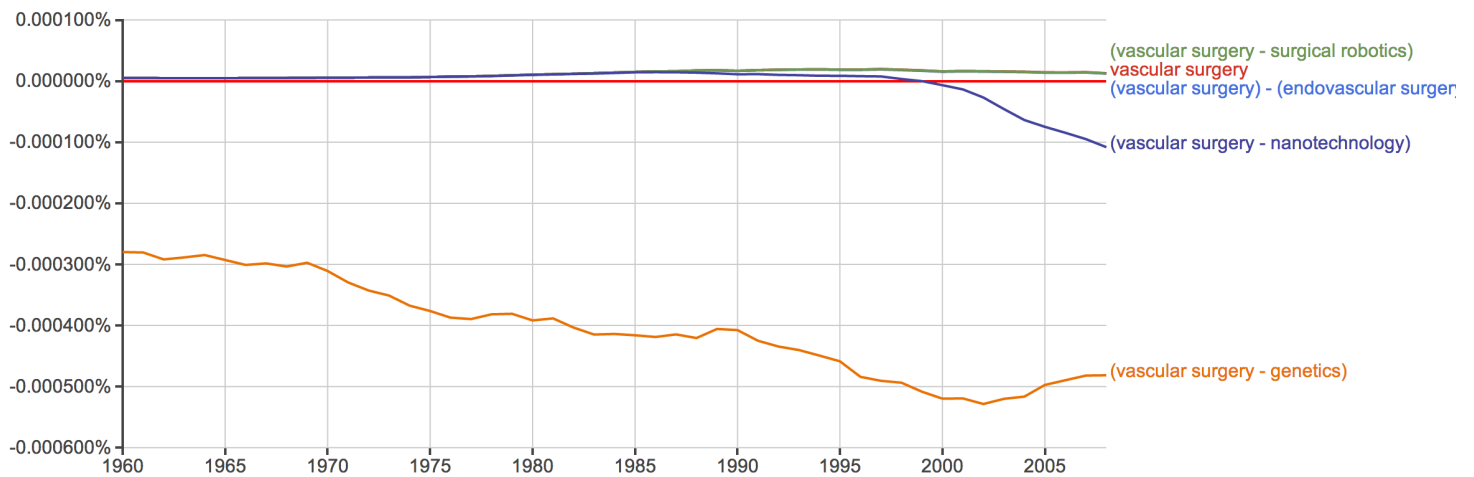


Master Thesis in Endovascular Techniques

Mining big data for knowledge extraction



Louizos-Alexandros Louizos
2014

Περίληψη

Τα τελευταία χρόνια παρατηρούμε την εκθετική αύξηση της γνώσης. Τα δεδομένα αφθονούν. Υπολογίζεται ότι υπάρχουν 2.7 Zetabytes από δεδομένα στο ψηφιακό σύμπαν. Δύο δεκαετίες πριν υπήρχε η ανάγκη για δημιουργία δεδομένων και ύπαρξη υπολογιστικής ισχύς. Σήμερα τα δεδομένα και η υπολογιστική ισχύς αφθονούν και έχουμε περάσει στην εποχή όπου η επεξεργασία αυτών και η ανάλυση της γνώσης που υπάρχει κρυμμένη σε αυτά είναι πιο σημαντική. Η επιστήμη ανάλυσης πολλών δεδομένων (big data science) δημιουργήθηκε για να εξυπηρετήσει αυτή την ανάγκη και έχει αποδώσει καρπούς σε περιοχές όπως η αστροφυσική, η επιστήμη υλικών και τα κοινωνικά δίκτυα. Σε αυτή την εργασία παρουσιάζουμε την ανάλυση 5 δισεκατομμυρίων λέξεων από όλα τα βιβλία που έχουν παραχθεί ποτέ με σκοπό να προσπαθήσουμε να βρούμε κρυμμένους μέσα στον όγκο των δεδομένων συσχετισμούς ως προς την πορεία της αγγειοχειρουργικής μέσα στο χρόνο καθώς και τι μπορεί να την επηρεάζει και τι μπορεί να επιφυλάσσει το μέλλον για την επιστήμη μας. Η ανάλυση γίνεται μεθοδικά με τη χρήση τεχνητής νοημοσύνης και τεχνικών επεξεργασίας φυσικής γλώσσας. Επιπλέον επιστρατεύτηκε παράλληλη επεξεργασία σε δίκτυο υπερυπολογιστών για την πραγματοποίηση της ανάλυσης. Τα αποτελέσματα παρουσιάζονται σε χρονογραφήματα των N-grams ως προς τη συχνότητα εμφάνισης της λέξης εντός του κειμένου.

ΕΙΣΑΓΩΓΗ

Ζούμε στην εποχή των πολλών δεδομένων. Υπάρχουν σήμερα 2.7 Zetabytes δεδομένων στον ψηφιακό κόσμο. 235 Terabytes έχουν συλλεχθεί από την βιβλιοθήκη του αμερικάνικου κογκρέσου τον Απρίλη του 2011. Το κοινωνικό δίκτυο facebook χειρίζεται τουλάχιστον 30 Petabytes δεδομένων από τους χρήστες του. Ο μεγάλος αδρονικός δακτύλιος παράγει κάθε έτος 15 Petabytes δεδομένων.

Είναι κατανοητό ότι τα δεδομένα είναι ο σύγχρονος χρυσός. Τα δεδομένα από μόνα τους είναι σαν ακατέργαστο πετρέλαιο. Αυτό που έχει πραγματική αξία είναι η επεξεργασία αυτών των δεδομένων. Για την επεξεργασία χρειάζονται νέες μέθοδοι. Δεν μπορεί κανείς να χρησιμοποιήσει SPSS ή EXCEL για να επεξεργαστεί πάνω από λίγα gigabyte από δεδομένα, ενώ στο χώρο των πολλών δεδομένων λίγα gigabyte δεν θεωρούνται καν πολλά.

Για την επεξεργασία όλων αυτών των τεραστίων δεδομένων χρειάζονται νέα εργαλεία και μέθοδοι. Αυτά τα εργαλεία έχουν δημιουργηθεί για τις ανάγκες επεξεργασίας των γιγάντων του διαδικτύου αρχικά από τη yahoo και μετά από τη google. Αυτό που έγινε κατανοητό το 2000 ήταν ότι δε γίνεται να δημιουργηθεί ένα και μοναδικό υπερμηχάνημα με τεράστια μνήμη και υπολογιστική ισχύ. Εκείνο που είναι όμως εφικτό ήταν η χρήση χαμηλού κόστους υπολογιστών εν παραλλήλω. Με τον τρόπο αυτό δημιουργούνται ομάδες παράλληλης επεξεργασίας από μερικές δεκάδες ως και χιλιάδες μηχανήματα. Η μέθοδος αυτή είναι φθηνότερη και ανεκτική σε λάθη καθώς δημιουργούνται πολλά αντίγραφα των δεδομένων. Αυτές οι σειρές υπολογιστών ονομάζονται GRID clusters.

Για την επεξεργασία των δεδομένων χρησιμοποιούνται ειδικά λογισμικά που είναι σε θέση να αναλύσουν αυτό τον όγκο εν παραλλήλω στα clusters. Το λογισμικό αυτό ονομάζεται Hadoop. Το hadoop αποτελεί ένα οικοσύστημα από λογισμικά που έχουν δημιουργηθεί για την επεξεργασία των δεδομένων αυτών ενώ απαιτείται ειδική γνώση για την περαιτέρω επεξεργασία αυτών πέρα από την αρχική οργάνωση που ονομάζεται map/reduce. Το map αποτελεί τη διαδικασία όπου τα δεδομένα οργανώνονται ανα τύπο ενώ στο reduce γίνεται η τελική πρόσθεση και οργάνωση όλων των δεδομένων στην τελική τους μορφή.

Από κει και έπειτα χρειάζονται τεχνικές machine learning για να μπορέσει κανείς να αναλύσει τα δεδομένα. Η μηχανική εκμάθηση αποτελεί κλάδο της τεχνητής νοημοσύνης και αφορά την κατασκευή και τη μελέτη συστημάτων που μπορούν να μάθουν από τα δεδομένα. Για παράδειγμα ένα τέτοιο σύστημα μπορεί να

εκπαιδευθεί ώστε να διαχωρίζει το ηλεκτρονικό ταχυδρομείο σε χρήσιμο και μη. Μετά τη διαδικασία εκπαίδευσης μπορεί να χρησιμοποιηθεί ώστε να διαχωρίζει και νέα μηνύματα ηλεκτρονικού ταχυδρομείου.

Η βασικές αρχές του machine learning είναι η αναπαράσταση και η γενίκευση. Η αναπαράσταση των δεδομένων γίνεται με ειδικές εξισώσεις και μεθόδους που πολλές φορές προσπαθούν να προσομοιάσουν τη λειτουργία νευρώνων. Η γενίκευση είναι η ιδιότητα του συστήματος να λειτουργεί καλά σε δεδομένα που δεν έχει ξανασυναντήσει. Αυτό αποτελεί και το αντικείμενο της επιστήμης της υπολογιστικής εκμάθησης.

Κλασικό παράδειγμα μηχανικής μάθησης αποτελεί η αναγνώριση φωνής ή η οπτική αναγνώριση κειμένου που συναντούμε συχνά σε σύγχρονα συστήματα.

Υπάρχει μια βασική διαφορά μεταξύ των μεθόδων που χρησιμοποιούνται στους τομείς της τεχνητής νοημοσύνης και της ανάλυσης των δεδομένων.

Η τεχνητή νοημοσύνη στοχεύει στην πρόβλεψη βασισμένη σε γνωστές ιδιότητες από τα δεδομένα στα οποία εκπαιδεύεται.

Η ανάλυση δεδομένων βασίζεται στην ανίχνευση αγνώστων ιδιοτήτων μέσα στα δεδομένα.

Οι δύο παραπάνω τομείς αλληλοεπικαλύπτονται με πολλούς τρόπους καθώς χρησιμοποιούν και οι δύο μεθόδους μηχανικής εκμάθησης αλλά διαφέρουν στον τελικό σκοπό. Η τεχνητή νοημοσύνη προσπαθεί να αναπαράγει γνωστό περιεχόμενο ενώ η ανάλυση δεδομένων προσπαθεί να παράγει νέα γνώση.

ΥΛΙΚΑ - ΜΕΘΟΔΟΙ

Για την αλίευση γνώσης χρησιμοποιήθηκε ψηφιοποιημένο υλικό από το Google Books Project το οποίο έχει ψηφιοποιήσει 5.2 εκατομμύρια βιβλία που έχουν δημοσιευθεί μεταξύ 1500 και 2008. Ο συνολικός αριθμός των λέξεων σε αυτά τα βιβλία είναι 500 δισεκατομμύρια.

Ο όγκος των αρχείων που χρησιμοποιήθηκαν ήταν περίπου 15 Gigabyte. Αρχικά χρησιμοποιήθηκε το Natural Language Processing Kit of Python για να διαχωριστεί το κείμενο σε μικρότερες μονάδες (tokens) που να αποτελούν κομμάτι λεξολογικά και να επεξεργαστούμε το κείμενο ώστε όλα τα γράμματα να είναι μικρά χωρίς διάκριση μεταξύ κεφαλαίων ή μικρών για την κάθε λέξη (normalization).

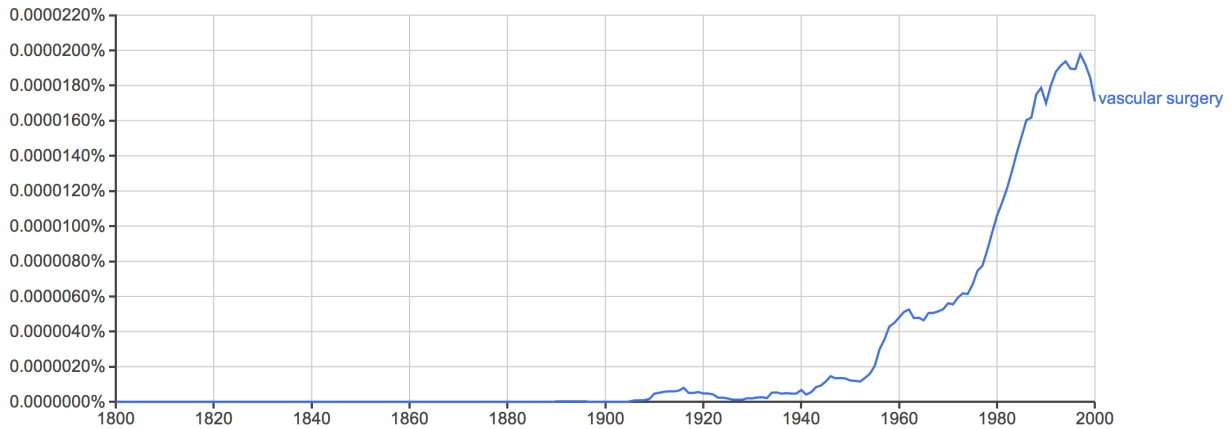
Για την επεξεργασία των αρχείων χρησιμοποιήθηκε ένα Hadoop cluster με 40 υπολογιστές και 320 GB μνήμη από την Amazon με συνολικό χρόνο επεξεργασίας τις 48 ώρες. Χρησιμοποιήθηκαν μόνο τα βιβλία που περιέχουν αγγλική γλώσσα. Αρχικά έλαβε χώρα map/reduce διαδικασία για όλους τους συνδυασμούς ενός και δύο λέξεων ώστε για κάθε λέξη ή δύο λέξεις (monogram and bi-gram) να δημιουργηθεί ο παρακάτω αλφαβητικός πίνακας για παράδειγμα για το συνδυασμό vascular & surgery.

WORD	YEAR	TIMES	TOTAL BOOKS
άλφα	1980	665	191
άλφα	1981	564	193

Που σημαίνει ότι η λέξη π.χ. άλφα εμφανίστηκε το 1980, 665 φορές σε σύνολο 191 βιβλίων.

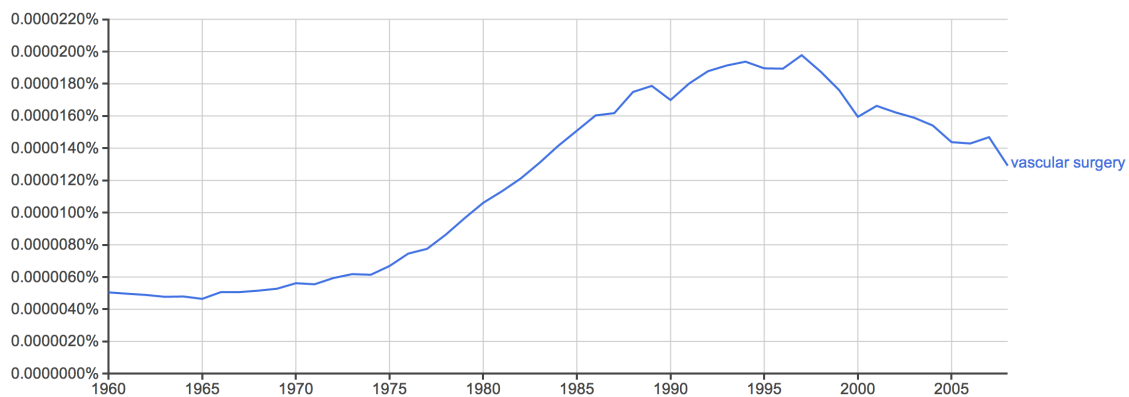
Από το παραπάνω δημιουργήσαμε γραφήματα x-y με x άξονα τη χρονολογία και y άξονα τη συχνότητα εμφάνισης μια λέξης επί του συνόλου των λέξεων. Τα γραφήματα αυτά υποθέτουμε ότι μπορούν να αποτελέσουν ένδειξη προς συγκεκριμένες τάσεις ιδιαίτερα ανάλογα με το βαθμό συσχέτισης των καμπυλών. Το insight στην περίπτωση αυτή προέρχεται από το βαθμό συσχέτισης συγκεκριμένων λέξεων μεταξύ τους ως προς τις καμπύλες τους. Οι καμπύλες αυτές αποτελούν ουσιαστικά το σώμα της συνολικής γνώσης διαθέσιμης στην ανθρωπότητα.

ΑΠΟΤΕΛΕΣΜΑΤΑ



Εικόνα 1: Η εμφάνιση της λέξης αγγειακή χειρουργική από το 1800 και έπειτα

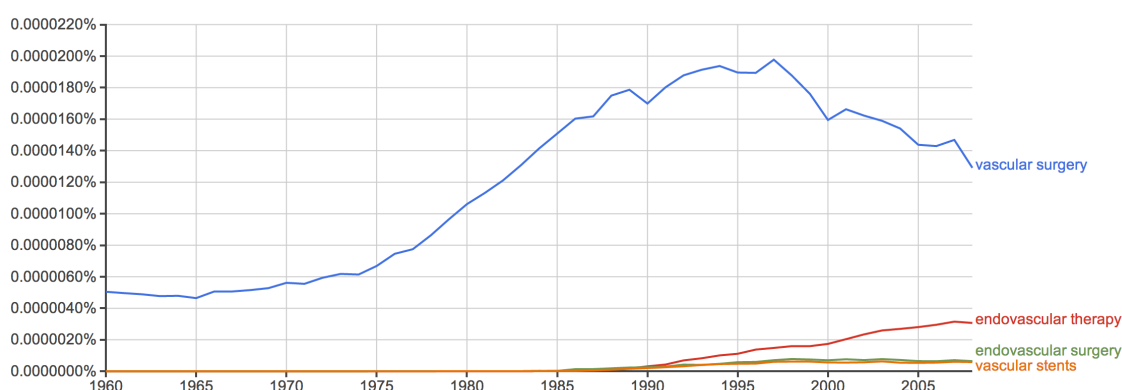
Παρατηρούμε ότι η εμφάνιση της λέξης είναι αντίστοιχη με την ιστορία της αγγειακής χειρουργικής ενώ εμφανίζει μία κορυφή το 1990 όπου από εκεί και έπειτα ξεκινάει πτώση της συχνότητας ως το 2008 που υπάρχουν διαθέσιμα δεδομένα για βιβλία.



Εικόνα 2: Γράφημα από το 1960 και έπειτα για το δίγραμμο vascular surgery

Μπορεί βέβαια να αναρωτηθεί κανείς τι είναι αυτό που έχει προκαλέσει τη σχετική πτώση στην ανοιχτή χειρουργική. Αυτό βέβαια που έρχεται πρώτο στο νου είναι η έλευση των ενδαγγειακών μεθόδων και τεχνικών. Λιγότερη ανοιχτή χειρουργική λοιπόν.

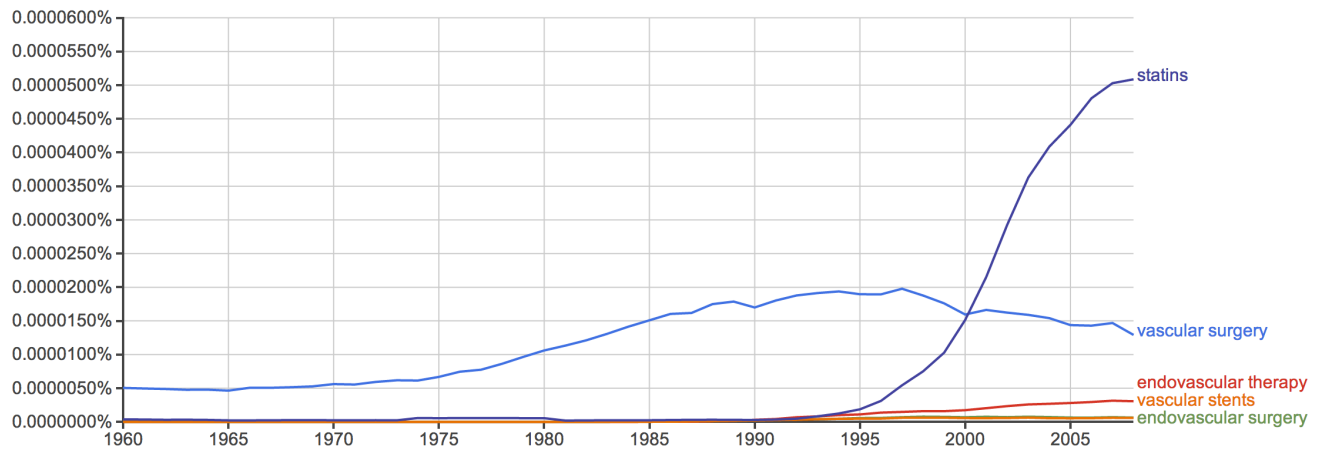
Για να δούμε τι δείχνει το σώμα της γνώσης που έχουμε ανιχνεύσει για την επίπτωση που έχει η ενδαγγειακή αντιμετώπιση στην ανοιχτή αγγειακή χειρουργική.



Εικόνα 3: Σύγχρονο γράφημα ανοιχτής και ενδαγγειακής χειρουργικής

Παρατηρούμε ότι η άνοδος της ενδαγγειακής χειρουργικής από το 1990 και έπειτα οδήγησε σε χαρακτηριστική πτώση της ανοιχτής χειρουργικής που συμφωνεί με το ένστικτο μας.

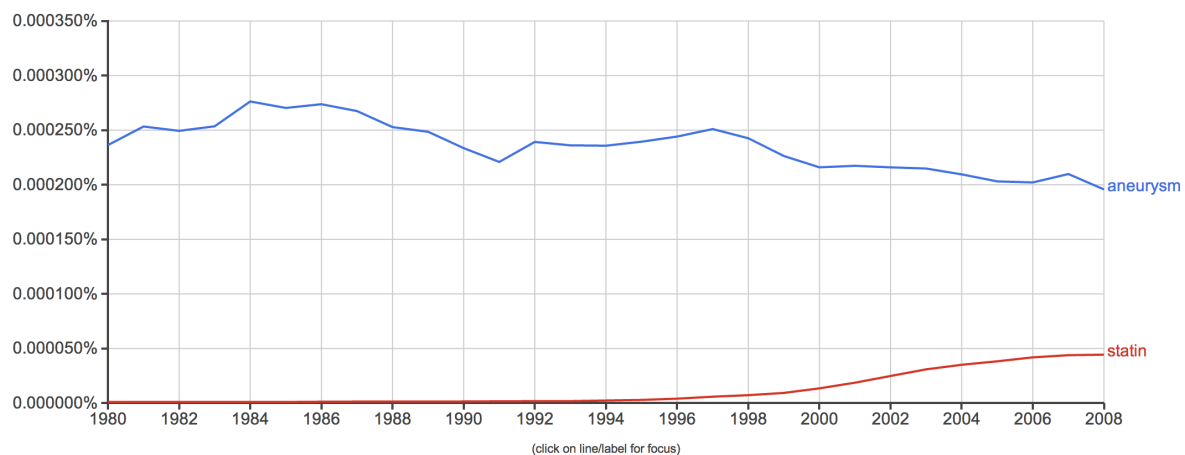
Ας δούμε όμως την πιθανή επίπτωση που μπορεί να έχει όμως η στατίνη, η οποία καταπολεμά το κυρίαρχο αίτιο για το οποίο οι χειρουργοί αγγείων είναι χρήσιμοι: την αθηροσκλήρωση.



Εικόνα 4: Η επίδραση της στατίνης σε σχέση με την ενδαγγειακή χειρουργική.

Παρατηρούμε πως συμπίπτει η άνοδο στη συχνότητα αναφοράς της στατίνης εκθετικά και πως φαίνεται να έχει άμεση σχέση με τη μείωση στη συχνότητα αναφοράς στην ανοιχτή χειρουργική. Παρατηρούμε επίσης πόσο μεγαλύτερη είναι αυτή η επίδραση σε σχέση με την ενδαγγειακή χειρουργική.

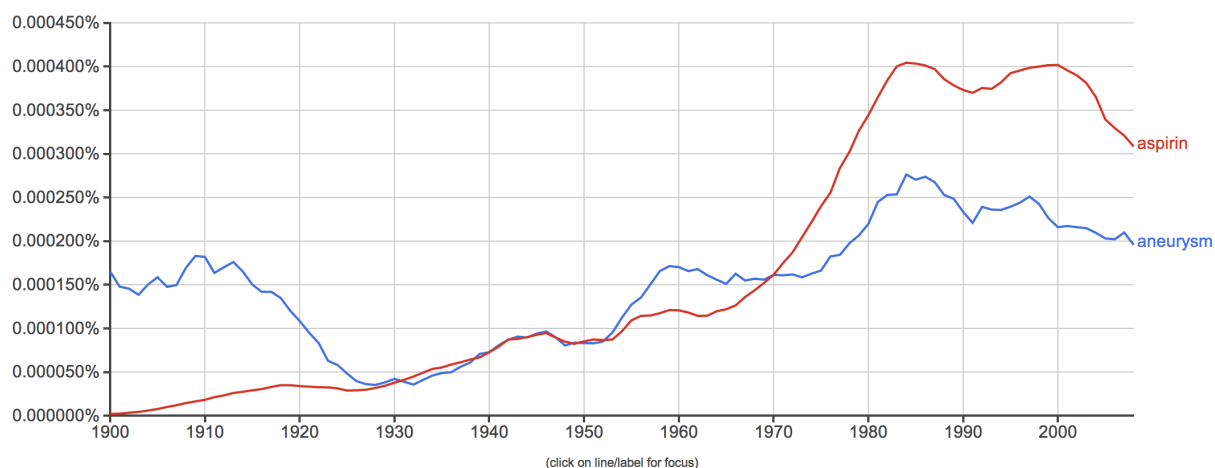
Μπορούμε επίσης να ανιχνεύσουμε την πιθανή συσχέτιση μεταξύ ανευρύσματος και στατίνης η οποία δε φαίνεται να είναι ιδιαίτερα ισχυρή.



Εικόνα 5: Σχέση ανευρύσματος και στατίνης

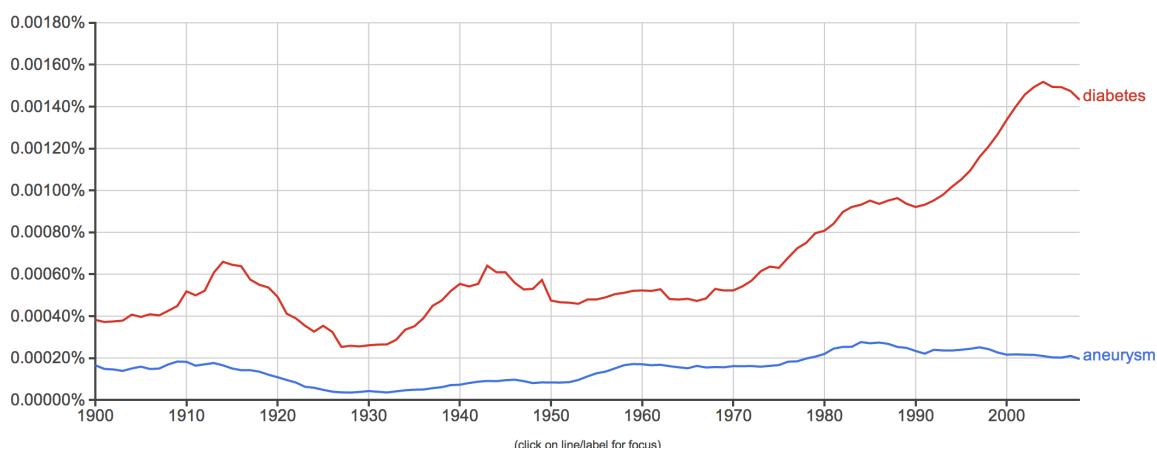
Εκείνο όμως το φάρμακο που απρόσμενα φαίνεται να επηρεάζει και επηρεάζεται από τη λέξη ανεύρυσμα είναι η γνωστή μας ασπιρίνη. Είναι μάλιστα εντυπωσιακό πως αλληλεπιδρούν αυτές οι δυο λέξεις στο γράφημα και παρουσιάζουν ισχυρή

συσχέτιση. Πιθανώς αυτό να αποτελεί ένδειξη για περαιτέρω διερεύνηση αυτής της σχέσης.



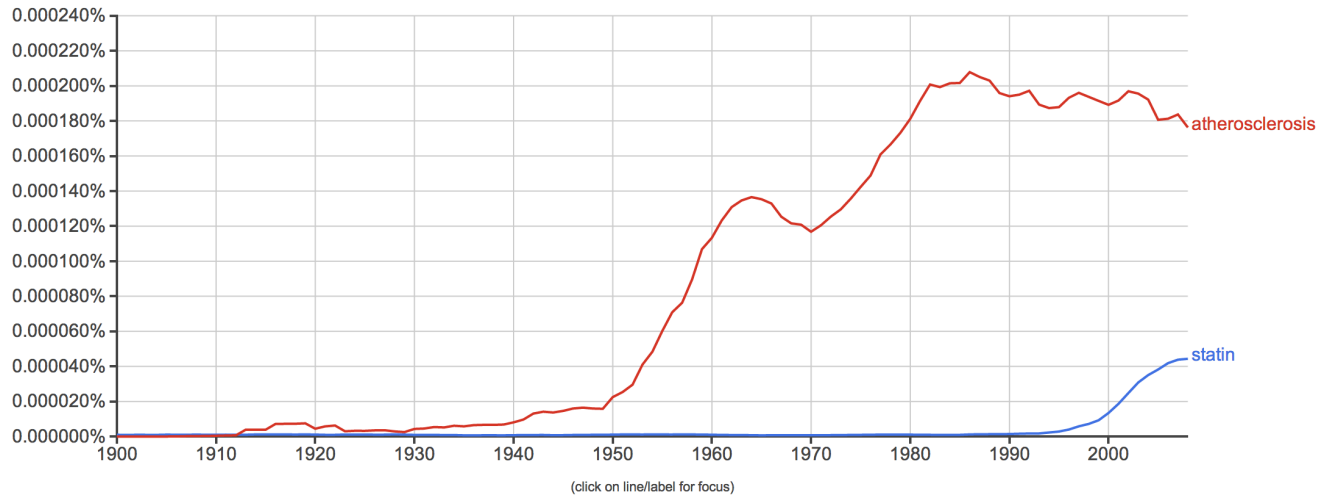
Εικόνα 6: Συσχέτιση ανευρύσματος με ασπιρίνη.

Σε αντίθεση με την επίδραση του διαβήτη στο ανεύρυσμα όπου παρόλη την αύξηση στην αναφορά αυτού δεν παρατηρείται καμμία συσχέτιση με το ανεύρυσμα.



Εικόνα 7: Συσχέτιση διαβήτη και ανευρύσματος

και για του λόγου το αληθές η στατίνη όντως επιδρά στην αθηροσκλήρωση σε σημαντικό βαθμό.



Εικόνα 8: Συσχέτιση αθηροσκλήρωσης με στατίνη.

Τι μας επιφυλάσσει το μέλλον της αγγειακής χειρουργικής?

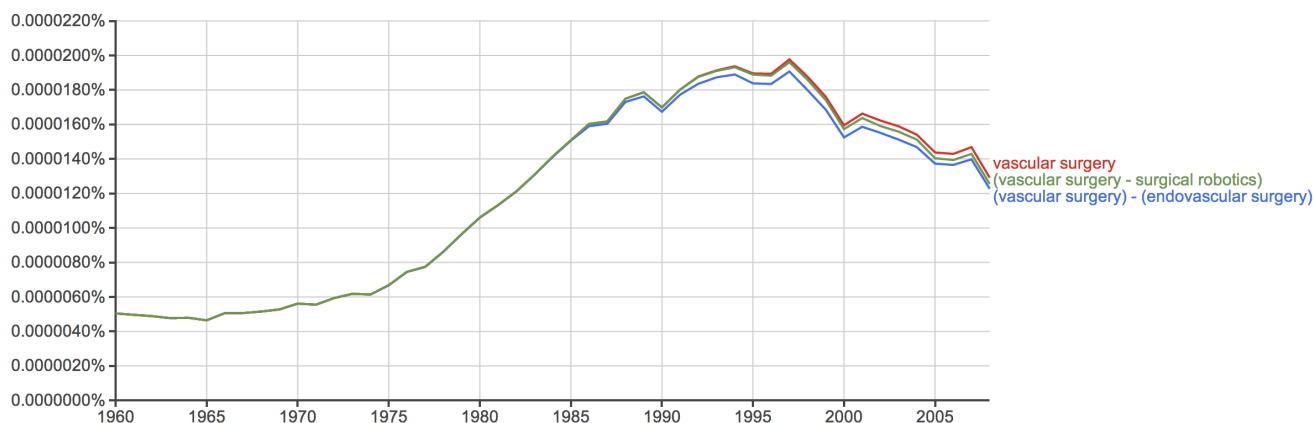
Τι θα γίνει τα επόμενα χρόνια? Έχουμε βιώσει μια σημαντική αλλαγή με την έλευση των ενδαγγειακών τεχνικών.

Είναι αυτές το μέλλον?

Στην εικόνα 9 φαίνεται η επίδραση της ρομποτικής και της ενδαγγειακής χειρουργικής στην ανοιχτή χειρουργική η οποία δείχνει να είναι σημαντική και γραμμική.

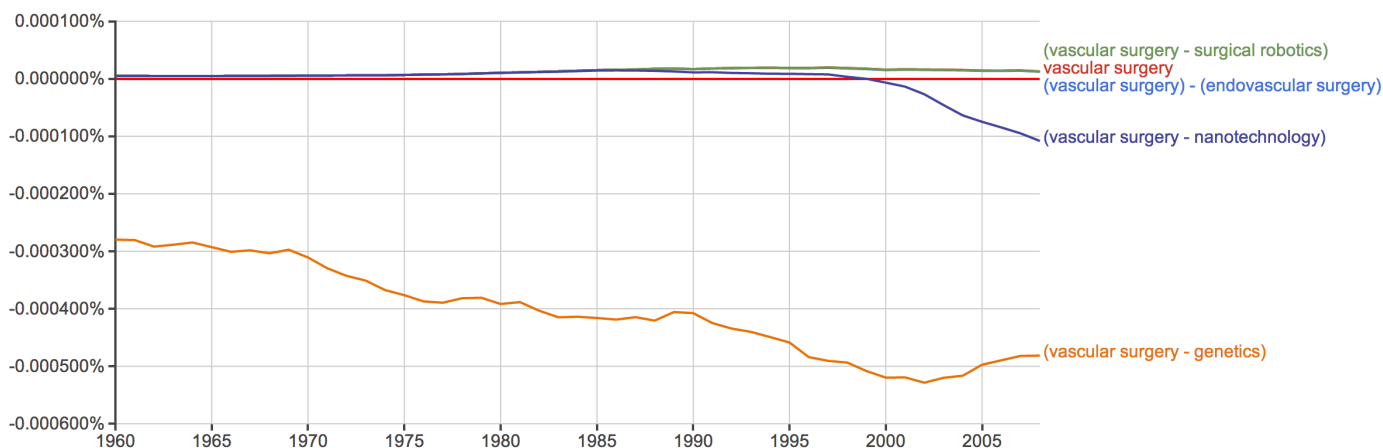
Μπορεί κανείς με ασφάλεια να υποθέσει ότι αυτές οι τεχνικές θα συνεχίσουν να επηρεάζουν την αγγειακή χειρουργική στο κοντινό μέλλον.

Δεν φαίνεται όμως να είναι οι τεχνολογίες που θα προκάλεσουν την επανάσταση στην αγγειακή χειρουργική.



Εικόνα 9: Επίδραση στην αγγειακή χειρουργική της ρομποτικής και της ενδαγγειακής χειρουργικής

Ποιες λοιπόν θα μπορούσαν να είναι αυτές οι τεχνικές που θα φέρουν την επανάσταση και θα αλλάξουν το μέλλον της αγγειακής χειρουργικής?



Εικόνα 10: Πιθανές επαναστατικές αλλαγές στο μέλλον της αγγειακής χειρουργικής

Από το παραπάνω γράφημα φαίνεται η τεράστια επίδραση που έχουν τεχνολογίες όπως η γενετική με κορυφή το 2000 και έπειτα σταθεροποίηση καθώς και η νανοτεχνολογία ιδιαίτερα από το 2000 και μετά.

Οι μελλοντικές τάσεις είναι ξεκάθαρες αν και δεν έχουν περάσει ακόμα στην κλινική πράξη. Για το μέλλον ίσως θα ήταν καλύτερα κάποιος να γίνει νανοτεχνολόγος ή γενετιστής αντί για αγγειακός χειρουργός. Ίσως όμως η εμπειρία της χειρουργικής να μην αντικαθιστάται με τίποτα. Ίσως και μεις οι νεότεροι να ήρθε η ώρα να αντικατασταθούμε από τη νέα γενιά. Είναι γεγονός όμως ότι η μελλοντική έρευνα δε θα είναι τυφλή αναζήτηση συσχετίσεων σε δεδομένα αλλά το αντίστροφο: θα ρωτάμε αρχικά ποια δεδομένα φαίνεται να έχουν συσχέτιση ώστε μετά να σχεδιάζουμε τη μελέτη για να αποσαφηνίσουμε αυτή τη συσχέτιση. Το μέλλον είναι εδώ, απλά δεν έχει ακόμα διανεμηθεί ισότιμα (William Gibson, Neuromancer).

BIBΛΙΟΓΡΑΦΙΑ

1. Jean-Baptiste Michel*, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, William Brockman, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden*. Quantitative Analysis of Culture Using Millions of Digitized Books. Science (Published online ahead of print: 12/16/2010)
2. Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, William Brockman, Slav Petrov. Syntactic Annotations for the Google Books Ngram Corpus. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics Volume 2: Demo Papers (ACL '12) (2012)
3. hadoop.com
4. Python Natural Processing Kit. <http://www.nltk.org>
5. Amazon Elastic Computing <http://aws.amazon.com>