

ΠΜΣ ΒΙΟΣΤΑΤΙΣΤΙΚΗ

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΙΑΤΡΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ

ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΛΟΥΜΠΙΑΝΑ ΚΟΥΛΑΙ

Ελλείπουσες τιμές λόγω περικοπής μετρήσεων σε διαχρονικά συνεχή δεδομένα: χρήση μοντέλων για πληροφοριακή περικοπή όταν η αποκοπή είναι αγνοήσιμη.

ΑΘΗΝΑ, 2013

Περιεχόμενα

Περίληψη.....	6
Abstract	8
Κεφάλαιο 1: Εισαγωγή	9
1.1 Διαχρονικές επαναλαμβανόμενες μελέτες και ελλείπουσες τιμές	9
1.2 Ελλείπουσες τιμές σε διαχρονικές μελέτες.	12
1.3 HIV λοίμωξη και CD4 κύτταρα.....	13
1.4 Σκοπός	15
Κεφάλαιο 2: Μέθοδοι για την ανάλυση διαχρονικών δεδομένων.....	16
2.1 Εισαγωγή	16
2.2 Ανάλυση σε δύο στάδια	16
2.2.1 1 ^ο Στάδιο	16
2.2.2 2 ^ο Στάδιο	17
2.3 Γραμμικό μοντέλο τυχαίων επιδράσεων	17
2.4 Εκτίμηση παραμέτρων στο γραμμικό μικτό μοντέλο	20
2.4.1 Μέθοδοι βασισμένοι στην Μέγιστη Πιθανοφάνεια ML/REML	20
2.4.2 Επαναληπτικές μέθοδοι βασισμένες στην θεωρία γενικευμένων ελαχίστων τετραγώνων IGLS/RIGLS	22
Κεφάλαιο 3: Ελλείπουσες τιμές (Missing Data)	27
3.1 Εισαγωγή	27
3.2 Ορολογία	27
3.3 Μηχανισμοί Ελλειπουσών Τιμών.....	28
3.3.1 Εντελώς τυχαίες (Missing Completely At Random - MCAR).....	28
3.3.2 Τυχαίες (Missing At Random - MAR).....	29
3.3.3 Μη αγνοήσιμες (Missing Not At Random - MNAR).....	29
3.4 Μέθοδοι διαχείρισης των ελλειπουσών τιμών	30
3.4.1 Ανάλυση πλήρων δεδομένων (Complete Case Analysis)	30
3.4.2 Ανάλυση διαθέσιμων κατά περίπτωση δεδομένων (Available Case Analysis).....	30
3.4.3 Υποκατάσταση ελλειπουσών τιμών (Imputation Methods)	31
3.5 Επίδραση ελλειπουσών τιμών στην συμπερασματολογία.....	34
3.6 Σύνοψη μεθόδων για την ανάλυση δεδομένων με μη αγνοήσιμο μηχανισμό δημιουργίας ελλειπουσών τιμών.....	37

Κεφάλαιο 4: Joint Models	42
4.1 Εισαγωγή	42
4.2 Random effects selection models	43
4.3 Περιγραφή του Joint Multivariate Random Effects μοντέλου (JMRE)	45
4.4 Διόρθωση του πίνακα διασποράς συνδιασποράς	48
Κεφάλαιο 5: Διερεύνηση διαχρονικών δεδομένων – Η μελέτη AMACS	52
5.1 Εισαγωγή	52
5.2 Μελέτη AMACS	53
5.3 Περιγραφή χρόνου, συχνότητας και αριθμού μετρήσεων CD4 λεμφοκυττάρων	56
5.4 Μοντέλο μικτών επιδράσεων	62
5.5 Λογιστική παλινδρόμηση σε επαναλαμβανόμενα δεδομένα	65
5.6 Ανάλυση Επιβίωσης	69
5.7 JMRE μοντέλο σε πραγματικά δεδομένα	71
Κεφάλαιο 6 : Προσομοιώσεις	75
Προσομοιωμένα δεδομένα	75
6.1 Σύγκριση μοντέλου μικτών επιδράσεων σε σχέση με το JMRE μοντέλο για τα 4 διαφορετικά σενάρια θεραπείας και με πιθανότητα προαιρετικής επίσκεψης 10%.	78
6.1.1 Αποτελέσματα προσομοιώσεων	78
6.1.2 Box plots για την μεροληψία της σταθεράς και της κλίσης των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων	79
6.1.3 95% διάστημα εμπιστοσύνης για την σταθερά και την κλίση των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων.	80
6.2 Σύγκριση μοντέλου μικτών επιδράσεων σε σχέση με το JMRE μοντέλο για τα 4 διαφορετικά σενάρια θεραπείας και με πιθανότητα προαιρετικής επίσκεψης που εξαρτάται από τα παρατηρηθέντα CD4 στην προηγούμενη επίσκεψη	81
6.2.1 Αποτελέσματα προσομοιώσεων	81
6.2.2 Box plots για την μεροληψία της σταθεράς και της κλίσης των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων.	82
6.2.3 95% διάστημα εμπιστοσύνης για την σταθερά και την κλίση των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων.	83
6.3 Σύγκριση μοντέλου μικτών επιδράσεων σε σχέση με το JMRE μοντέλο για τα 4 διαφορετικά σενάρια θεραπείας και με πιθανότητα προαιρετικής επίσκεψης που εξαρτάται από τα παρατηρηθέντα CD4 στην προηγούμενη επίσκεψη και έναν όρο frailty.	84
6.3.1 Αποτελέσματα προσομοιώσεων	84

6.3.2 Box plots για την μεροληψία της σταθεράς και της κλίσης των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων.	85
6.3.3 95% διάστημα εμπιστοσύνης για την σταθερά και την κλίση των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.	86
6.4 Σύγκριση μοντέλου μικτών επιδράσεων σε σχέση με το JMRE μοντέλο για τα 4 διαφορετικά σενάρια θεραπείας, με πιθανότητα προαιρετικής επίσκεψης 10%, και με αρνητική συσχέτιση μεταξύ frailty και random effects.	89
6.4.1 Αποτελέσματα προσομοιώσεων	89
6.4.2 Box plots για την μεροληψία της σταθεράς και της κλίσης των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.	90
6.4.3 95% διάστημα εμπιστοσύνης για την σταθερά και την κλίση των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.	91
6.5 Σύγκριση μοντέλου μικτών επιδράσεων σε σχέση με το JMRE μοντέλο για τα 4 διαφορετικά σενάρια θεραπείας, με πιθανότητα προαιρετικής επίσκεψης που εξαρτάται από τα παρατηρηθέντα CD4 στην προηγούμενη επίσκεψη και με αρνητική συσχέτιση μεταξύ frailty και random effects.	92
6.5.1 Αποτελέσματα προσομοιώσεων	92
6.5.2 Box plots για την μεροληψία της σταθεράς και της κλίσης των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.	93
6.5.3 95% διάστημα εμπιστοσύνης για την σταθερά και την κλίση των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.	94
6.6 Σύγκριση μοντέλου μικτών επιδράσεων σε σχέση με το JMRE μοντέλο για τα 4 διαφορετικά σενάρια θεραπείας, με πιθανότητα προαιρετικής επίσκεψης που εξαρτάται από τα παρατηρηθέντα CD4 στην προηγούμενη επίσκεψη και από έναν όρο frailty, και με αρνητική συσχέτιση μεταξύ frailty και random effects.	95
6.6.1 Αποτελέσματα προσομοιώσεων	95
6.6.2 Box plots για την μεροληψία της σταθεράς και της κλίσης των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.	96
6.6.3 95% διάστημα εμπιστοσύνης για την σταθερά και την κλίση των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.	97

6.7 Σύγκριση μοντέλου μικτών επιδράσεων σε σχέση με το JMRE μοντέλο για τα 4 διαφορετικά σενάρια θεραπείας, με πιθανότητα προαιρετικής επίσκεψης 10%, και με θετική συσχέτιση μεταξύ frailty και random effects.....	100
6.7.1 Αποτελέσματα προσομοιώσεων	100
6.7.2 Box plots για την μεροληψία της σταθεράς και της κλίσης των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.....	101
6.7.3 95% διάστημα εμπιστοσύνης για την σταθερά και την κλίση των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.....	102
6.8 Σύγκριση μοντέλου μικτών επιδράσεων σε σχέση με το JMRE μοντέλο για τα 4 διαφορετικά σενάρια θεραπείας, με πιθανότητα προαιρετικής επίσκεψης που εξαρτάται από τα παρατηρηθέντα CD4 στην προηγούμενη επίσκεψη και με θετική συσχέτιση μεταξύ frailty και random effects.	103
6.8.1 Αποτελέσματα προσομοιώσεων	103
6.8.2 Box plots για την μεροληψία της σταθεράς και της κλίσης των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.....	104
6.8.3 95% διάστημα εμπιστοσύνης για την σταθερά και την κλίση των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.....	105
6.9 Σύγκριση μοντέλου μικτών επιδράσεων σε σχέση με το JMRE μοντέλο για τα 4 διαφορετικά σενάρια θεραπείας, με πιθανότητα προαιρετικής επίσκεψης που εξαρτάται από τα παρατηρηθέντα CD4 στην προηγούμενη επίσκεψη και από έναν όρο frailty, και με αρνητική συσχέτιση μεταξύ frailty και random effects.	106
6.9.1 Αποτελέσματα προσομοιώσεων	106
6.9.2 Box plots για την μεροληψία της σταθεράς και της κλίσης των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.....	107
6.9.3 95% διάστημα εμπιστοσύνης για την σταθερά και την κλίση των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.....	108
Συμπεράσματα – Συζήτηση.....	109
ΒΙΒΛΙΟΓΡΑΦΙΑ	111

Περίληψη

Η ανάλυση πολλαπλών ανά άτομο, διαχρονικών μετρήσεων συχνά γίνεται πιο περίπλοκη λόγω ύπαρξης ελλειπουσών τιμών. Μια συνηθισμένη περίπτωση είναι αυτή κατά την οποία η προγραμματισμένη σειρά μετρήσεων διακόπτεται πρόωρα εξαιτίας της εμφάνισης κάποιου συμβάματος. Χαρακτηριστικό παράδειγμα αποτελούν οι μελέτες φυσικής ιστορίας της HIV-1 λοίμωξης όπου οι μετρήσεις ενός βιολογικού δείκτη διακόπτονται λόγω θανάτου του ασθενούς. Παράλληλα, μετρήσεις πέρα από κάποιο χρονικό σημείο αποκόπτονται τεχνητά, κατά το στάδιο της ανάλυσης, επειδή ο ασθενής ανέπτυξε κλινικό AIDS ή άρχισε να λαμβάνει αντιρετροϊκή θεραπεία (και οι δυο αυτοί παράγοντες σηματοδοτούν εξ' ορισμού την φυσική ιστορία της νόσου). Στην περίπτωση όπου μηχανισμοί αποκοπής όπως οι παραπάνω σχετίζονται με μη παρατηρούμενες ποσότητες, η περικοπή χαρακτηρίζεται ως πληροφοριακή και η εξαγωγή αμερόληπτων εκτιμήσεων για την χρονική εξέλιξη των επιπέδων του δείκτη απαιτεί χρήση τεχνικών από κοινού μοντελοποίησης (joint models) του δείκτη και του συμβάματος που οδηγεί στην περικοπή των μετρήσεων. Αντιθέτως, στην περίπτωση όπου οι μηχανισμοί περικοπής εξαρτώνται μόνο από ήδη παρατηρημένες τιμές του μελετώμενου δείκτη, η χρήση μεθόδων μέγιστης πιθανοφάνειας οδηγεί σε αμερόληπτες εκτιμήσεις.

Η εργασία αυτή επικεντρώνεται στην ανάλυση του απόλυτου αριθμού CD4 λεμφοκυττάρων που αποτελεί σημαντικό ανοσολογικό δείκτη κατά την φυσική ιστορία της HIV-1 λοίμωξης. Σε τέτοια δεδομένα, ο κύριος μηχανισμός περικοπής μετρήσεων, για άτομα που μολύνθηκαν μετά το 1996, έτος όπου η αντιρετροϊκή θεραπεία υψηλής δραστηριότητας άρχισε να διατίθεται ευρέως, είναι η έναρξη θεραπείας. Το γεγονός ότι η απόφαση για έναρξη θεραπείας βασίζεται κυρίως σε ήδη παρατηρημένες μετρήσεις αριθμού CD4 λεμφοκυττάρων, κάνει τον μηχανισμό αυτό θεωρητικά τουλάχιστον αγνοήσιμο. Τα πραγματικά δεδομένα που χρησιμοποιήθηκαν προέρχονται από την Ελληνική προοπτική μελέτη AMACS, ενώ οι παράμετροι για την δημιουργία των προσομοιωμένων δεδομένων προέκυψαν από την Ευρωπαϊκή μελέτη CASCADE. Εφαρμόστηκε ένα μοντέλο τυχαίων επιδράσεων και ένα μοντέλο (JMRE) που συνδυάζει ένα μικτό μοντέλο για την εκτίμηση της μεταβολής του δείκτη και ένα λογαριθμικό

μοντέλο επιβίωσης για την πληροφοριακή απόσυρση. Η απόδοση των μοντέλων αυτών αξιολογήθηκε σε προσομοιωμένα δεδομένα. Κατασκευάστηκαν τέσσερα διαφορετικά σενάρια που αντανακλούν την περικοπή παρατηρήσεων λόγω έναρξη θεραπείας.

Ο μηχανισμός παραγωγής των ελλειπουσών τιμών καθορίζει την καταλληλότητα της μεθόδου ανάλυσης που πρέπει να χρησιμοποιηθεί. Το μοντέλο τυχαίων επιδράσεων δίνει αμερόληπτες εκτιμήσεις παρουσία εντελώς τυχαίων (MCAR) ή τυχαίων ελλειπουσών (MAR) τιμών. Στις περιπτώσεις που οι ελλείπουσες τιμές οφείλονται σε δεξιά πληροφοριακή απόσυρση (MNAR) ενδείκνυται η χρήση ενός μοντέλου όπως το JMRE το οποίο μειώνει την μεροληψία σε σύγκριση με το μοντέλο τυχαίων επιδράσεων. Τέλος, η χρήση του JMRE μοντέλου σε περιπτώσεις τυχαίων ελλειπουσών τιμών (MAR) οδηγεί σε μεροληπτικά αποτελέσματα.

Abstract

The analysis of longitudinal data is more complicated when there are missing data. It is a well known fact that in most cases the repeated measurements are truncated due to a clinical event. When studying the CD4 count evolution during HIV natural history, the series of this marker's measurements can be naturally censored due to death or lost to follow up and artificially censored due to clinical AIDS onset or antiretroviral treatment (ART) initiation. In the combined ART (cART) era (i.e. post 1996), the most prevalent censoring mechanism is cART initiation. If the censoring mechanism is informative then models that jointly model the marker's evolution and the drop-out mechanism such as the Joint Multivariate Random Effects (JMRE) model, will give unbiased estimations. On the other hand, if the censoring mechanism is non-informative then models which are based in maximum likelihood estimations lead to unbiased estimations. The aim of the present study is to investigate the performance of linear mixed models and JMRE models when analyzing such data under various plausible scenarios regarding cART initiation and timing.

This study focuses on the analysis of CD4 cell count, which is an important marker of the HIV natural history. The decision for cART initiation is based on observed CD4 cell count and the censoring mechanism is considered as ignorable. We used real data from Athens Multicenter AIDS Cohort Study – AMACS. We applied a linear mixed model and a JMRE model in the real data. However, the performance of the above methods is based on simulated data drawn from the European CASCADE study. Four distinct scenarios which illustrate the censoring mechanism due to cART initiation have been explored.

The missingness process plays a vital role on determination of the ideal method to be applied. The random effect model always results in unbiased estimates in presence of completely at random (MCAR) and missing at random (MAR) data, while in presence of non-ignorable missingness it could result in biased estimates. In case of right informative censoring, the application of a JMRE model that minimizes bias is highly recommended.

Κεφάλαιο 1: Εισαγωγή

1.1 Διαχρονικές επαναλαμβανόμενες μελέτες και ελλείπουσες τιμές

Είναι ευρέως γνωστό ότι οι επαναλαμβανόμενες μετρήσεις χρησιμοποιούνται σχεδόν σε όλες τις εφαρμοσμένες επιστήμες καθώς και στις βιοϊατρικές μελέτες. Ιδιαίτερα, στις επιδημιολογικές μελέτες και στις κλινικές δοκιμές η χρήση επαναλαμβανόμενων μετρήσεων είναι συχνή διότι στην πρώτη περίπτωση συχνά το ενδιαφέρον επικεντρώνεται στην διαχρονική πορεία της νόσου ενώ στην δεύτερη περίπτωση στόχος είναι η αξιολόγηση των νέων θεραπειών ειδικά σε χρόνιες νόσους. Με τον όρο επαναλαμβανόμενες μετρήσεις εννοούμε μετρήσεις της ίδιας μεταβλητής (ή μεταβλητών) επί της ίδιας ερευνητικής μονάδας. Η επανάληψη των μετρήσεων μπορεί να αφορά σε διαφορετικά χωρικά σημεία και σε αυτή την περίπτωση χαρακτηρίζονται ως χωρικά δεδομένα (spatial data) ή σε διαφορετικά χρονικά σημεία και χαρακτηρίζονται ως διαχρονικά δεδομένα (longitudinal data) και οι αντίστοιχες μελέτες ονομάζονται διαχρονικές μελέτες (longitudinal studies).

Στις βιοϊατρικές επιστήμες χρησιμοποιούνται όλο και περισσότερο διαχρονικές μελέτες. Σε αυτές τις μελέτες παρακολουθείται μια ομάδα ατόμων για ένα συγκεκριμένο χρονικό διάστημα και για κάθε άτομο λαμβάνονται επαναλαμβανόμενες μετρήσεις για ένα ή περισσότερα χαρακτηριστικά. Υπάρχουν δύο βασικές κατηγορίες στις οποίες ταξινομούνται οι επιδημιολογικές μελέτες: παρεμβατικές και μη παρεμβατικές. Στις παρεμβατικές μελέτες υπάρχει κάποια παρέμβαση του ερευνητή (π.χ. θεραπεία), σε αυτή την κατηγορία ανήκουν και οι κλινικές δοκιμές που είναι συνήθως τυχαιοποιημένες και στοχεύουν στην αξιολόγηση των αποτελεσμάτων συγκεκριμένων παρεμβάσεων οι οποίες αποτελούν και το αντικείμενο της μελέτης και ορίζονται εκ των προτέρων. Στις μη παρεμβατικές μελέτες δεν πραγματοποιείται καμία παρέμβαση από τον ερευνητή, σε αυτή την κατηγορία ανήκουν οι προοπτικές, οι αναδρομικές και οι συγχρονικές μελέτες. Οι μη παρεμβατικές μελέτες έχουν συνήθως προοπτικό χαρακτήρα, σημείο εκκίνησης μπορεί να είναι η εμφάνιση κάποιας νόσου ή η ημερομηνία μόλυνσης από κάποιο

λοιμώδες νόσημα. Στην περίπτωση όπου τα στοιχεία συλλέγονται αναδρομικά από ιατρικούς φακέλους οι μελέτες αυτές ονομάζονται αναδρομικές προοπτικές μελέτες (retrospective cohort study).

Στις διαχρονικές μελέτες λόγω των επαναλαμβανόμενων μετρήσεων δίνεται η δυνατότητα της διερεύνησης των αλλαγών του υπό μελέτη χαρακτηριστικού (ή χαρακτηριστικών) ανά τον χρόνο, ενώ στις συγχρονικές (cross-sectional) κάτι τέτοιο είναι αδύνατο διότι λαμβάνεται μόνο μια μέτρηση ανά άτομο. Είναι γεγονός ότι για την πλειονότητα των μεταβλητών παρατηρείται μεγαλύτερη μεταβλητότητα μεταξύ των ατόμων μιας μελέτης ("μεταξύ των ατόμων" μεταβλητότητα - between subject variability) παρά μεταξύ των επαναλαμβανόμενων μετρήσεων του ίδιου ατόμου ("εντός του ατόμου" μεταβλητότητα - within subject variability). Η μεταξύ των ατόμων μεταβλητότητα οφείλεται σε διαφορές μεταξύ των ατόμων που είτε δεν καταγράφονται (π.χ. κοινωνικό-οικονομικές διαφορές) είτε δεν είθισται να καταγραφούν (π.χ. γενετικές διαφορές), αλλά παραμένουν σταθερές στο χρόνο. Ωστόσο, στις διαχρονικές μελέτες η επίδραση τέτοιων παραγόντων μπορεί να εξαλειφθεί σε ικανοποιητικό βαθμό αφού η συσχέτιση της "εντός του ατόμου" μεταβλητότητας μπορεί να ενσωματωθεί στην ανάλυση και έτσι κάθε άτομο λειτουργεί ως μάρτυρας (control) του εαυτού του.

Τα χαρακτηριστικά που μετρώνται σε μια διαχρονική μελέτη μπορεί να είναι σταθερά στο χρόνο (π.χ. δημογραφικά χαρακτηριστικά) τα οποία καταγράφονται μόνο στην έναρξη της μελέτης, ή να μεταβάλλονται διαχρονικά οπότε απαιτούνται επαναλαμβανόμενες στο χρόνο μετρήσεις. Τα χαρακτηριστικά που μεταβάλλονται στο χρόνο μπορεί να είναι είτε οι κύριες μεταβλητές των οποίων οι μεταβολές στο χρόνο αποτελούν το αντικείμενο μελέτης είτε μεταβλητές που μπορεί να παίζουν ρόλο στην εξέλιξη των κύριων χαρακτηριστικών και είναι αναγκαίο να ληφθούν υπόψη στην ανάλυση των δεδομένων.

Σε πολλές επιδημιολογικές μελέτες και ιδιαίτερα σε διαχρονικές μελέτες παρατήρησης που μελετούν την φυσική πορεία της νόσου χρόνιων νοσημάτων, τα χαρακτηριστικά κύριου ενδιαφέροντος που καταγράφονται είναι συχνά κάποιοι δείκτες που σχετίζονται με την πορεία της συγκεκριμένης νόσου. Η χρήση τέτοιων δεικτών είναι χρήσιμη για την

παρακολούθηση της κατάστασης ενός ατόμου στην διάρκεια του χρόνου και σχετίζεται με πολλές εφαρμογές όπως (Jewell and Klabfleisch, 1992):

- κατασκευή μοντέλων σταδίων
- κατανόηση αιτιολογικών σχέσεων στην εξέλιξη της νόσου
- άντληση προγνωστικών πληροφοριών σχετικά με κλινικές εκβάσεις
- παρακολούθηση της πορείας της νόσου και καθοδήγηση σχετικά με την θεραπευτική στρατηγική
- σύνδεση των επιπέδων των δεικτών με άλλες παραμέτρους της φυσικής ιστορίας της νόσου.

Εκτός από τη μελέτη της φυσικής ιστορίας χρόνιων νοσημάτων, οι επαναλαμβανόμενες μετρήσεις δεικτών χρησιμοποιούνται ευρέως ως αναπληρωτές δείκτες (surrogate markers) για τελικές εκβάσεις σε τυχαιοποιημένες κλινικές δοκιμές που έχουν ως σκοπό την αξιολόγηση θεραπευτικών παρεμβάσεων σε ασθενείς με μακροχρόνια πρόγνωση. Σε αυτές τις μελέτες το ενδιαφέρον επικεντρώνεται στην εκτίμηση και την σύγκριση του ρυθμού μεταβολής των επιπέδων του δείκτη μεταξύ ομάδων ασθενών που λαμβάνουν διαφορετικές θεραπείες. Σύμφωνα με πολλούς ερευνητές, το βασικό κίνητρο για την χρήση αναπληρωτών δεικτών σε κλινικές δοκιμές είναι η αναμενόμενη μείωση στο απαιτούμενο μέγεθος δείγματος ή/ και στο χρόνο διεξαγωγής της μελέτης. Μειώσεις στο μέγεθος του δείγματος παίζουν καθοριστικό ρόλο στο κόστος της κλινικής δοκιμής αλλά ακόμα και στην ικανότητα πραγματοποίησης της (Ellenberg and Hamilton, 1989; Wittes et al, 1989; Hillis and Siegel, 1989). Επιπρόσθετα, υπάρχουν περιπτώσεις όπου η καταγραφή της τελικής έκβασης δεν είναι δυνατή για ποικίλους λόγους και είναι προτιμότερο να αντικατασταθεί από έναν αναπληρωτή δείκτη έστω και αν με αυτό τον τρόπο εισάγεται ένα σφάλμα στην εκτίμηση του χρόνου εμφάνισης της πραγματικής τελικής έκβασης (Prentice, 1989).

1.2 Ελλείπουσες τιμές σε διαχρονικές μελέτες.

Οι διαχρονικές μελέτες κατά κύριο λόγο διεξάγονται βάσει συγκεκριμένου πρωτοκόλλου, όπου τα δεδομένα των ατόμων που συμμετέχουν καταγράφονται σε προκαθορισμένα χρονικά διαστήματα μετά την έναρξη της μελέτης. Ωστόσο, στις μελέτες που διεξάγονται σε ανθρώπους είναι σχεδόν βέβαιο ότι τα πρωτόκολλα αυτά δεν θα τηρηθούν με απόλυτη ακρίβεια, και αυτό μπορεί να οφείλεται σε σταδιακή είσοδο των ατόμων στην μελέτη, σε ελλείπουσες επισκέψεις, σε προσέλευση για μετρήσεις σε διαφορετικά από τα προκαθορισμένα χρονικά διαστήματα, σε αποχώρηση από την μελέτη ή γενικά σε τερματισμό των μετρήσεων για διάφορους λόγους. Αυτό έχει ως αποτέλεσμα να δημιουργούνται μη πλήρη σύνολα δεδομένων στα οποία για κάθε άτομο υπάρχουν διαφορετικά σύνολα χρονικών μετρήσεων και μπορεί ο συνολικός αριθμός μετρήσεων ανά άτομο να διαφέρει σημαντικά. Εστιάζοντας σε αυτό το πρόβλημα από την πλευρά της στατιστικής προκύπτει ότι τέτοιου είδους δεδομένα δημιουργούν διάφορα προβλήματα στην στατιστική ανάλυση και για αυτό πρέπει να χειριστούν με ιδιαίτερη προσοχή. Έχουν προταθεί αρκετές μέθοδοι που μπορούν να εφαρμοστούν σε τέτοια σύνολα δεδομένων οι οποίες αξιοποιούν το σύνολο της παρεχόμενης πληροφορίας (Laird and Ware, 1982; Lindstrom and Bates, 1988; Stiratelli et al., 1984; Harville and Mee, 1984; Zeger and Karim, 1991; Zeger et al., 1988; Lange et al., 1992; Vonesh και Carter, 1992). Για την εκτίμηση των παραμέτρων μοντέλων για δεδομένα με ελλείπουσες μετρήσεις είναι απαραίτητη η εφαρμογή κατάλληλων στατιστικών μεθόδων όπως μέθοδοι μέγιστης πιθανοφάνειας, γενικευμένες εξισώσεις εκτίμησης (GEE) ή Μπεϋζιανή στατιστική. Η αμεροληψία των εκτιμητών των μεθόδων αυτών σε περιπτώσεις όπου υπάρχουν ελλείπουσες τιμές εξαρτάται από τον μηχανισμό δημιουργίας των ελλειπουσών τιμών. Τα μοντέλα αυτά διαφέρουν ως προς τις μεθόδους εκτίμησης, τις προϋποθέσεις για την εφαρμογή τους, το εύρος των περιπτώσεων που μπορεί να χρησιμοποιηθούν και την ευκολία εφαρμογής τους στην καθημερινή πρακτική. Στην περίπτωση όπου η πιθανότητα να απουσιάζει μια μέτρηση δεν εξαρτάται από τις μη παρατηρηθείσες ποσότητες, οι μέθοδοι που βασίζονται σε μεγιστοποίηση της πιθανοφάνειας ή Μπεϋζιανή στατιστική δίνουν αμερόληπτες εκτιμήσεις και τα περικομμένα δεδομένα χαρακτηρίζονται ως αγνοήσιμα (ignorable missing data). Στην

αντίθετη περίπτωση, δηλαδή η πιθανότητα να απουσιάζει μια μέτρηση σχετίζεται και με τις μη παρατηρηθείσες ποσότητες, τα περικομμένα δεδομένα χαρακτηρίζονται ως μη αγνοήσιμα (non ignorable missing data) και όλες οι ευρέως χρησιμοποιούμενες μέθοδοι μπορούν να καταλήξουν σε μεροληπτικές εκτιμήσεις (Laird, 1988). Στην βιβλιογραφία προτείνεται η εφαρμογή κατάλληλων μοντέλων που λαμβάνουν υπόψη τόσο τις διαχρονικές μετρήσεις του υπό εξέταση χαρακτηριστικού όσο και τον μηχανισμό που οδηγεί στην ύπαρξη ελλειπυσών μετρήσεων (Little and Rubin, 1987; Little, 1995; Laird, 1988). Τα τελευταία χρόνια έχουν προταθεί διάφορα τέτοια μοντέλα (Wu Carroll, 1988; Schlutcher, 1992; Wu and Bailey, 1989; Diggle and Kenward, 1994; Degruittola και Tu, 1994; Faucett και Thomas, 1996; Touloumi et al., 1999; Scharfstein et al., 1999; Henderson et al., 2000; Roy και Lin, 2002) τα οποία διαφέρουν ως προς τις μεθόδους εκτίμησης, τις προϋποθέσεις εφαρμογής τους, τις περιπτώσεις στις οποίες δύναται να εφαρμοστούν και την ευκολία εφαρμογής τους στην καθημερινή πρακτική.

1.3 HIV λοίμωξη και CD4 κύτταρα

Ο ιός της ανθρώπινης ανοσοανεπάρκειας HIV (Human Immunodeficiency Virus) ανήκει στην κατηγορία των ρετροϊών και προκαλεί το σύνδρομο επίκτητης ανοσοποιητικής ανεπάρκειας (Acquired Immune Deficiency Syndrome - AIDS), κατάσταση κατά την οποία το ανθρώπινο ανοσοποιητικό σύστημα καταρρέει οδηγώντας σε ευκαιριακές λοιμώξεις οι οποίες πολύ συχνά μπορεί να οδηγήσουν στο θάνατο. Έχουν αναγνωριστεί δυο είδη του ιού HIV, ο HIV-1 και ο HIV-2. Ο HIV-1 είναι αυτός που ευθύνεται για την μεγάλη πλειονότητα των μολύνσεων παγκόσμια και είναι αυτός με τις περισσότερες μεταλλάξεις και υποτύπους (περίπου 10 υποτύπους). Στην συνέχεια όταν αναφερόμαστε στον HIV ιό θα εννοούμε τον HIV-1. Τα άτομα που έχουν μολυνθεί με τον ιό HIV αναπτύσσουν AIDS και τελικά πεθαίνουν σε περίπτωση που δεν χορηγηθεί κατάλληλη αντιρετροϊκή θεραπεία. Έχει παρατηρηθεί ότι ο χρόνος από την πρωτολοίμωξη μέχρι την ανάπτυξη AIDS και το θάνατο ποικίλει αρκετά. Ο διάμεσος χρόνος από την πρωτολοίμωξη μέχρι την ανάπτυξη κλινικού AIDS εκτιμάται ότι είναι περίπου 10 χρόνια απουσία θεραπείας.

Για την πρόγνωση και την εξέλιξη της HIV λοίμωξης έχουν προταθεί και χρησιμοποιηθεί διάφοροι ανοσολογικοί και ιολογικοί δείκτες όπως ο απόλυτος αριθμός ή το ποσοστό των CD4 λεμφοκυττάρων, το ιικό φορτίο HIV-RNA στο πλάσμα, τα επίπεδα β₂ μικροσφαιρίνης και νεοπερίνης, το HIV-1 p24 αντιγόνο (Lang et al, 1989; Fahey et al, 1990; Stein et al, 1992). Μεταξύ αυτών ο καλύτερος προγνωστικός δείκτης είναι ο απόλυτος αριθμός των CD4 λεμφοκυττάρων. Τα CD4 λεμφοκύτταρα χρησιμοποιούνται ευρέως στην παρακολούθηση ατόμων με HIV λοίμωξη, λαμβάνονται υπόψη στις αποφάσεις σχετικά με την θεραπευτική αγωγή καθώς και στην αξιολόγηση της επίδρασης θεραπείας στις κλινικές δοκιμές.

Το CD4 (Cluster of differentiation 4) λεμφοκύτταρο είναι μια γλυκοπρωτεΐνη που εκφράζεται στην επιφάνεια των T βοηθητικών κυττάρων, μονοκυττάρων, μακροφάγων και στα δενδριτικά κύτταρα. Ανακαλύφθηκε το 1970 και ήταν γνωστό ως leu-3 και T₄ πριν ονομαστεί CD4 το 1984. Τα CD4 λεμφοκύτταρα παίζουν σημαντικό ρόλο στην λειτουργία του ανοσοποιητικού συστήματος αλλά είναι επίσης και τα πρώτα κύτταρα που προσβάλλει ο ιός HIV-1 (Roitt et al, 1996). Πολλές επιδημιολογικές μελέτες έχουν δείξει ότι τα CD4 λεμφοκύτταρα μειώνονται σταδιακά κατά την πορεία της HIV λοίμωξης (Eyster et al, 1987; Detels et al, 1988; Munoz et al, 1988; Lang et al, 1989; Biggar et al, 1990; Phillips et al, 1991; Margolick et al, 1993; Phillips et al, 1994). Έχει αποδειχτεί ότι χαμηλές τιμές των CD4 λεμφοκυττάρων αποτελούν ισχυρή ένδειξη αυξημένου κινδύνου ανάπτυξης κλινικού AIDS. Η καταλληλότητα του αριθμού των CD4 λεμφοκυττάρων ως αναπληρωτή δείκτη σε κλινικές δοκιμές έχει αξιολογηθεί από πολλούς ερευνητές (Tsiatis et al, 1992; Choi et al, 1993; Moss et al, 1994; Henrard et al, 1995; Mellors et al, 1995; Mellors et al, 1996, O'Brein et al, 1996). Τα CD4 λεμφοκύτταρα μετρώνται βάση μιας τεχνικής που ονομάζεται κυτταρομετρία ροής και οι μετρήσεις εκφράζονται σε μονάδες (κύτταρα/μl). Ωστόσο έχει αποδειχθεί ότι οι μετρήσεις αυτές παρουσιάζουν σημαντική εγγενή μεταβλητότητα λόγω βιολογικών διαφορών και σφαλμάτων μέτρησης (Aledort et al, 1992; Hughes et al, 1994).

1.4 Σκοπός

Η εργασία αυτή θα επικεντρωθεί στην ανάλυση του απόλυτου αριθμού CD4 λεμφοκυττάρων κατά την φυσική ιστορία της HIV-1 λοίμωξης. Σε τέτοιες αναλύσεις, ο κύριος μηχανισμός περικοπής μετρήσεων, για άτομα που μολύνθηκαν μετά το 1996, έτος κατά το οποίο η συνδυασμένη αντιρετροϊκή αγωγή (cART) έγινε ευρέως διαθέσιμη, είναι η έναρξη αντιρετροϊκής θεραπείας. Το γεγονός ότι η απόφαση για έναρξη θεραπείας βασίζεται κυρίως σε ήδη παρατηρημένες μετρήσεις αριθμού CD4 λεμφοκυττάρων κάνει τον μηχανισμό αυτό, θεωρητικά τουλάχιστον, αγνοήσιμο. Προηγούμενες μελέτες προσομοίωσης έδειξαν ότι μικτά γραμμικά μοντέλα δίνουν όντως αμερόληπτες εκτιμήσεις σε τέτοιες περιπτώσεις. Σε αναλύσεις πραγματικών δεδομένων φάνηκε ότι η χρήση από κοινού μοντέλων για διόρθωση πληροφοριακής αποκοπής οδηγεί σε μεγάλες αποκλίσεις στην εκτίμηση του ρυθμού μεταβολής του αριθμού των CD4 λεμφοκυττάρων σε σχέση με τις αντίστοιχες εκτιμήσεις από τα απλούστερα μικτά γραμμικά μοντέλα. Σκοπός της παρούσας εργασίας θα είναι να διερευνηθεί η πιο πάνω διαφωνία καθώς και η γενικότερη επίδοση των προαναφερθέντων μοντέλων σε περιπτώσεις περικοπής μετρήσεων λόγω έναρξης αντιρετροϊκής θεραπείας. Η διερεύνηση θα βασιστεί σε αναλύσεις πραγματικών δεδομένων και σε μελέτες προσομοίωσης.

Η παρούσα διπλωματική χωρίζεται σε 6 μέρη. Στο κεφάλαιο 2 παρουσιάζονται μοντέλα για την ανάλυση διαχρονικών δεδομένων. Στο κεφάλαιο 3 παρουσιάζεται το θεωρητικό υπόβαθρο των ελλειπουσών τιμών, η επίδραση τους στη συμπερασματολογία καθώς και μέθοδοι διαχείρισης τους. Στο κεφάλαιο 4 παρουσιάζεται το Joint Multivariate Random Effect μοντέλο. Στο κεφάλαιο 5 γίνεται αναφορά στα πραγματικά δεδομένα της μελέτης AMACS και η ανάλυση τους με διαφορετικές μεθόδους. Στο κεφάλαιο 6 πραγματοποιούνται μελέτες προσομοίωσης κάτω από τέσσερα διαφορετικά σενάρια ελλειπουσών τιμών με στόχο τη σύγκριση ως προς την μεροληψία και την αποτελεσματικότητα των μοντέλων του κεφαλαίου 5. Τέλος, στο κεφάλαιο 7 παρουσιάζεται μια συζήτηση πάνω στα αποτελέσματα της μελέτης προσομοίωσης.

Κεφάλαιο 2: Μέθοδοι για την ανάλυση διαχρονικών δεδομένων

2.1 Εισαγωγή

Στις εφαρμοσμένες επιστήμες συνήθως συλλέγονται δεδομένα που υπόκεινται σε κάποιου είδους συσχέτιση όπως πολυδιάστατες παρατηρήσεις, κατά συστάδες (clustered) δεδομένα, επαναλαμβανόμενες μετρήσεις, διαχρονικά δεδομένα ή χωρικά δεδομένα. Στη παρούσα διπλωματική εργασία θα ασχοληθούμε με διαχρονικά δεδομένα που είναι επαναλαμβανόμενες μετρήσεις συσχετισμένες ως προς τον χρόνο.

Στην πράξη, οι επαναλαμβανόμενες μετρήσεις δεν είναι ισορροπημένα δεδομένα (balanced data) δηλαδή δεν υπάρχει ίδιος αριθμός μετρήσεων για όλες τις παρατηρήσεις ή/και οι μετρήσεις δεν γίνονται τις ίδιες χρονικές στιγμές. Ένας τρόπος προσέγγισης αυτών των δεδομένων είναι η χρήση γραμμικής παλινδρόμησης για την απεικόνιση των ατομικών-προφίλ των επαναλαμβανόμενων μετρήσεων. Σε πρώτο στάδιο, για κάθε άτομο το διάνυσμα των επαναλαμβανόμενων μετρήσεων συμπυκνώνεται σε ένα διάνυσμα εκτιμώμενων συντελεστών παλινδρόμησης. Σε δεύτερο στάδιο, χρησιμοποιούνται τεχνικές πολλαπλής παλινδρόμησης για να σχετίσουν αυτές τις εκτιμήσεις με γνωστούς παράγοντες όπως θεραπεία, χαρακτηριστικά κατά την έναρξη της μελέτης κ.ο.κ.

2.2 Ανάλυση σε δύο στάδια

2.2.1 1^ο Στάδιο

Έστω η τυχαία μεταβλητή Y_{ij} είναι η μεταβλητή απόκρισης για το άτομο i την χρονική στιγμή t_{ij} , $i=1,\dots,N$ και $j=1,\dots,n_i$. Το διάνυσμα $\mathbf{Y}_i$ είναι το n_i -διάστατο διάνυσμα όλων των επαναλαμβανόμενων μετρήσεων για το άτομο i , όπου $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \dots, Y_{in_i})'$. Το πρώτο στάδιο υποθέτει ότι το $\mathbf{Y}_i$ ικανοποιεί την γραμμική παλινδρόμηση:

$$Y_i = Z_i \beta_i + \varepsilon_i \quad (2.1)$$

όπου Z_i είναι ένας πίνακας ($n_i * q$) γνωστών παραγόντων, η οποία δείχνει πως εξελίσσεται η μεταβλητή απόκρισης με το χρόνο για το άτομο i . Επιπλέον, β_i είναι q -διάστατο διάνυσμα των άγνωστων ατομικών παραμέτρων παλινδρόμησης, και ε_i είναι ένα διάνυσμα των τυχαίων καταλοίπων ε_{ij} , $j=1, \dots, n_i$. Συνήθως υποθέτουμε ότι όλα τα κατάλοιπα είναι ανεξάρτητα και ακολουθούν κανονική κατανομή $\varepsilon_i \sim N(0, \sigma^2 I_{n_i})$.

2.2.2 2ο Στάδιο

Στο δεύτερο στάδιο, χρησιμοποιείται πολλαπλή παλινδρόμηση της μορφής:

$$\beta_i = K_i \beta + b_i \quad (2.2)$$

για να εξηγήσει την παρατηρηθείσα μεταβλητότητα μεταξύ των ατόμων σε σχέση με τους δικούς τους ατομικούς συντελεστές παλινδρόμησης β_i . Το K_i είναι ένας πίνακας ($q * p$) γνωστών παραγόντων και β είναι p -διάστατο διάνυσμα των άγνωστων παραμέτρων. Υποθέτουμε ότι τα b_i είναι ανεξάρτητα και ακολουθούν q -διάστατη κανονική κατανομή με μέσο το διάνυσμα 0 και πίνακα διασποράς-συνδιασποράς D $b_i \sim N(0, D)$.

2.3 Γραμμικό μοντέλο τυχαίων επιδράσεων

Για να συνδυάσουμε τα δυο μοντέλα από την ανάλυση δυο σταδίων αντικαθιστούμε το β_i στην εξίσωση (2.1) με την εξίσωση (2.2) και προκύπτει:

$$Y_i = X_i \beta + Z_i b_i + \varepsilon_i \quad (2.3)$$

Όπου $X_i = Z_i * K_i$ είναι ένας πίνακας ($n_i * p$) γνωστών παραγόντων. Η εξίσωση 2.3 ονομάζεται γραμμικό μοντέλο μικτών επιδράσεων με σταθερές επιδράσεις β και

ατομικές επιδράσεις b_i . Τα b_i υποθέτουμε ότι είναι τυχαία και συχνά ονομάζονται τυχαίες επιδράσεις. Στο μοντέλο η βασική υπόθεση είναι ότι το διάνυσμα των επαναλαμβανόμενων μετρήσεων του κάθε ατόμου ακολουθεί ένα γραμμικό μοντέλο παλινδρόμησης, του οποίου κάποιες παράμετροι είναι χαρακτηριστικές του πληθυσμού (άρα είναι κοινές για όλα τα άτομα) και άλλες είναι χαρακτηριστικές του κάθε ατόμου. Σε γενικές γραμμές, γραμμικό μοντέλο μικτών επιδράσεων ονομάζεται οποιοδήποτε μοντέλο το οποίο ικανοποιεί τα εξής (Laird and Ware 1982):

$$\left\{ \begin{array}{l} \mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\varepsilon}_i \\ \mathbf{b}_i \sim N(\mathbf{0}, D) \\ \boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \Sigma_i) \\ (\mathbf{b}_1, \dots, \mathbf{b}_N, \boldsymbol{\varepsilon}_1, \dots, \boldsymbol{\varepsilon}_N \text{ ανεξάρτητα}) \end{array} \right\} \quad (2.4)$$

Όπου D είναι ένας πίνακας ($q \times q$) με (i, j) στοιχεία $d_{ij} = d_{ji}$ και Σ_i είναι ένας $(n_i \times n_i)$ πίνακας διακύμανσης συνδιακύμανσης το οποίο εξαρτάται από τα i μόνο μέσω του n_i .

Από την 2.4 προκύπτει ότι δοθέντος των τυχαίων επιδράσεων $\mathbf{b}_i$, η κατανομή των $\mathbf{Y}_i$ είναι κανονική με μέσο το διάνυσμα $\mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i$ και με πίνακα διακύμανσης συνδιακύμανσης Σ_i . Έστω $f(\mathbf{y}_i | \mathbf{b}_i)$ και $f(\mathbf{b}_i)$ είναι οι αντίστοιχες συναρτήσεις πυκνότητας πιθανότητας για τα $\mathbf{Y}_i$ και $\mathbf{b}_i$. Τότε η περιθώρια συνάρτηση πυκνότητας πιθανότητας των $\mathbf{Y}_i$ δίνεται από την σχέση:

$$f(\mathbf{y}_i) = \int f(\mathbf{y}_i | \mathbf{b}_i) f(\mathbf{b}_i) d\mathbf{b}_i$$

η οποία είναι συνάρτηση πυκνότητας πιθανότητας μιας n_i -διάστατης κανονικής κατανομής με μέσο $\mathbf{X}_i \boldsymbol{\beta}$ και πίνακα διακύμανσης συνδιακύμανσης $V_i = \mathbf{Z}_i D \mathbf{Z}_i' + \Sigma_i$. Με αυτό τον τρόπο η περιθώρια μορφή που προκύπτει από το γραμμικό μικτό μοντέλο υπονοεί συγκεκριμένες δομές για το μέσο και τη διακύμανση οι οποίες εξαρτώνται άμεσα από τους πίνακες σχεδιασμού $\mathbf{X}_i$ και $\mathbf{Z}_i$. Είναι γνωστό ότι οι συντελεστές παλινδρόμησης $\boldsymbol{\beta}$ στα μοντέλα τυχαίων επιδράσεων αφορούν στο "μέσο άτομο" ("μέσο" με την έννοια ότι οι τυχαίες επιδράσεις $\mathbf{b}_i = \mathbf{0}$ για αυτό το άτομο) ενώ οι αντίστοιχες ποσότητες στα περιθώρια μοντέλα αφορούν πληθυσμιακές μέσες τιμές. Επισημαίνεται ότι σε κανονικά κατανεμημένα δεδομένα, όπως στα μοντέλα που μελετάμε εδώ, όπου η

συνδετική συνάρτηση της μέσης τιμής με το γραμμικό εκτιμητή είναι η συνάρτηση ταυτότητας, ταυτίζονται οι πληθυσμιακοί μέσοι με τις εκτιμήσεις για το τυπικό άτομο.

Πολύ συχνά στο μοντέλο (2.4) το Σ_i επιλέγεται έτσι ώστε να είναι ίσο με $\sigma^2 I_{n_i}$. Σε αυτή την περίπτωση το μοντέλο ονομάζεται μοντέλο υπό συνθήκη ανεξαρτησίας καθώς οι n_i μετρήσεις του ατόμου i θεωρούνται ανεξάρτητες δεδομένων των β και $\mathbf{b}_i$. Στις περιπτώσεις όπου παραβιάζεται η υπόθεση της υπό συνθήκη ανεξαρτησίας έχουν προταθεί επεκτάσεις που υποθέτουν ότι ο όρος ε_i μπορεί να διαχωριστεί σε δυο όρους $\varepsilon_{(1)i}$ και $\varepsilon_{(2)i}$ εκ των οποίων ο όρος $\varepsilon_{(2)i}$ αντικατοπτρίζει σειριακή συσχέτιση μεταξύ των διαδοχικών μετρήσεων του ατόμου i , υπονοώντας έτσι ότι ένα μέρος της παρατηρηθείσας ανά άτομο διαχρονικής εξέλιξης οφείλεται σε κάποια χρονικά μεταβαλλόμενη στοχαστική ανέλιξη. Επιπρόσθετα, ο όρος $\varepsilon_{(1)i}$ περιγράφει καθαρό σφάλμα μετρήσεων το οποίο θεωρείται ανεξάρτητο από το $\varepsilon_{(2)i}$. Το γραμμικό μικτό μοντέλο που προκύπτει στην περίπτωση αυτή είναι:

$$\left\{ \begin{array}{l} \mathbf{Y}_i = X_i \beta + Z_i \mathbf{b}_i + \varepsilon_{(1)i} + \varepsilon_{(2)i} \\ \mathbf{b}_i \sim N(\mathbf{0}, D) \\ \varepsilon_{(1)i} \sim N(\mathbf{0}, \sigma^2 I_{n_i}) \\ \varepsilon_{(2)i} \sim N(\mathbf{0}, \tau^2 H_i) \\ \mathbf{b}_1, \dots, \mathbf{b}_N, \varepsilon_{(1)1}, \dots, \varepsilon_{(1)N}, \varepsilon_{(2)1}, \dots, \varepsilon_{(2)N} \text{ ανεξάρτητα} \end{array} \right\} \quad (2.5)$$

Σε αυτά τα μοντέλα (Mansour et al., 1985; Diem kai Liukkonen, 1988; Diggle, 1988; Chi kai Reinsel, 1989; Rochon, 1992; Nunezanton kai Woodworth, 1994) συνήθως υποτίθεται μια συγκεκριμένη δομή για τον $n_i \times n_i$ πίνακα συσχέτισης H_i , δανεισμένη συχνά από το πεδίο ανάλυσης χρονοσειρών. Συγκεκριμένα υποτίθεται ότι το στοιχείο (j,k) του πίνακα H_i (h_{ijk}) εξαρτάται από την απόλυτη απόσταση μεταξύ των αντίστοιχων χρονικών στιγμών t_{ij} t_{ik} δηλαδή $h_{ijk} = g(|t_{ij} - t_{ik}|)$ όπου $g(\cdot)$ είναι φθίνουσα συνάρτηση με $g(0)=1$.

Ο Diggle et al. (1994) που αναφέρονται σε αυτά τα μοντέλα, αναφέρουν ότι στην πράξη η επίδραση της σειριακής συσχέτισης δεν μπορεί να διαχωριστεί εύκολα από αυτή των τυχαίων επιδράσεων και των σφαλμάτων μέτρησης. Το γεγονός αυτό οδηγεί συχνά σε

υπολογιστικά προβλήματα όταν υπάρχουν αρκετοί όροι τυχαίων επιδράσεων, σφάλμα μέτρησης και σειριακή συσχέτιση.

2.4 Εκτίμηση παραμέτρων στο γραμμικό μικτό μοντέλο

Στο κεφάλαιο 2.3 παρατηρήσαμε ότι το γραμμικό μοντέλο μικτών επιδράσεων (2.4) υπονοεί το περιθώριο μοντέλο:

$$Y_i \sim N(X_i\beta, Z_i D Z_i' + \Sigma_i) \quad (2.6)$$

Συνήθως η συμπερασματολογία στηρίζεται στην περιθώρια κατανομή των Y_i εκτός και αν η ανάλυση γίνεται στο πλαίσιο της Μπεϋζιανής στατιστικής (Gelman et al. 1995). Οι τυχαίες επιδράσεις b_i δεν λαμβάνονται υπόψη παρά μόνο έμμεσα μέσω της δομής του πίνακα διακύμανσης συνδιακύμανσης $V_i = Z_i D Z_i' + \Sigma_i$. Ωστόσο οι τυχαίες επιδράσεις b_i μπορούν να εκτιμηθούν εκ των υστέρων λαμβάνοντας υπόψη ότι τα Y_i ακολουθούν το μοντέλο (2.4).

Οι Laird και Ware (1982) περιέγραψαν μια ενοποιημένη μέθοδο για την εκτίμηση αυτών των μοντέλων, βασιζόμενοι σε μεθόδους μέγιστης πιθανοφάνειας, εμπειρική Μπεϋζιανή στατιστική και τον EM (Expectation-Maximization) αλγόριθμο. Ο Longford (1993) συνόψισε τις μεθόδους για την εκτίμηση των παραμέτρων του γραμμικού μικτού μοντέλου επικεντρώνοντας στους εξής αλγορίθμους: Newton-Raphson και Fisher scoring, Γενικευμένα Ελάχιστα Τετράγωνα, EM αλγόριθμος και μέθοδος Περιορισμένης Μέγιστης Πιθανοφάνειας (REML). Παρακάτω θα αναφερθούν οι μέθοδοι βασισμένες σε Μέγιστη Πιθανοφάνεια και οι επαναληπτικές μέθοδοι που βασίζονται στην θεωρία Γενικευμένων Ελαχίστων Τετραγώνων.

2.4.1 Μέθοδοι βασισμένοι στην Μέγιστη Πιθανοφάνεια ML/REML

Έστω α το διάνυσμα που περιλαμβάνει όλες τις παραμέτρους διακύμανσης ή συνδιακύμανσης του πίνακα $V_i = Z_i D Z_i' + \Sigma_i$, το α περιλαμβάνει $q(q+1)/2$ διαφορετικά στοιχεία του πίνακα D και όλες τις παραμέτρους του Σ_i . Επιπλέον, έστω $\theta = (\beta', \alpha')'$ το s-διάστατο διάνυσμα όλων των παραμέτρων στο μοντέλο (2.6) και $\theta = \theta_\beta \times \theta_\alpha$ ο αντίστοιχος παραμετρικός χώρος όπου $\theta_\beta = R^p$ και θ_α το σύνολο των τιμών του α για το οποίο ο πίνακας D και όλοι οι πίνακες Σ_i είναι θετικά (ημι-) ορισμένοι.

Η κλασσική προσέγγιση βασίζεται σε εκτιμητές που προκύπτουν από την μεγιστοποίηση της περιθώριας συνάρτησης πιθανοφάνειας:

$$L_{ML}(\theta) = \prod_{i=1}^N \left\{ (2\pi)^{-n_i/2} |V_i(\alpha)|^{-\frac{1}{2}} \times \exp \left(-\frac{1}{2} (Y_i - X_i \beta)' V_i^{-1}(\alpha) (Y_i - X_i \beta) \right) \right\} \quad (2.7)$$

ως προς το θ . Έστω ότι το α γνωστό, τότε ο εκτιμητής μέγιστης πιθανοφάνειας του β με δεδομένο το α δίνεται από την σχέση (Laird and Ware, 1982):

$$\hat{\beta}(\alpha) = (\sum_{i=1}^N X_i' W_i X_i)^{-1} \sum_{i=1}^N X_i' W_i y_i \quad (2.8)$$

όπου $W_i = V_i^{-1}$.

Όταν α είναι άγνωστο, αλλά υπάρχει εκτιμητής του $\hat{\alpha}$ μπορούμε να αντικαταστήσουμε τον πίνακα W_i στην εξίσωση (2.8) με τον πίνακα $\widehat{W}_i = V_i(\hat{\alpha})^{-1}$. Για την εκτίμηση του α χρησιμοποιείται η μέθοδος Μέγιστης Πιθανοφάνειας (ML) και η μέθοδος Περιορισμένης Μέγιστης Πιθανοφάνειας (REML).

Ο εκτιμητής Μέγιστης Πιθανοφάνειας του α προκύπτει από την μεγιστοποίηση της (2.7) αφού πρώτα αντικατασταθεί το β με βάση την (2.8). Αυτή η προσέγγιση θεωρείται σωστή όταν εκτιμούμε ταυτόχρονα το β και το α μεγιστοποιώντας ταυτόχρονα την από κοινού πιθανοφάνεια (2.7).

Η μέθοδος Περιορισμένης Μέγιστης Πιθανοφάνειας λαμβάνει επιπλέον υπόψη την απώλεια σε βαθμούς ελευθερίας κατά την εκτίμηση των σταθερών επιδράσεων β έτσι ώστε να γίνει αμερόληπτη εκτίμηση των παραμέτρων α μέσω μιας προσέγγισης αντιδιαστολών των υπολοίπων (error contrasts). Αρχικά συνδυάζουμε όλες τις N ατομικές παλινδρομήσεις της (2.4) σε ένα μοντέλο:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon} \quad (2.9)$$

όπου τα διανύσματα $\mathbf{Y}$, $\mathbf{b}$, $\boldsymbol{\varepsilon}$ και ο πίνακας $\mathbf{X}$ προκύπτουν συνδυάζοντας το ένα κάτω από το άλλο τα διανύσματα $\mathbf{Y}_i$, $\mathbf{b}_i$, $\boldsymbol{\varepsilon}_i$ και τους πίνακες $\mathbf{X}_i$ αντίστοιχα και $\mathbf{Z}$ είναι ο μπλοκ-διαγώνιος πίνακας με μπλοκ τα $\mathbf{Z}_i$. Η διάσταση του $\mathbf{Y}$ είναι $n = \sum_{i=1}^N n_i$. Η περιθώρια κατανομή του $\mathbf{Y}$ είναι $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, V(\mathbf{a}))$ με τον $V(\mathbf{a})$ να είναι μπλοκ διαγώνιος με μπλοκ τα V_i στην διαγώνιο και μηδέν οπουδήποτε αλλού. Ο εκτιμητής περιορισμένης μέγιστης πιθανοφάνειας για το $\mathbf{a}$ προκύπτει μεγιστοποιώντας την συνάρτηση πιθανοφάνειας ενός συνόλου από αντιδιαστολές σφαλμάτων (error contrasts) $\mathbf{U} = \mathbf{A}'\mathbf{Y}$ όπου $\mathbf{A}$ είναι ένας $(n \times (n-p))$ πλήρους τάξης πίνακας με τις στήλες του ανεξάρτητες από τις στήλες του $\mathbf{X}$. Το διάνυσμα $\mathbf{U}$ ακολουθεί κανονική κατανομή με μέσο το διάνυσμα $\mathbf{0}$ και πίνακα διακύμανσης-συνδιακύμανσης $\mathbf{A}'V(\mathbf{a})\mathbf{A}$

που πλέον δεν εξαρτάται από το $\boldsymbol{\beta}$, $\mathbf{U} \sim N(\mathbf{0}, \mathbf{A}'V(\mathbf{a})\mathbf{A})$.

2.4.2 Επαναληπτικές μέθοδοι βασισμένες στην θεωρία γενικευμένων ελαχίστων τετραγώνων IGLS/RIGLS

Η επαναληπτική μέθοδος γενικευμένων ελαχίστων τετραγώνων (Iterative Generalized Least Squares - IGLS) και η αντίστοιχη περιορισμένη μέθοδος (Restricted Iterative Generalized Least Squares - RIGLS) προτάθηκαν από τον Goldstein (1986, 1989). Οι εκτιμητές της μεθόδου IGLS είναι εκτιμητές μέγιστης πιθανοφάνειας αντίστοιχοι με αυτούς της μεθόδου ML που περιγράφηκαν στην ενότητα 2.4.1, αν και μόνο αν ισχύει η πολυδιάστατη κανονικότητα. Παρομοίως, η RIGLS μέθοδος αντιμετωπίζει την μεροληψία των παραμέτρων διακύμανσης-συνδιακύμανσης δίνοντας αμερόληπτους εκτιμητές αντίστοιχους με αυτούς της μεθόδου REML.

Έστω ένα γραμμικό μικτό μοντέλο δυο επιπέδων (διαχρονικές μετρήσεις εντός ατόμων) με δυο τυχαίες επιδράσεις (σταθερά και κλίση) και ομοσκεδαστικά υπόλοιπα επιπέδου-1. Στο μοντέλο (2.4) θεωρούμε ότι ο πίνακας $\mathbf{Z}_i$ έχει διαστάσεις $n_i \times 2$ με την πρώτη στήλη να αποτελείται από μονάδες και η δεύτερη από τους διαδοχικούς χρόνους μετρήσεων

$t_{ij} = (j = 1, 2, \dots, n_i)$ ενώ παράλληλα ισχύει ότι $\Sigma_i = \sigma^2 I_{n_i}$. Με βάση τα παραπάνω το μοντέλο γράφεται ως:

$$\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b} + \boldsymbol{\varepsilon}_i$$

με πίνακες σχεδιασμού:

$$\mathbf{X}_i = \begin{pmatrix} x_{i11} & x_{i12} & \dots & x_{i1p} \\ x_{i21} & x_{i22} & \dots & x_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{in_i1} & x_{in_i2} & \dots & x_{in_ip} \end{pmatrix}$$

και

$$\mathbf{Z}_i = \begin{pmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ \vdots & \vdots \\ 1 & t_{in_i} \end{pmatrix}$$

Οι δυο πρώτες στήλες του πίνακα $\mathbf{X}_i$ είναι ακριβώς ίδιες με τις αντίστοιχες του πίνακα $\mathbf{Z}_i$, ώστε οι αντίστοιχες σταθερές επιδράσεις να αναφέρονται στην πληθυσμιακή μέση αρχική τιμή και κλίση της εξαρτημένης μεταβλητής. Ο πίνακας διακύμανσης συνδιακύμανσης των τυχαίων επιδράσεων είναι ο:

$$\mathbf{D} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$$

ενώ ο πίνακας διακύμανσης-συνδιακύμανσης των υπολοίπων του πρώτου επιπέδου είναι ο $n_i \times n_i$:

$$\Sigma_i = \sigma^2 I_{n_i} = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix}$$

Ο πίνακας διακύμανσης-συνδιακύμανσης του $\mathbf{Y}_i$ είναι:

$$\text{Var}(\mathbf{Y}_i) = \mathbf{V}_i = \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i' + \sigma^2 I_{n_i}$$

όπου αν κάνουμε τις πράξεις προκύπτει:

$$V_i = \begin{pmatrix} \sigma_{11} + 2t_{i1}\sigma_{12} + t_{i1}^2\sigma_{22} + \sigma^2 & \sigma_{11} + (t_{i1} + t_{i2})\sigma_{12} + t_{i1}^2t_{i2}^2\sigma_{22} & \dots & \sigma_{11} + (t_{i1} + t_{in_i})\sigma_{12} + t_{i1}^2t_{in_i}^2\sigma_{22} \\ \sigma_{11} + (t_{i1} + t_{i2})\sigma_{12} + t_{i1}^2t_{i2}^2\sigma_{22} & \sigma_{11} + 2t_{i2}\sigma_{12} + t_{i2}^2\sigma_{22} + \sigma^2 & \dots & \sigma_{11} + (t_{i2} + t_{in_i})\sigma_{12} + t_{i2}^2t_{in_i}^2\sigma_{22} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{11} + (t_{i1} + t_{in_i})\sigma_{12} + t_{i1}^2t_{in_i}^2\sigma_{22} & \sigma_{11} + (t_{i2} + t_{in_i})\sigma_{12} + t_{i2}^2t_{in_i}^2\sigma_{22} & \dots & \sigma_{11} + 2t_{in_i}\sigma_{12} + t_{in_i}^2\sigma_{22} + \sigma^2 \end{pmatrix} \quad (2.10)$$

Αν $\mathbf{Y} = (\mathbf{Y}'_1, \dots, \mathbf{Y}'_N)'$ το συνολικό διάνυσμα όλων των μετρήσεων, ο αντίστοιχος πίνακας διακύμανσης-συνδιακύμανσης θα είναι :

$$Var(\mathbf{Y}) = V = \begin{pmatrix} V_1 & 0 & \dots & 0 \\ 0 & V_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & V_N \end{pmatrix}$$

Η μπλοκ διαγώνια δομή του V αντανακλά την εξάρτηση των μετρήσεων εντός του ίδιου ατόμου αλλά ταυτόχρονα την ανεξαρτησία μεταξύ των ατόμων.

Οι μέθοδοι IGLS και RIGLS είναι επαναληπτικές, δηλαδή για γνωστές παραμέτρους $\boldsymbol{\theta} = (\sigma_{11}, \sigma_{12}, \sigma_{22}, \sigma^2)$ οι παράμετροι $\boldsymbol{\beta}$ εκτιμώνται στο πρώτο βήμα και στο δεύτερο βήμα για γνωστό $\boldsymbol{\beta}$ το $\boldsymbol{\theta}$ επανεκτιμάται. Τα δυο αυτά βήματα επαναλαμβάνονται μέχρι να επιτευχθεί η σύγκλιση. Για γνωστό $\boldsymbol{\theta}$, άρα για γνωστό V η εκτίμηση του $\boldsymbol{\beta}$ γίνεται ως:

$$\widehat{\boldsymbol{\beta}} = (X'V^{-1}X)^{-1}X'V^{-1}\mathbf{Y} \quad (2.11)$$

με πίνακα διακύμανσης-συνδιακύμανσης $(X'V^{-1}X)^{-1}$. Στο δεύτερο βήμα για γνωστό πλέον $\boldsymbol{\beta}$, για να εκτιμηθεί το $\boldsymbol{\theta}$ αρχικά κατασκευάζεται το διάνυσμα των αδρών υπολοίπων του σταθερού μέρους του μοντέλου $\tilde{\mathbf{Y}}_i = \mathbf{Y}_i - X_i\widehat{\boldsymbol{\beta}}$ και στη συνέχεια ο παρατηρηθείς πίνακας διαγωνίων γινομένων των αδρών υπολοίπων $Y_i^* = \tilde{\mathbf{Y}}_i\tilde{\mathbf{Y}}_i'$ του i ατόμου, με αναμενόμενη τιμή $E(Y_i^*) = V_i$. Έστω $\mathbf{Y}^{**}$ το διάνυσμα που προκύπτει από την κατακόρυφη διάταξη των στηλών του $\mathbf{Y}_i^*$ τότε προκύπτει:

$$\mathbf{Y}_i^{**} = (r_{i1}^2, r_{i1}r_{i2}, \dots, r_{i1}r_{in_i}, r_{i2}r_{i1}, r_{i2}^2, r_{i2}r_{i3}, \dots, r_{i2}r_{in_i}, \dots, r_{in_i}r_{i1}, r_{in_i}r_{i2}, \dots, r_{in_i}^2)$$

όπου το r_{ij} είναι το j-οστό αδρό υπόλοιπο του i-οστού ατόμου. Η αναμενόμενη τιμή του $\mathbf{Y}^{**}$ μπορεί να γραφτεί ως γραμμικός συνδυασμός των στοιχείων του $\boldsymbol{\theta}$, $E(\mathbf{Y}^{**}) = \mathbf{Z}^*\boldsymbol{\theta}$ όπου

Z^* ο πίνακας σχεδιασμού που συνδέει το Y^{**} με τον V . Οι παράμετροι του θ μπορούν να εκτιμηθούν μέσω της θεωρίας γενικευμένων ελαχίστων τετραγώνων:

$$\hat{\theta} = (Z^{*'}V^{*-1}Z^*)^{-1}Z^{*'}V^{*-1}Y^{**} \quad (2.12)$$

όπου $V^* = \{V_i \otimes V_i\} \otimes I_N$ όπου $\otimes$ είναι το γινόμενο Kronecker.

Η μέθοδος επαναλαμβάνεται μεταξύ των δυο βημάτων (2.11) και (2.12) μέχρι να συγκλίνει, δηλαδή μέχρι η σχετική διαφορά σε όλες τις παραμέτρους μεταξύ δυο επαναλήψεων να είναι μικρότερη από ένα καθορισμένο όριο π.χ. 1%.

Για το μοντέλο που υποθέσαμε παραπάνω ο πίνακας Z^* για το άτομο θα έχει την μορφή:

$$Z_i^* = \begin{pmatrix} 1 & 2t_{i1} & t_{i1}^2 & 1 \\ 1 & t_{i1} + t_{i2} & t_{i1}t_{i2} & 0 \\ & \vdots & \vdots & \vdots \\ \vdots & t_{i1} + t_{in_1-1} & t_{i1}t_{in_1-1} & \vdots \\ 1 & t_{i1} + t_{in_1} & t_{i1}t_{in_1} & 0 \\ 1 & t_{i1} + t_{i2} & t_{i1}t_{i2} & 0 \\ 1 & 2t_{i2} & t_{i2}^2 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_{i2} + t_{in_1-1} & t_{i2}t_{in_1-1} & 0 \\ 1 & t_{i2} + t_{in_1} & t_{i2}t_{in_1} & 0 \\ 1 & t_{i1} + t_{in_i} & t_{i1}t_{in_i} & 0 \\ 1 & t_{i2} + t_{in_i} & t_{i2}t_{in_i} & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_{in_i-1} + t_{in_1} & t_{in_i-1}t_{in_1} & 0 \\ 1 & 2t_{in_i} & t_{in_i}^2 & 1 \end{pmatrix}$$

Η μορφή του Z_i^* καθορίζεται από τον πίνακα V_i . Διατάσσοντας τις στήλες του V_i κατακόρυφα είναι εμφανές ότι οι τέσσερις στήλες του Z_i^* αντιστοιχούν στους πολλαπλασιαστές των παραμέτρων του $\theta = (\sigma_{11}, \sigma_{12}, \sigma_{22}, \sigma^2)$.

Εφόσον οι εκτιμήσεις της μεθόδου IGLS, υπό την προϋπόθεση της πολυδιάστατης κανονικής κατανομής για τις τυχαίες επιδράσεις και τα υπόλοιπα πρώτου επιπέδου, αντιστοιχούν σε εκτιμήσεις μέγιστης πιθανοφάνειας εμφανίζουν και αυτές πρόβλημα μεροληψίας στις εκτιμήσεις των παραμέτρων διακύμανσης-συνδιακύμανσης. Για να

έχουμε αμερόληπτες εκτιμήσεις μπορεί να χρησιμοποιηθεί η μέθοδος RIGLS οι εκτιμήσεις της οποίας είναι ισοδύναμες με εκείνες της μεθόδου περιορισμένης μέγιστης πιθανοφάνειας (Goldstein 1989). Σύμφωνα με την μέθοδο RIGLS σε κάθε επανάληψη τροποποιείται η εκτίμηση του θ με την πρόσθεση του όρου $X_i(X_i'\widehat{V_i^{-1}}X_i)'X_i'$ στον πίνακα διαγωνίων γινομένων των αδρών υπολοίπων Y_i^* και έτσι παίρνουμε αμερόληπτες εκτιμήσεις για τις παραμέτρους διακύμανσης-συνδιακύμανσης.

Κεφάλαιο 3: Ελλείπουσες τιμές (Missing Data)

3.1 Εισαγωγή

Συχνά παρατηρούνται σύνολα δεδομένων που είναι ελλιπή με την έννοια ότι δεν πραγματοποιούνται όλες οι προγραμματισμένες μετρήσεις ανά άτομο ή ανά χρονική στιγμή. Αυτό παρατηρείται ιδιαίτερα σε μελέτες όπου συμμετέχουν ανθρώπινα όντα. Ωστόσο, σε όλες τις μορφές ερευνών υπάρχουν κατά κανόνα ελλείπουσες τιμές και η πιθανότητα ύπαρξής τους είναι ιδιαίτερα υψηλή σε διαχρονικές επιδημιολογικές μελέτες. Η αποχώρηση ατόμων από την μελέτη ή η απουσία τους από προκαθορισμένες ενδιάμεσες επισκέψεις στην κλινική οδηγούν σε ελλείπουσες τιμές. Ο μηχανισμός που οδηγεί στην δημιουργία ελλειπουσών τιμών πρέπει να λαμβάνεται υπόψη διότι παρουσία αυτών η επιλογή των μεθόδων ανάλυσης παίζει σημαντικό ρόλο στη συμπερασματολογία.

3.2 Ορολογία

Η παρακάτω ορολογία στηρίζεται στο μοντέλο που εισήγαγαν οι Rubin (1976) και Little and Rubin (2002). Έστω ότι για κάθε ανεξάρτητο άτομο $i=1, \dots, N$ στη μελέτη έχουμε σχεδιάσει να συλλέξουμε μια σειρά μετρήσεων $Y_{ij} = (j = 1, \dots, n_i)$. Το διάνυσμα $Y_{ij} = (Y_{i1}, \dots, Y_{in_i})'$ περιλαμβάνει όλες τις μετρήσεις ενός δείκτη για το άτομο i , ο πίνακας X_i είναι ο $n_i \times p$ πίνακας σχεδιασμού των επεξηγηματικών μεταβλητών και το θ είναι ένα διάνυσμα άγνωστων παραμέτρων που συνδέουν το Y_i με το X_i . Το διάνυσμα $R_i = (R_{i1}, \dots, R_{in_i})$ είναι ένα διάνυσμα $n_i \times 1$ το οποίο δηλώνει την παρουσία ελλειπουσών τιμών. Για κάθε επίσκεψη j ορίζουμε:

$$R_{ij} = \begin{cases} 1 & \text{αν το } Y_{ij} \text{ παρατηρηθεί} \\ 0 & \text{αλλιώς} \end{cases}$$

Έστω ϕ ένα διάνυσμα άγνωστων παραμέτρων για το μοντέλο της μη ανταπόκρισης και W_i ο πίνακας σχεδιασμού που συνδέει το R_i με το ϕ . Η πολυδιάστατη συνάρτηση

πυκνότητας πιθανότητας του $\mathbf{Y}$ δεδομένων των X και θ είναι: $f(\mathbf{Y}|X, \theta)$, ενώ η πολυδιάστατη συνάρτηση πιθανότητας του $\mathbf{R}$ δεδομένων των $\mathbf{Y}$, X , W και φ είναι: $f(\mathbf{R}|\mathbf{Y}, X, W, \varphi)$.

Μπορούμε να διαχωρίσουμε το πλήρες διάνυσμα της εξαρτημένης μεταβλητής $\mathbf{Y}$ σε δυο υποδιανύσματα $\mathbf{Y}_o$ και $\mathbf{Y}_m$, το πρώτο περιλαμβάνει τις παρατηρηθείσες (observed) και το δεύτερο της ελλείπουσες (missing) τιμές. Το διάνυσμα $\mathbf{R}_i$ και η διαδικασία που οδηγεί στην δημιουργία των $\mathbf{R}_i$ αναφέρονται ως μηχανισμός παραγωγής ελλειπουσών τιμών. Τα παρατηρηθέντα δεδομένα είναι τα $\mathbf{Y}_o$ και $\mathbf{R}$ και η συνάρτηση πυκνότητας πιθανότητας των παρατηρηθέντων δεδομένων μπορεί να βρεθεί ολοκληρώνοντας τις μη παρατηρηθείσες ποσότητες στην έκφραση της από κατανομής $f(\mathbf{Y}, \mathbf{R})$:

$$f(\mathbf{Y}_o, \mathbf{R}|X, W, \theta, \varphi) = \int_{\mathbf{R}} f(\mathbf{Y}_o, \mathbf{R}|X, W, \theta, \varphi) d\mathbf{Y}_m \quad (3.2.1)$$

Η οποία μπορεί να γραφεί ως:

$$\begin{aligned} f(\mathbf{Y}_o, \mathbf{R}|X, W, \theta, \varphi) &= \int_{\mathbf{R}} f(\mathbf{Y}_o, \mathbf{R}|X, W, \theta, \varphi) d\mathbf{Y}_m \\ &= \int_{\mathbf{R}} f(\mathbf{Y}|X, \theta) f(\mathbf{R}|\mathbf{Y}, X, W, \varphi) d\mathbf{Y}_m \end{aligned} \quad (3.2.2)$$

3.3 Μηχανισμοί Ελλειπουσών Τιμών

Οι ελλείπουσες τιμές χωρίζονται σε τρεις κατηγορίες:

3.3.1 Εντελώς τυχαίες (Missing Completely At Random - MCAR)

Κάτω από τον MCAR μηχανισμό, η πιθανότητα μια παρατήρηση να είναι missing είναι ανεξάρτητη από το διάνυσμα της εξαρτημένης μεταβλητής $\mathbf{Y}$.

$$f(\mathbf{R}|\mathbf{Y}, X, W, \varphi) = f(\mathbf{R}|X, W, \varphi) \quad (3.3.1)$$

Σε αυτή την περίπτωση οποιοσδήποτε τρόπος και αν επιλεγεί για την ανάλυση των δεδομένων, ο μηχανισμός δημιουργίας ελλειπουσών τιμών μπορεί να αγνοηθεί. Παράδειγμα MCAR ελλειπουσών τιμών αποτελεί η κλιμακωτή ένταξη ατόμων σε κάποια μελέτη (Lange et al., 1989). Γενικά αν οι ελλείπουσες τιμές δημιουργούνται λόγω του σχεδιασμού της μελέτης και όχι λόγω χαρακτηριστικών των ερευνητικών μονάδων είναι πολύ πιθανό οι ελλείπουσες τιμές να είναι MCAR (Little, 1995).

3.3.2 Τυχαιές (Missing At Random - MAR)

Κάτω από τον μηχανισμό MAR, η πιθανότητα μια παρατήρηση να είναι missing εξαρτάται από το παρατηρηθέν τμήμα του $\mathbf{Y}$ ($\mathbf{Y}_o$) αλλά όχι από το μη παρατηρηθέν ($\mathbf{Y}_m$).

$$f(\mathbf{R}|\mathbf{Y}, \mathbf{X}, \mathbf{W}, \varphi) = f(\mathbf{R}|\mathbf{Y}_o, \mathbf{X}, \mathbf{W}, \varphi) \quad (3.3.2)$$

Σε αυτή την περίπτωση μπορεί να αγνοηθεί ο μηχανισμός δημιουργίας των ελλειπουσών τιμών αλλά όχι οι ίδιες ή τα άτομα που έχουν ελλείπουσες τιμές. Χαρακτηριστικό παράδειγμα τυχαίου μηχανισμού δημιουργίας ελλειπουσών τιμών (MAR) αποτελούν περιπτώσεις όπου τα άτομα αποχωρούν από την μελέτη αφού η υπό εξέταση εξαρτημένη μεταβλητή φτάσει σε κάποια προκαθορισμένα επίπεδα (Murray & Findlay, 1988)

3.3.3 Μη αγνοήσιμες (Missing Not At Random - MNAR)

Κάτω από τον μηχανισμό MNAR, η πιθανότητα μια παρατήρηση να είναι missing εξαρτάται από το μη παρατηρηθέν τμήμα του $\mathbf{Y}$ ($\mathbf{Y}_m$).

Δεν είναι δυνατή η απλοποίηση της από κοινού κατανομής. Η από κοινού κατανομή των παρατηρηθέντων μετρήσεων και του μηχανισμού δημιουργίας ελλειπουσών τιμών γράφεται ως:

$$f(Y_o, R|X, W, \theta, \varphi) = \int f(Y|X, \theta)f(R|Y, W, \varphi)dY_m \quad (3.3.3)$$

Εάν ο μηχανισμός δημιουργίας ελλειπουσών τιμών είναι MNAR τότε δεν πρέπει να αγνοηθεί και πρέπει να εφαρμόζονται κατάλληλες μέθοδοι στατιστικής ανάλυσης.

3.4 Μέθοδοι διαχείρισης των ελλειπουσών τιμών

3.4.1 Ανάλυση πλήρων δεδομένων (Complete Case Analysis)

Στην ανάλυση συμπεριλαμβάνονται τα άτομα για τα οποία έχουν καταγραφεί όλες οι μετρήσεις και παραλείπονται τα άτομα που έχουν έστω και μια ελλείπουσα τιμή οποιαδήποτε χρονική στιγμή. Τα πλεονεκτήματα της μεθόδου είναι α) η απλότητα εφαρμογής της εφόσον μπορούν να χρησιμοποιηθούν γνωστές στατιστικές μέθοδοι χωρίς παραλλαγές και η συγκρισιμότητα των μονοπαραγοντικών στατιστικών μέτρων εφόσον όλες έχουν υπολογιστεί βασισμένες σε ένα συγκεκριμένο δείγμα μετρήσεων. Τα μειονεκτήματα της μεθόδου είναι η απώλεια πληροφορίας που οδηγεί συχνά σε μη αποτελεσματικούς εκτιμητές και σε μείωση ισχύος. Επιπλέον, σε περίπτωση MAR μηχανισμού ελλειπουσών τιμών η μέθοδος αυτή οδηγεί σε μεροληπτικές εκτιμήσεις. Τέλος, η μέθοδος αυτή είναι συνεπής μόνο σε περίπτωση MCAR μηχανισμού ελλειπουσών τιμών.

3.4.2 Ανάλυση διαθέσιμων κατά περίπτωση δεδομένων (Available Case Analysis)

Σε αυτή την κατηγορία ανήκουν διάφορες μέθοδοι που χρησιμοποιούν την διαθέσιμη πληροφορία για να εκτιμήσουν τον μέσο και την συνδιακύμανση. Αυτή η μέθοδος είναι πιο αποτελεσματική από την complete case μέθοδο διότι χρησιμοποιείται παραπάνω πληροφορία. Μια δημοφιλής μέθοδος είναι η pair-wise deletion method όπου ο πίνακας

διακύμανσης συνδιακύμανσης (ή συσχέτισης) υπολογίζεται έτσι ώστε κάθε στοιχείο του να βασίζεται στο συνολικό αριθμό περιπτώσεων με πλήρη δεδομένα για κάθε ζεύγος μεταβλητών. Το βασικό μειονέκτημα είναι ότι το μέγεθος παρατηρήσεων κάθε μεταβλητής διαφέρει αναλόγως με το μοτίβο των ελλειπουσών τιμών και είναι έγκυρη μόνο εάν ο μηχανισμός δημιουργίας ελλειπουσών τιμών είναι MCAR.

3.4.3 Υποκατάσταση ελλειπουσών τιμών (Imputation Methods)

Ένας εναλλακτικός τρόπος για να αποκτήσουμε ένα πλήρες σετ δεδομένων είναι αντί να διαγράψουμε τις ελλείπουσες τιμές να τις συμπληρώσουμε με κάποιο τρόπο έτσι ώστε να έχουμε την δυνατότητα να τις χρησιμοποιήσουμε σε κλασσικές μεθόδους ανάλυσης. Η χρήση μεθόδων υποκατάστασης (imputation) επιφυλάσσει διάφορους κινδύνους. Πρώτον, το μοντέλο υποκατάστασης μπορεί να είναι λανθασμένο με αποτέλεσμα να λάβουμε μεροληπτικές εκτιμήσεις. Δεύτερον, ακόμα και αν το μοντέλο υποκατάστασης είναι σωστό υπάρχει αβεβαιότητα που προέρχεται από τον μηχανισμό δημιουργίας ελλειπουσών τιμών. Οι περισσότερες μέθοδοι απαιτούν ο μηχανισμός δημιουργίας ελλειπουσών τιμών να είναι MCAR, ενώ κάποιες όπως η LCOF¹ απαιτεί ακόμα περισσότερες και ισχυρότερες υποθέσεις για να εφαρμοστεί.

3.4.3.1 Υποκατάσταση ελλειπουσών τιμών χωρίς συνθήκες (Unconditional mean imputation)

Κάθε ελλείπουσα τιμή αντικαθιστάται από την μέση τιμή των παρατηρηθέντων τιμών της ίδιας μεταβλητής. Ο όρος unconditional αναφέρεται στο γεγονός ότι δεν χρησιμοποιείται καμία πληροφορία σχετικά με το άτομο στο οποίο γίνεται η υποκατάσταση (imputation) της ελλείπουσας τιμής, δηλ οι τιμές που έχουν υποκατασταθεί (imputed values) δεν σχετίζονται με τις άλλες τιμές του ίδιου ατόμου

¹ LCOF: Last observation carried forward

(Little and Rubin 2002). Και σε αυτή την περίπτωση απαιτείται ο μηχανισμός δημιουργίας ελλειπουσών τιμών να είναι MCAR.

3.4.3.2 Μέθοδος υποκατάστασης μέσης τιμής υπό συνθήκη (*Conditional mean imputation – Buck’s method*)

Είναι περισσότερο περίπλοκη μέθοδος σε σχέση με την unconditional mean imputation. Στο πρώτο στάδιο εκτιμάται η μέση τιμή και ο πίνακας διακύμανσης συνδιακύμανσης των πλήρων δεδομένων. Στο δεύτερο στάδιο υπολογίζεται ο δεσμευμένος μέσος από την παλινδρόμηση των ελλειπουσών παραγόντων στις παρατηρημένες τιμές και αντικαθιστάται για τις ελλείπουσες τιμές. Η μέθοδος είναι έγκυρη αν ο μηχανισμός δημιουργίας ελλειπουσών τιμών είναι MCAR (Buck 1960) ενώ ο Little and Rubin (2002) έδειξαν ότι η μέθοδος είναι έγκυρη ακόμα και σε μερικές περιπτώσεις MAR μηχανισμού δημιουργίας ελλειπουσών τιμών.

3.4.3.3 Υποκατάσταση Hot-Deck

Η μέθοδος Hot Deck imputation περιλαμβάνει την υποκατάσταση των ελλειπουσών τιμών μιας ή περισσότερων μεταβλητών ενός ατόμου που δεν ανταποκρίθηκε, με τις παρατηρηθείσες τιμές από ένα άτομο που ανταποκρίθηκε και έχει παρόμοια χαρακτηριστικά με τον πρώτο. Σε μερικές περιπτώσεις διαλέγεται τυχαία ένα άτομο από ένα σύνολο ατόμων, ενώ σε άλλες περιπτώσεις εντοπίζουμε ένα άτομο, τον πιο κοντινό σε σχέση με τα χαρακτηριστικά του ατόμου για τον οποίο έχουμε ελλείπουσες τιμές και συμπληρώνουμε τις ελλείπουσες τιμές από αυτό το άτομο τον οποίο ονομάζουμε ως “κοντινός γείτονας” (Andridge and Little, 2010).

3.4.3.4 Τελευταία παρατήρηση προβαλλόμενη προοπτικά (Last Observation Carried Forward (LCOF))

Σύμφωνα με αυτή την προσέγγιση κάθε ελλείπουσα τιμή αντικαθιστάται από την τελευταία παρατηρηθείσα τιμή του ίδιου ατόμου. Για να εφαρμοστεί αυτή η μέθοδος πρέπει να ισχύουν πολύ ισχυρές προϋποθέσεις. Πρώτον, προϋποθέτει ότι η τιμή της μεταβλητής απόκρισης δεν μεταβάλλεται μετά την εμφάνιση της ελλείπουσας τιμής. Επιπρόσθετα, υπερεκτιμάει την ακρίβεια των εκτιμήσεων αφού χειρίζεται τις τιμές που έχουν συμπληρωθεί ύστερα από υποκατάσταση (imputation) αλλά και τις παρατηρηθείσες τιμές επί ίσοις όροις. Είναι σημαντικό να επισημανθεί ότι η LCOF δεν δίνει εσφαλμένες εκτιμήσεις ακόμα και υπό MCAR ελλείπουσες τιμές σε αντίθεση με την complete case προσέγγιση.

3.4.3.5 Πολλαπλές υποκαταστάσεις (Multiple Imputation (MI))

Αυτή είναι η πιο δημοφιλής μέθοδος για να διαχειριστεί κανείς ελλείπουσες τιμές. Η κύρια ιδέα πίσω από το multiple imputation είναι να αντικατασταθεί κάθε ελλείπουσα τιμή από ένα σετ M εύλογων τιμών. Κάθε ελλείπουσα τιμή αντικαθιστάται από δυο ή περισσότερες αποδεκτές τιμές απεικονίζοντας μια κατανομή πιθανοτήτων. Η διαδικασία αυτή αποτελείται από 3 φάσεις:

1. Τα ελλιπή δεδομένα συμπληρώνονται M φορές για να δημιουργηθούν M πλήρη σετ δεδομένων.
2. Τα M πλήρη σετ δεδομένων αναλύονται χρησιμοποιώντας κοινές διαδικασίες.
3. Τα αποτελέσματα από τα M πλήρη σετ δεδομένων συνδυάζονται για το τελικό συμπέρασμα.

Το πλεονέκτημα της μεθόδου είναι ότι εφόσον δημιουργηθεί το σετ με τα δεδομένα που έχουν υποκατασταθεί, η ανάλυση μπορεί να πραγματοποιηθεί σε οποιοδήποτε στατιστικό πακέτο. Το multiple imputation μπορεί να είναι αρκετά αποτελεσματικό ακόμα και αν ο

αριθμός των imputation είναι σχετικά μικρός, ειδικά όταν η διακύμανση μεταξύ των υποκαταστάσεων δεν είναι πολύ μεγάλη. Υπάρχει αβεβαιότητα που οφείλεται στις ελλείπουσες τιμές επειδή η ανάλυση δεν διακρίνει μεταξύ των παρατηρηθέντων και των τιμών που έχουν υποκατασταθεί, ωστόσο αυτή η αβεβαιότητα διορθώνεται με τον κλασσικό εκτιμητή της διακύμανσης σε MI ανάλυση. Σε σχέση με την μέθοδο single imputation το μόνο μειονέκτημα είναι ότι απαιτείται περισσότερη δουλειά για να δημιουργηθούν τα M πλήρη σετ δεδομένων και να αναλυθούν τα αποτελέσματα.

3.5 Επίδραση ελλειπουσών τιμών στην συμπερασματολογία

Ο μηχανισμός δημιουργίας ελλειπουσών τιμών μπορεί να επηρεάσει τη συμπερασματολογία σχετικά με την εξαρτημένη μεταβλητή (Laird, 1988). Στην περίπτωση που οι ελλείπουσες τιμές είναι εντελώς τυχαίες (MCAR), όλες σχεδόν οι στατιστικές μέθοδοι δίνουν έγκυρα αποτελέσματα (Laird & Ware, 1982; Lindstrom & Bates, 1988; Stiratelli et al., 1984; Breslow, 1984; Harville & Mee, 1984). Ωστόσο, κάθε στατιστική ανάλυση που χρησιμοποιείται υπό αυτό τον μηχανισμό πρέπει να υλοποιείται με ιδιαίτερη προσοχή διότι υπάρχει ανομοιομορφία στον αριθμό των μετρήσεων κάθε ατόμου.

Σε αυτή την περίπτωση επειδή τα $\mathbf{Y}_0$ και $\mathbf{R}$ είναι ανεξάρτητα η εξίσωση 3.2.2 μπορεί να γραφτεί ως:

$$f(\mathbf{Y}_0, \mathbf{R} | X, W, \theta, \varphi) = f(\mathbf{R} | X, W, \varphi) \int_R f(\mathbf{Y} | X, \theta) d\mathbf{Y}_m \quad (3.5.1)$$

και κατά συνέπεια η συνάρτηση πυκνότητας πιθανότητας του $\mathbf{Y}$ είναι η συνήθης περιθώρια συνάρτηση πυκνότητας πιθανότητας:

$$f(\mathbf{Y}_0 | X, \theta) = \int_R f(\mathbf{Y} | X, \theta) d\mathbf{Y}_m \quad (3.5.2)$$

Χαρακτηριστική ιδιότητα των μεθόδων που βασίζονται στη μεγιστοποίηση της πιθανοφάνειας είναι ότι εφαρμόζονται εύκολα και σε ελλειπή δεδομένα (Dempster et al., 1977; Meng & Rubin, 1991, 1993; Gilks et al., 1993). Επιπλέον συνεπείς εκτιμητές, υπό MCAR μηχανισμό δημιουργίας ελλειπουσών τιμών, δίνουν ακόμα και μέθοδοι που δεν βασίζονται στην πιθανοφάνεια όπως αυτές που προτάθηκαν από τους Liang & Zieger, (1986); Stram et al., (1988); Wei & Stram, (1988) ή μέθοδοι που βασίζονται σε συνοπτικές στατιστικές (Dawson & Lagakos, 1993; Frison & Pocock, 1992).

Στην περίπτωση όπου η ελλείπουσες τιμές είναι τυχαίες (MAR), η συνάρτηση πυκνότητας πιθανότητας των παρατηρηθέντων δεδομένων παραγοντοποιείται σε δυο ξεχωριστούς όρους, ο ένας αντιστοιχεί στο παρατηρηθέν τμήμα της $\mathbf{Y}$ ($\mathbf{Y}_0$) και ο άλλος αντιστοιχεί στον μηχανισμό παραγωγής ελλειπουσών τιμών (Little & Rubin, 1987; Laird, 1988, Little, 1995).

$$f(\mathbf{Y}_0|\mathbf{X}, \boldsymbol{\theta})f(\mathbf{R}|\mathbf{Y}_0, \mathbf{X}, \mathbf{W}, \varphi) \quad (3.5.3)$$

Με αυτό τον τρόπο, οι μέθοδοι μέγιστης πιθανοφάνειας μπορούν να εκτιμήσουν το $\boldsymbol{\theta}$ αγνοώντας τον μηχανισμό που δημιουργεί τις ελλείπουσες τιμές $f(\mathbf{R}|\mathbf{Y}_0, \mathbf{X}, \mathbf{W}, \varphi)$. Για αυτό το λόγο, ο μηχανισμός παραγωγής τους ονομάζεται *αγνοήσιμος* (ignorable), δηλαδή δεν χρειάζεται να εφαρμόσει κανείς ένα μοντέλο για τις ελλείπουσες τιμές για να πάρει έγκυρες εκτιμήσεις για το $\boldsymbol{\theta}$. Επομένως, επειδή ο μηχανισμός ονομάζεται αγνοήσιμος αυτό δεν σημαίνει ότι οι μέθοδοι που δεν βασίζονται στην πιθανοφάνεια των δεδομένων δίνουν έγκυρες εκτιμήσεις. Πρέπει να υπογραμμιστεί ότι αυτό που μπορεί να αγνοηθεί είναι ο μηχανισμός δημιουργίας ελλειπουσών τιμών αλλά όχι οι ίδιες οι ελλείπουσες τιμές ή τα άτομα με ελλείπουσες τιμές.

Οι μέθοδοι που δεν βασίζονται στην μέγιστη πιθανοφάνεια μπορεί να οδηγήσουν σε μεροληπτικούς εκτιμητές όταν οι ελλείπουσες τιμές είναι τυχαίες (MAR). Από τον δεύτερο όρο της έκφρασης 3.5.3 είναι φανερό ότι τα $\mathbf{Y}_0$ και $\mathbf{R}$ δεν είναι ανεξάρτητα, άρα η συνάρτηση πυκνότητας πιθανότητας του $\mathbf{Y}_0$ δεδομένου του $\mathbf{R}$ δεν ισούται με $f(\mathbf{Y}_0|\mathbf{X}, \boldsymbol{\theta})$, αλλά εξαρτάται από τον μηχανισμό παραγωγής των ελλειπουσών τιμών.

Αυτή είναι η αιτία για την οποία ο ασυμπτωτικός πίνακας διακύμανσης-συνδιακύμανσης του εκτιμητή μέγιστης πιθανοφάνειας του θ δεν δίνεται με τον συνήθη τρόπο, δηλαδή με τον αντίστροφο του αναμενόμενου πίνακα πληροφορίας του Fisher. Η Laird (1988) πρότεινε οι εκτιμητές ακρίβειας και οι αντίστοιχες στατιστικές συναρτήσεις ελέγχων να βασίζονται στον παρατηρηθέντα και όχι στον αναμενόμενο πίνακα πληροφορίας. Ένα μειονέκτημα των μεθόδων μέγιστης πιθανοφάνειας είναι η ευαισθησία τους στον ορισμό του μοντέλου, καθώς αυτό χρησιμοποιείται για να υποκαταστήσει τις ελλείπουσες τιμές με βάση την συνάρτηση $f(Y_o|X, \theta)$. Για αυτό, συνιστάται η χρήση τεχνικών επιβεβαίωσης του μοντέλου.

Το γραμμικό μικτό μοντέλο, το οποίο βασίζεται στην πιθανοφάνεια δίνει έγκυρες εκτιμήσεις σε περίπτωση τυχαίων ελλειπουσών τιμών MAR. Ενώ οι μέθοδοι εκτίμησης που βασίζονται σε γενικευμένες εξισώσεις εκτίμησης (Generalized Estimating Equations-GEE) αλλά όχι στην πιθανοφάνεια απαιτούν πλήρως τυχαίες ελλείπουσες τιμές MCAR για να δώσουν συνεπείς εκτιμήσεις (Fitzmaurice et al., 1993; Kenward et al., 1994; Little, 1995). Επιπλέον, έχουν προταθεί μέθοδοι ανάλυσης που δεν βασίζονται στην πιθανοφάνεια αλλά δίνουν συνεπείς εκτιμήσεις και στην περίπτωση MAR ελλειπουσών τιμών. Οι συγκεκριμένες μέθοδοι βασίζονται στη στάθμιση με βάρη αντιστρόφως ανάλογα της εκτιμηθείσας πιθανότητας ανταπόκρισης (Robins et al., 1995; Heyting et al., 1992, Rotnitzky & Wypij, 1994).

Ο μηχανισμός παραγωγής ελλειπουσών τιμών χαρακτηρίζεται εντελώς τυχαίος (MCAR) ακόμα και αν η πιθανότητα εμφάνισής του $f(R)$ εξαρτάται μόνο από τις επεξηγηματικές μεταβλητές X και W . Ωστόσο, σε αυτή την περίπτωση ο διαχωρισμός μεταξύ MCAR και MAR δεν είναι εύκολος για αυτό ο Little (1995) πρότεινε τον όρο “εξαρτώμενος από συμμεταβλητές” (covariate dependent) για τον συγκεκριμένο μηχανισμό. Συγκρίνοντας την κατανομή της εξαρτημένης μεταβλητής κατά την χρονική στιγμή t_j μεταξύ των ατόμων με παρατηρηθείσες ή ελλείπουσες τιμές την χρονική στιγμή t_{j+1} , μπορούμε να διακρίνουμε μεταξύ “αγνοήσιμου” και “μη αγνοήσιμου” μηχανισμού δημιουργίας ελλειπουσών τιμών. Οι covariate dependent ελλείπουσες τιμές μπορούν να ελεγχθούν μέσω σύγκρισης των κατανομών της εξαρτημένης μεταβλητής ανάλογα με τα μοτίβα των ελλειπουσών τιμών μετά από στάθμιση για επεξηγηματικές μεταβλητές (Little, 1995). Οι

Park & Davis (1993) και Park & Lee (1997), έχουν προτείνει ελέγχους για MCAR ελλείπουσες τιμές σε περιπτώσεις επαναλαμβανόμενων διαχρονικών μετρήσεων.

Στην περίπτωση όπου ο μηχανισμός δημιουργίας ελλειπουσών τιμών είναι μη αγνοήσιμος πρέπει υποχρεωτικά να οριστεί ένα μοντέλο μη ανταπόκρισης για να εξαχθούν έγκυρα αποτελέσματα. Με την εξέταση των δεδομένων, όπως προαναφέρθηκε, δεν μπορεί κανείς να διερευνήσει την φύση ενός μη αγνοήσιμου μηχανισμού. Ο διαχωρισμός ενός MAR και ενός μη αγνοήσιμου μηχανισμού (MNAR) είναι δύσκολος διότι στην πραγματικότητα βασίζεται σε υποθέσεις οι οποίες δεν είναι δυνατό να ελεγχθούν (Curran et al., 1998). Για αυτές τις περιπτώσεις έχουν προταθεί μοντέλα που περιγράφουν ταυτόχρονα τις επαναλαμβανόμενες μετρήσεις και τον μηχανισμό δημιουργίας ελλειπουσών τιμών. Αλλά υπάρχουν περιπτώσεις όπου απλές αναλύσεις μπορούν να δώσουν έγκυρους εκτιμητές (Baker & Laird, 1988).

Οι Touloumi et al. (2001) χρησιμοποίησαν μελέτες προσομοίωσης για να διερευνήσουν την επίδραση των διάφορων ειδών ελλειπουσών τιμών, μεταξύ των οποίων και μη αγνοήσιμες, στην εκτίμηση του ρυθμού μεταβολής ενός ποσοτικού δείκτη χρησιμοποιώντας μεγάλη γκάμα μοντέλων. Σύμφωνα με τους Touloumi et al., τα μοντέλα μικτών επιδράσεων μπορούν να οδηγήσουν σε σημαντική μεροληψία όταν οι ελλείπουσες τιμές είναι μη αγνοήσιμες (MNAR), ενώ το ίδιο μπορεί να συμβεί ακόμα και αν ο μηχανισμός δημιουργίας ελλειπουσών τιμών είναι αγνοήσιμος (MAR) όταν χρησιμοποιούνται μοντέλα βασισμένα σε γενικές εξισώσεις εκτίμησης (Generalized Estimating Equations-GEE). Κατά συνέπεια συνιστούν οι ελλείπουσες τιμές να αντιμετωπίζονται με ιδιαίτερη προσοχή και να χρησιμοποιούνται κατάλληλα μοντέλα, στο πλαίσιο τουλάχιστον μιας ανάλυσης ευαισθησίας.

3.6 Σύνοψη μεθόδων για την ανάλυση δεδομένων με μη αγνοήσιμο μηχανισμό δημιουργίας ελλειπουσών τιμών

Κατά την διάρκεια των δυο τελευταίων δεκαετιών, το ενδιαφέρον στην στατιστική βιβλιογραφία για τις ελλείπουσες τιμές έχει επικεντρωθεί περισσότερο σε περιπτώσεις

των μη αγνοήσιμων ελλειπουσών τιμών, λόγω του πρόωρου τερματισμού της σειράς μετρήσεων ενός συνεχούς δείκτη, παρά σε ενδιάμεσες ελλείπουσες τιμές ενός δείκτη, οι οποίες συνήθως θεωρούνται τυχαίες (MAR) και άρα ο μηχανισμός παραγωγής τους θεωρείται αγνοήσιμος.

Σε αυτό το χρονικό διάστημα έχουν δημοσιευτεί αρκετά μοντέλα για την ανάλυση της διαχρονικής εξέλιξης των μετρήσεων του υπό μελέτη δείκτη και του μηχανισμού που οδηγεί στην περικοπή των μετρήσεων του (Wu & Carroll, 1988; Murray & Findlay, 1988; Wu & Bailey, 1988,1989; Schlutcher, 1992; Mori et al., 1992; Diggle & Kenward, 1994; Degrytola & Tu, 1994; Follman & Wu, 1995; Faucet & Thomas, 1996; Rotnitzky et al., 1998a; Touloumi et al., 1999; Scharfstein et al., 1999; Henderson et al., 2000, Song et al.,2002; Stubbendick & Ibrahim, 2003; Roy, 2003, Hogan et al., 2004; Vonesh et al., 2006; Pantazis et al., 2010; Crowther et al., 2012). Σύμφωνα με τον Little (1995) όλες αυτές οι μέθοδοι μπορούν να χωριστούν σε δυο μεγάλες υποκατηγορίες: τα μοντέλα μικτών προτύπων (pattern-mixture models) και τα μοντέλα επιλογής (selection models). Οι κατηγορίες αυτές προκύπτουν ανάλογα με την παραγοντοποίηση της από κοινού συνάρτησης πυκνότητας πιθανότητας (Laird, 1988; Little, 1995).

Στα pattern-mixture models η από κοινού συνάρτηση πυκνότητας πιθανότητας παραγοντοποιείται ως εξής:

$$f(Y, R|X, W, \theta_p, \varphi_p) = f(Y|R, X, W, \theta_p)f(R|X, \varphi_p) \quad (3.6.1)$$

Στα selection models η από κοινού συνάρτηση πυκνότητας πιθανότητας παραγοντοποιείται ως εξής:

$$f(Y, R|X, W, \theta_s, \varphi_s) = f(R|Y, X, W, \varphi_s)f(Y|X, \theta_s) \quad (3.6.2)$$

Οι δείκτες s και p αναφέρονται στα selection models και στα pattern-mixture models αντίστοιχα. Όταν οι ελλείπουσες τιμές είναι MCAR, οι 3.6.1 και 3.6.2 είναι ισοδύναμες αν $\theta_p = \theta_s$ και $\varphi_p = \varphi_s$. Στην περίπτωση όπου οι ελλείπουσες τιμές δεν είναι MCAR, οι 3.6.1 και 3.6.2 ενδέχεται να οδηγήσουν σε διαφορετικά μοντέλα (Little, 1995).

Τα pattern mixture models χαρακτηρίζονται από διαστρωμάτωση του υπό μελέτη πληθυσμού ανάλογα με συγκεκριμένα πρότυπα-δομές (patterns) του μηχανισμού δημιουργίας ελλειπυσών τιμών και μοντελοποιούν τις επαναλαμβανόμενες μετρήσεις μέσα σε κάθε ξεχωριστό pattern. Ο αριθμός των προτύπων δεν πρέπει να είναι πολύ μεγάλος διότι έτσι δεν είναι δυνατό να πραγματοποιηθεί διαστρωμάτωση γιατί θα υπάρχει μικρός αριθμός ατόμων σε κάθε πρότυπο. Αυτό το χαρακτηριστικό κάνει τα pattern mixture models καταλληλότερα για διαχρονικές μελέτες με διακριτές και καλά καθορισμένες χρονικές στιγμές μετρήσεων για όλα τα άτομα και οι πρόωρες αποχωρήσεις να συμβαίνουν σε μικρό αριθμό τέτοιων χρονικών στιγμών. Σε μελέτες που περιλαμβάνουν ανθρώπους, η παραπάνω απαίτηση δεν ικανοποιείται και αυτό έχει ως συνέπεια τα μοντέλα αυτά να μην εφαρμόζονται ευρέως. Ωστόσο, υπάρχει διαθέσιμη στατιστική βιβλιογραφία για αυτά τα μοντέλα, ο Dawson (1994) και Stavola & Christensen (1996) χρησιμοποίησαν τέτοια μοντέλα για να συνδυάσουν διαστρωμένες στατιστικές συνοπτικές με σκοπό τη σύγκριση υποπληθυσμών υπό διαφορετικά θεραπευτικά σχήματα. Οι Shih & Quan (1997) ανέπτυξαν μεθόδους για τον έλεγχο υποθέσεων για την σύγκριση υποομάδων, οι οποίες είναι εμπνευσμένες από τα pattern mixture models, ενώ η Rotnitzky et. al. (2001), παρουσιάζουν μια μεθοδολογία βασισμένη σε τέτοιου είδους μοντέλα κατάλληλη για ανάλυση ευαισθησίας.

Τα μοντέλα επιλογής (selection models) χωρίζονται σε δύο υποκατηγορίες ανάλογα με την βασική παραδοχή σχετικά με τη σχέση που συνδέει τις επαναλαμβανόμενες μετρήσεις με το μηχανισμό παραγωγής των ελλειπυσών τιμών. Στην μια ομάδα βασική παραδοχή αποτελεί η άμεση εξάρτηση του πρόωρου τερματισμού των μετρήσεων ενός ατόμου από το πλήρες σύνολο των μετρήσεών του, παρατηρηθέντων και μη (outcome dependent selection models). Τέτοιου είδους μοντέλα έχουν προταθεί από τους Diggle & Kenward (1994), Rotnitzky et al., (1998b), Scharfestein et al., (1999) ενώ η ευαισθησία τέτοιων μοντέλων μελετήθηκε από τους Verbeke et al. (2001). Η εφαρμογή τέτοιου είδους μοντέλων (outcome dependent selection models) δεν είναι εύκολη διότι οι προϋποθέσεις εφαρμογής τους, δηλαδή η παρακολούθηση των ατόμων στη μελέτη για ίσο χρόνο με κοινά σε όλους χρονικά σημεία μετρήσεις, δεν ικανοποιούνται σε πολλές περιπτώσεις που προκύπτουν στην κλινική πράξη.

Στην δεύτερη ομάδα των selection models η βασική παραδοχή αποτελεί η σύνδεση της πρόωρης αποχώρησης από την μελέτη και της διαχρονικής εξέλιξης του μετρούμενου δείκτη σε κάθε άτομο μέσω κοινών τυχαίων παραμέτρων (random effects selection models). Τα παραπάνω μοντέλα εφαρμόζονται ευρέως διότι δεν θέτουν περιορισμούς όσο αφορά τους χρόνους των μετρήσεων ή το συνολικό χρόνο παρακολούθησης. Επιπρόσθετα, αυτά τα μοντέλα δείχνουν καταλληλότερα σε περιπτώσεις όπου η πρόωρη αποχώρηση οφείλεται σε επιδείνωση της κατάστασης του ασθενούς ή σε θάνατο. Για αυτά τα μοντέλα χρησιμοποιείται και ο όρος "μοντέλα δεξιάς πληροφοριακής αποκοπής" (right informative censoring models) που εισήγαγαν ο Wu & Bailey, (1989) ενώ χρησιμοποιείται επίσης και ο όρος "non-ignorable random coefficient based drop out models" που εισήγαγε ο Little (1995). Οι Wu & Carroll (1988) πρότειναν ένα μοντέλο που χρησιμοποιεί ένα γραμμικό μοντέλο τυχαίων επιδράσεων για τις επαναλαμβανόμενες μετρήσεις ενός δείκτη και ένα probit μοντέλο για τις πρόωρες αποχωρήσεις το οποίο εξαρτάται από τις ατομικές παραμέτρους για την αρχική τιμή και για την κλίση της ευθείας που χαρακτηρίζει τη διαχρονική μεταβολή του εν λόγω χαρακτηριστικού. Το συγκεκριμένο μοντέλο επεκτάθηκε από τον Schlutcher (1992) μοντελοποιώντας ταυτόχρονα τις επαναλαμβανόμενες μετρήσεις του δείκτη και τον ατομικό χρόνο επιβίωσης, ενώ παρόμοια προσέγγιση ακολουθήθηκε από τους Deguttola & Tu (1994). Ο E-M αλγόριθμος χρησιμοποιείται για την εκτίμηση των παραμέτρων μέγιστης πιθανοφάνειας του μοντέλου. Η εκτίμηση των παραμέτρων παρόμοιου μοντέλου προσεγγίστηκε και από την σκοπιά της Μπεϋζιανής στατιστικής από τους Faucett & Thomas (1996).

Μια άλλη κατηγοριοποίηση τέτοιου είδους μοντέλων είναι ανάλογα με το αν το ενδιαφέρον εστιάζεται στην ανάλυση επιβίωσης ή στη μοντελοποίηση των επαναλαμβανόμενων μετρήσεων του δείκτη (Hogan & Laird, 1997b). Οι ίδιοι συγγραφείς δημοσίευσαν ένα αντίστοιχο μοντέλο και εκτιμητή δίνοντας έμφαση στην ανάλυση επιβίωσης (Hogan & Laird, 1997a). Οι Touloumi et al., (1999, 2003) ανέπτυξαν μια νέα μεθοδολογία για την εκτίμηση ενός αντίστοιχου μοντέλου βασισμένη στην επαναληπτική μέθοδο των γενικευμένων ελαχίστων τετραγώνων (IGLS-RIGLS) και στον Expectation-Maximization αλγόριθμο μαζί με μεθοδολογία για την εκτίμηση των τυπικών σφαλμάτων. Τέλος, οι Lyles et al. (2000) χρησιμοποιώντας επαναλαμβανόμενες

μετρήσεις ιικού φορτίου σε άτομα με HIV-1 λοίμωξη, παρουσίασαν μια επέκταση του μοντέλου ενσωματώνοντας σε αυτό τεχνικές που λάμβαναν υπόψη τους τις αριστερά περικομμένες τιμές ιικού φορτίου λόγω ύπαρξης κατώτατου ορίου ανίχνευσης στις χρησιμοποιούμενες μεθόδους μέτρησης του.

Το θέμα της αντιμετώπισης του προβλήματος που προκύπτει από την παρουσία μη αγνοήσιμων ελλειπών τιμών είναι υπό διερεύνηση στην διεθνή βιβλιογραφία. Αξίζει να σημειωθεί ότι κάποια μοντέλα είναι καταλληλότερα σε σχέση με άλλα ανάλογα με την μορφή των δεδομένων που αναλύονται και τους πιθανούς μηχανισμούς που οδηγούν σε πρόωρη αποχώρηση κάποιων ατόμων από την μελέτη.

Κεφάλαιο 4: Joint Models

4.1 Εισαγωγή

Στις διαχρονικές μελέτες συνήθως μελετάμε δυο ειδών αποκρίσεις, επαναλαμβανόμενες μετρήσεις ενός χαρακτηριστικού (δείκτη) και το χρόνο μέχρι το συμβάν όπως ο θάνατος, ανάπτυξη της νόσου ή απόσυρση από την μελέτη. Αυτές οι δυο αποκρίσεις αναλύονται συνήθως ξεχωριστά χρησιμοποιώντας μοντέλα μικτών επιδράσεων για τις επαναλαμβανόμενες μετρήσεις και μοντέλα ανάλυσης επιβίωσης για το χρόνο μέχρι το συμβάν. Ωστόσο, σε μερικές περιπτώσεις χρειάζεται η από κοινού μοντελοποίηση τους (joint modeling):

- i. Η πρώτη περίπτωση είναι όταν το ενδιαφέρον εστιάζεται στο χρόνο μέχρι το συμβάν και επιθυμούμε να λάβουμε υπόψη την επίδραση των μετρούμενου δείκτη ως χρόνο-εξαρτώμενης μεταβλητής. Σε αυτή την περίπτωση οι κλασσικές μέθοδοι δεν πρέπει να εφαρμόζονται διότι βασική προϋπόθεση αυτών των μεθόδων είναι ότι οι χρόνο-εξαρτώμενες μεταβλητές είναι εξωτερικές, δηλαδή η τιμή της μεταβλητής την χρονική στιγμή t δεν επηρεάζεται από τη εμφάνιση του συμβάντος την χρονική στιγμή u , $t > u$ (Klabfleisch and Prentice, 2002). Κατά συνέπεια οι χρόνο-εξαρτώμενες μεταβλητές, οι οποίες είναι αποτέλεσμα μιας στοχαστικής διαδικασίας που συνδέεται άμεσα με την περικοπή των παρατηρήσεων, δεν πληρούν αυτή την προϋπόθεση και για να έχουμε έγκυρα αποτελέσματα χρειάζεται ένα μοντέλο για την από κοινού κατανομή της διαχρονικής εξέλιξης του δείκτη και του χρόνου μέχρι το συμβάν.
- ii. Η δεύτερη περίπτωση είναι όταν το ενδιαφέρον εστιάζεται στη διαχρονική εξέλιξη ενός δείκτη. Σε αυτή την περίπτωση η εμφάνιση του συμβάντος οδηγεί σε περικοπή αφού δεν υπάρχουν διαθέσιμες επαναλαμβανόμενες μετρήσεις την στιγμή και ύστερα από το συμβάν. Όταν η περικοπή δεν είναι τυχαία τότε υπάρχει μεροληψία αν χρησιμοποιήσουμε μια μέθοδο η οποία αγνοεί τον μηχανισμό δημιουργίας ελλειπουσών τιμών. Σύμφωνα με Wulfsohn and Tsiatis (1997) και Henderson et al. (2000) χρησιμοποιώντας ένα joint μοντέλο το οποίο

αξιολογεί ταυτόχρονα τις διαχρονικές μετρήσεις αλλά και τον χρόνο επιβίωσης μπορούμε να μειώσουμε την μεροληψία των εκτιμήσεων και να βελτιώσουμε την ακρίβεια σε σχέση με απλούστερες μεθόδους ανάλυσης.

4.2 Random effects selection models

Στις διαχρονικές μελέτες είναι συνηθισμένο να παρατηρηθεί πρόωρη διακοπή των επαναλαμβανόμενων μετρήσεων ενός δείκτη. Αυτό μπορεί να οφείλεται σε επιδείνωση της νόσου ή σε θάνατο. Επομένως, είναι λογική η υπόθεση της σύνδεσης της κατανομής του χρόνου μέχρι το γεγονός που προκαλεί τον πρόωρο τερματισμό των μετρήσεων, με ατομικές παραμέτρους (random effects) που καθορίζουν την διαχρονική εξέλιξη του υπό μελέτη χαρακτηριστικού (δείκτη). Τα αντίστοιχα μοντέλα είναι ένα υποσύνολο των μοντέλων επιλογής (selection models), ενώ για την περιγραφή τους έχει προταθεί ο όρος random selection models.

Σε αυτά τα μοντέλα η περικοπή του χρόνου επιβίωσης λόγω τερματισμού της μελέτης ή άλλων αιτιών που δεν σχετίζονται με την υπό εξέταση νόσο, θεωρείται τυχαία. Αν ο χρόνος επιβίωσης του ατόμου i υποδηλώνεται ως T_i^S και ο περικομμένος χρόνος επιβίωσης ως C_i , τότε τα παρατηρηθέντα δεδομένα επιβίωσης για κάθε άτομο περιλαμβάνουν τον χρόνο $T_i = \min(T_i^S, C_i)$ και μια δείκτρια μεταβλητή δ_i η οποία παίρνει την τιμή 1 όταν $T_i = T_i^S$ και 0 όταν $T_i < T_i^S$. Επομένως, για ένα άτομο του οποίου οι μετρήσεις διακόπτονται λόγω θανάτου ή επιδείνωσης της νόσου θα ισχύει $T_i = T_i^S$ και $\delta_i = 1$, ενώ στις υπόλοιπες περιπτώσεις θα ισχύει $T_i = C_i$ και $\delta_i = 0$. Γενικά, τα C_i και T_i^S θεωρούνται ανεξάρτητα, δεδομένων των μετρήσεων του δείκτη Y_i και των υπόλοιπων συμμεταβλητών X_i , και εδώ εφαρμόζεται απλή ανάλυση επιβίωσης θεωρώντας τις τιμές του δείκτη ως μια επιπλέον συμμεταβλητή.

Ωστόσο, στις περιπτώσεις όπου ο χρόνος επιβίωσης σχετίζεται με ατομικές παραμέτρους της διαχρονικής εξέλιξης του υπό μελέτη χαρακτηριστικού (δείκτη) η περικοπή χαρακτηρίζεται ως *πληροφοριακή (informative censoring)* (Wu & Bailey, 1988; Wu & Carroll, 1988; Wu & Bailey, 1989). Με τον όρο αυτό χαρακτηρίζεται η περικοπή των

σειρών μετρήσεων του υπό μελέτη δείκτη $Y_i = (Y_{i1}, Y_{i2}, \dots, Y_{in_i})'$ και όχι η παρουσία περικομμένων χρόνων επιβίωσης. Ο όρος περικοπή (censoring) έχει καθιερωθεί στα πλαίσια της ανάλυσης επιβίωσης και για να μην υπάρχει σύγχυση έχει προταθεί η χρήση του όρου *πληροφοριακή απόσυρση* (informative drop-out) (Hogan & Laird, 1997). Επιπλέον, ο όρος πληροφοριακή απόσυρση χρησιμοποιείται στις περιπτώσεις όπου ο πρόωρος τερματισμός των μετρήσεων οδηγεί σε μη αγνοήσιμες ελλείπουσες τιμές (Diggle & Kenward, 1994). Η μη πληροφοριακή απόσυρση με μαθηματικές εκφράσεις είναι:

$$f(T^s|Y_0, Y_m) = f(T^s|Y_0) \quad (4.2.1)$$

$$f(Y_m|Y_0, T^s) = f(Y_m|Y_0) \quad (4.2.2)$$

Συνήθως στα random effects models το μοντέλο που χρησιμοποιείται για να περιγράψει τις επαναλαμβανόμενες μετρήσεις του υπό μελέτη δείκτη είναι το γραμμικό μοντέλο μεικτών επιδράσεων με ομοσκεδαστικά κατάλοιπα τύπου 1. Ο Schluchter (1992) και ο Degruetola & Tu (1994) παρουσίασαν μοντέλα με κοινό παρανομαστή τη βασική υπόθεση ότι ο χρόνος μέχρι το γεγονός που οδηγεί στον πρόωρο τερματισμό των μετρήσεων και οι τυχαίες επιδράσεις b_i ακολουθούν από κοινού μια πολυδιάστατη κανονική κατανομή. Σύμφωνα με την βασική υπόθεση αυτών των μοντέλων, ο χρόνος μέχρι το γεγονός που οδηγεί σε πρόωρο τερματισμό των μετρήσεων ή κάποιος μετασχηματισμός του, μπορεί να περιγραφεί ως γραμμικός συνδυασμός των τυχαίων επιδράσεων b_i συν ένα τυχαίο κανονικά κατανομημένο σφάλμα. Η ύπαρξη μη μηδενικών συνδιασπορών μεταξύ του χρόνου απόσυρσης και τουλάχιστον μια από τις τυχαίες επιδράσεις του δείκτη, συνεπάγεται μη αγνοήσιμο μηχανισμό παραγωγής ελλειπουσών τιμών. Η χρήση μεθόδων μέγιστης πιθανοφάνειας για την εκτίμηση των παραμέτρων των μοντέλων όταν υπάρχουν περικομμένοι χρόνοι απόσυρσης καθιστούν δύσκολη έως αδύνατη την εκτίμηση τους.

Η Touloumi et al (1999,2002) πρότειναν η εκτίμηση των παραμέτρων ενός αντίστοιχου μοντέλου να γίνεται βάσει μεθοδολογίας η οποία βασίζεται στον επαναληπτικό αλγόριθμο RIGLS σε συνδυασμό με ένα Expectation Maximazation αλγόριθμο. Για το

συγκεκριμένο μοντέλο χρησιμοποιείται ο όρος Joint Multivariate Random Effects Model-JMRE λόγω της πολυδιάστατης κανονικής κατανομής που ακολουθούν από κοινού οι τυχαίες επιδράσεις του υπό εξέταση δείκτη και τα κατάλοιπα του μοντέλου επιβίωσης, το οποίο χρησιμοποιείται για να περιγράψει την κατανομή του χρόνου απόσυρσης. Το μοντέλο αυτό επιτρέπει τη χρησιμοποίηση διαχρονικών δεδομένων χωρίς διακριτούς χρόνους για τις μετρήσεις του δείκτη και θεωρείται ευέλικτο ως προς τη μορφή της διαχρονικής εξέλιξης. Η εκτίμηση των παραμέτρων του είναι απλούστερη συγκριτικά με αυτή που βασίζεται σε μεθόδους μεγιστοποίησης της από κοινού πιθανοφάνειας ενώ η υλοποίηση της μεθόδου γίνεται στο στατιστικό πακέτο Stata (Pantazis & Touloumi, 2009).

4.3 Περιγραφή του Joint Multivariate Random Effects μοντέλου (JMRE)

Σύμφωνα με την Laird (1988) οι μέθοδοι που βασίζονται στην μέγιστη πιθανοφάνεια μπορούν να δώσουν αξιόπιστες εκτιμήσεις αγνοώντας τον μηχανισμό δημιουργίας ελλειπουσών τιμών μόνο στην περίπτωση όπου οι ελλείπουσες τιμές είναι MCAR ή MAR. Ενώ, στην περίπτωση MNAR μηχανισμού ελλειπουσών τιμών μπορεί να έχουμε σημαντική μεροληψία στις εκτιμήσεις. Σε αυτή την περίπτωση θα ήταν προτιμότερο να χρησιμοποιηθούν μοντέλα που μοντελοποιούν ταυτόχρονα την μεταβλητή απόκρισης και τον μηχανισμό δημιουργίας ελλειπουσών τιμών. Μια ειδική κατηγορία αυτών των μοντέλων είναι το Joint Multivariate Random Effects (JMRE) μοντέλο. Το JMRE μοντέλο υποθέτει ένα γραμμικό μοντέλο μικτών επιδράσεων για την πραγματική εξέλιξη του μελετώμενου δείκτη $\mathbf{Y}$ σε συνδυασμό με ένα λογαριθμοκανονικό (lognormal) μοντέλο επιβίωσης (Cox & Oakes, 1984) για τον χρόνο απόσυρσης.

Το γραμμικό μοντέλο μικτών επιδράσεων για την εξέλιξη του δείκτη $\mathbf{Y}$ περιγράφεται ως εξής:

$$Y_i = X_i\boldsymbol{\beta} + Z_i\mathbf{b}_i + \mathbf{e}_i \quad (4.3.1)$$

Όπου τα υπόλοιπα πρώτου επιπέδου θεωρούνται ανεξάρτητα με κατανομή $\mathbf{e}_i \sim N(\mathbf{0}, \sigma^2 I_{n_i})$ και οι τυχαίες επιδράσεις $\mathbf{b}_i \sim N(\mathbf{0}, \Sigma^m)$.

Ενώ για τους χρόνους απόσυρσης το lognormal μοντέλο επιβίωσης είναι:

$$\log(T_i^s) = \mathbf{x}_i^{s^T} \boldsymbol{\beta}^s + e_i^s \quad (4.3.2)$$

όπου $e_i^s \sim N(0, \sigma^s)$.

Για να διερευνήσουμε την ύπαρξη πληροφοριακής περικοπής στις μετρήσεις του δείκτη που μας ενδιαφέρει υποθέτουμε ότι οι τυχαίες επιδράσεις $\mathbf{b}_i$ και τα κατάλοιπα του μοντέλου επιβίωσης e_i^s ακολουθούν από κοινού την πολυμεταβλητή κανονική κατανομή:

$$\mathbf{b}_i^{joint} = \begin{pmatrix} b_i \\ e_i^s \end{pmatrix} \sim N \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma^{joint} \right\} \quad (4.3.3)$$

όπου $\Sigma^{joint} = \begin{Bmatrix} \Sigma^m & \\ \boldsymbol{\sigma}^{ms^T} & (\sigma^s)^2 \end{Bmatrix}$ και $\boldsymbol{\sigma}^{ms}$ είναι ένα διάνυσμα $\pi \times 1$ με τις συνδιακυμάνσεις των τυχαίων επιδράσεων $\mathbf{b}_i$ με τα κατάλοιπα του μοντέλου επιβίωσης.

Εάν $\boldsymbol{\sigma}^{ms} = \mathbf{0}$ τότε οι ελλείπουσες τιμές του δείκτη που προκύπτουν λόγω της εμφάνισης του συμβάντος προέρχονται από τυχαίο μηχανισμό.

Επομένως, το μοντέλο μπορεί να γραφτεί ως:

$$\mathbf{Y}_i^{joint} = X_i^{joint} \boldsymbol{\beta}^{joint} + Z_i^{joint} \mathbf{b}_i^{joint} + \mathbf{e}_i^{joint} \quad (4.3.4)$$

Όπου $\mathbf{Y}_i^{joint} = \{\mathbf{Y}_i^T, \ln(T_i^s)\}^T$, $\boldsymbol{\beta}^{joint} = (\boldsymbol{\beta}^T, \boldsymbol{\beta}^{s^T})^T$, $\mathbf{b}_i^{joint} = (\mathbf{b}_i^T, e_i^s)^T$, $\mathbf{e}_i^{joint} = (\mathbf{e}_i^T, 0)^T$, $X_i^{joint} = \begin{pmatrix} X_i & 0 \\ \mathbf{0}^T & X_i^s \end{pmatrix}$ και $Z_i^{joint} = \begin{pmatrix} Z_i & 0 \\ \mathbf{0}^T & 1 \end{pmatrix}$

Το $\boldsymbol{\beta}^{joint}$ περιλαμβάνει συντελεστές παλινδρόμησης τόσο για το μοντέλο της διαχρονικής εξέλιξης όσο και για το μοντέλο επιβίωσης. Ενώ, το $\mathbf{b}_i^{joint}$ περιλαμβάνει τις τυχαίες επιδράσεις για το μοντέλο του δείκτη και τα υπόλοιπα του μοντέλου επιβίωσης. Τέλος, το διάνυσμα $\mathbf{e}_i^{joint}$ περιλαμβάνει μόνο τα υπόλοιπα πρώτου επιπέδου του

μοντέλου του δείκτη αφού τα κατάλοιπα του μοντέλου επιβίωσης έχουν ήδη συμπεριληφθεί στο $\mathbf{b}_i^{joint}$.

Ο πίνακας διασποράς συνδιασποράς $V^{(joint)}$ για το από κοινού διάνυσμα των μετρήσεων $\mathbf{Y}^{(joint)}$ είναι μπλοκ-διαγώνιος, όπως και στο γραμμικό μικτό μοντέλο. Υποθέτοντας ένα απλό μοντέλο με τυχαία σταθερά και κλίση, που και οι δυο είναι συσχετισμένες με το χρόνο επιβίωσης, ο πίνακας διασποράς-συνδιασποράς $V_i^{(joint)}$ για το i άτομο έχει την μορφή:

$$V_i^{(joint)} = \begin{pmatrix} \sigma_{11}^2 + 2t_{i1}\sigma_{12} + t_{i1}^2\sigma_{22}^2 + \sigma^2 & \sigma_{11}^2 + (t_{i1} + t_{i2})\sigma_{12} + t_{i1}t_{i2}\sigma_{22}^2 & \dots & \sigma_{11}^2 + (t_{i1} + t_{in_i})\sigma_{12} + t_{i1}t_{in_i}\sigma_{22}^2 & \sigma_{1s} + t_{i1}\sigma_{2s} \\ \sigma_{11}^2 + (t_{i1} + t_{i2})\sigma_{12} + t_{i1}t_{i2}\sigma_{22}^2 & \sigma_{11}^2 + 2t_{i2}\sigma_{12} + t_{i2}^2\sigma_{22}^2 + \sigma^2 & \dots & \sigma_{11}^2 + (t_{i1} + t_{in_i})\sigma_{12} + t_{i2}t_{in_i}\sigma_{22}^2 & \sigma_{1s} + t_{i2}\sigma_{2s} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \sigma_{11}^2 + (t_{i1} + t_{in_i})\sigma_{12} + t_{i1}t_{in_i}\sigma_{22}^2 & \sigma_{11}^2 + (t_{i1} + t_{in_i})\sigma_{12} + t_{i2}t_{in_i}\sigma_{22}^2 & \dots & \sigma_{11}^2 + 2t_{in_i}\sigma_{12} + t_{in_i}^2\sigma_{22}^2 + \sigma^2 & \sigma_{1s} + t_{in_i}\sigma_{2s} \\ \sigma_{1s} + t_{i1}\sigma_{2s} & \sigma_{1s} + t_{i2}\sigma_{2s} & \dots & \sigma_{1s} + t_{in_{i-1}}\sigma_{2s} & \sigma_s^2 \end{pmatrix}$$

όπου σ_{1s} είναι η συνδιακύμανση των χρόνων επιβίωσης με τις ατομικές τυχαίες σταθερές και σ_{2s} των χρόνων επιβίωσης με τις ατομικές τυχαίες κλίσεις.

Στην περίπτωση όπου δεν υπάρχουν περικομμένοι χρόνοι απόσυρσης, δηλ. $T = T_i^s$, για κάθε i , όλες οι παράμετροι του μοντέλου, μετά από κατάλληλη αναδιάταξη, μπορούν να εκτιμηθούν με την μέθοδο RIGLS χρησιμοποιώντας οποιοδήποτε στατιστικό πακέτο που έχει δυνατότητες εκτίμησης μοντέλων τυχαίων επιδράσεων και ικανή ευελιξία στον ορισμό των πινάκων διακύμανσης-συνδιακύμανσης των τυχαίων επιδράσεων και των υπολοίπων πρώτου επιπέδου (π.χ. R, SAS, Splus, MLn, MLwiN).

Στην περίπτωση όπου υπάρχουν περικομμένοι χρόνοι απόσυρσης οι μέθοδοι εκτίμησης πρέπει να τροποποιηθούν, αλλιώς οι παρατηρηθέντες περικομμένοι χρόνοι απόσυρσης θα χρησιμοποιούνται ως πραγματικοί. Για αυτή την περίπτωση προτείνεται μια τροποποίηση του βασικού αλγορίθμου RIGLS, βασισμένη σε μια μορφή του EM αλγορίθμου, κατά την οποία σε κάθε επανάληψη του RIGLS αλγορίθμου, οι περικομμένοι χρόνοι επιβίωσης και τα αντίστοιχα αδρά υπόλοιπα αντικαθίστανται με τις δεσμευμένες αναμενόμενες τιμές τους, δεδομένων των παρατηρηθέντων δεδομένων και των τρεχουσών εκτιμήσεων των παραμέτρων (E-βήμα). Στη συνέχεια νέες εκτιμήσεις λαμβάνονται για τις παραμέτρους μέσω της RIGLS μεθόδου (M-βήμα). Οι εκτιμήσεις

των αναμενόμενων αυτών τιμών βασίζονται στις ιδιότητες της πολυδιάστατης κανονικής κατανομής και σε εκφράσεις για την πρώτη και δεύτερη ροπή της περικομμένης κανονικής κατανομής (Johnson & Kotz, 1970):

$$M_{1i} = E(\ln(T_i^s) | \ln(T_i^s) > \ln(C_i), Y_i) = \mu_{si} + S_{si} \frac{\varphi(u_i)}{1 - \Phi(u_i)} \quad (4.3.5)$$

$$M_{2i} = E(\{\ln(T_i^s)\}^2 | \ln(T_i^s) > \ln(C_i), Y_i) = \mu_{si}^2 + S_{si}^2 + S_{si}(\ln(C_i) + \mu_{si}) \frac{\varphi(u_i)}{1 - \Phi(u_i)} \quad (4.3.6)$$

όπου $u_i = \frac{\ln(C_i) - \mu_{si}}{S_{si}}$, $\varphi(u_i)$ και $\Phi(u_i)$ είναι η συνάρτηση πυκνότητας πιθανότητας και η αθροιστική συνάρτηση πιθανότητας της τυποποιημένης κανονικής κατανομής, αντίστοιχα. Τέλος, μ_{si} και S_{si}^2 είναι ο μέσος και η διασπορά του δεσμευμένου χρόνου επιβίωσης πάνω στις τιμές του δείκτη και μπορεί να εκτιμηθεί από την πολυδιάστατη κανονική θεωρία ως εξής:

$$\mu_{si} = E(\ln(T_i^s) | Y_i) = X_i^{sT} \mathbf{b}^s + \boldsymbol{\sigma}^{msT} Z_i' (Z_i \Sigma Z_i' + \sigma^2 \mathbf{I}_i)^{-1} (Y_i - X_i \hat{\mathbf{b}}) \quad (4.3.7)$$

$$S_{si}^2 = Var(\ln(T_i^s) | Y_i) = \sigma_s^2 - \boldsymbol{\sigma}^{msT} Z_i' (Z_i \Sigma Z_i' + \sigma^2 \mathbf{I}_i)^{-1} Z_i \boldsymbol{\sigma}^{msT} \quad (4.3.8)$$

Η μηδενική υπόθεση $\boldsymbol{\sigma}^{ms} = 0$ μπορεί να ελεγχθεί με χρήση λόγου πιθανοφανειών μεταξύ του μοντέλου (4.3.4) και ενός μοντέλου που οι παράμετροι $\boldsymbol{\sigma}^{ms}$ είναι δεσμευμένες στο μηδέν.

4.4 Διόρθωση του πίνακα διασποράς συνδιασποράς

Ο EM αλγόριθμος και οι παραλλαγές του είναι ένα σημαντικό εργαλείο που έχει χρησιμοποιηθεί για την ανάλυση δεδομένων υπό την παρουσία πληροφοριακής απόσυρσης. Ωστόσο, δεν παρέχει μια μέθοδος για την εκτίμηση του πίνακα πληροφορίας που σχετίζεται με τις εκτιμήτριες. Ο υπολογισμός της ακρίβειας των εκτιμητριών και οι αντίστοιχες στατιστικές για τον πίνακα υποθέσεων πρέπει να βασίζονται στα παρατηρηθέντα δεδομένα και όχι στον αναμενόμενο πίνακα πληροφορίας με τα

ψευδοπλήρη δεδομένα μετά την υποκατάσταση των ελλειπουσών τιμών. Ο ασυμπτωτικός πίνακας διασποράς-συνδιασποράς των εκτιμητριών μέγιστης πιθανοφάνειας δεν προκύπτει απλά από τον αντίστροφο του πίνακα πληροφορίας των ψευδοπλήρων δεδομένων κατά τη σύγκλιση του μοντέλου και ενώ έχει προηγηθεί η υποκατάσταση των ελλειπουσών τιμών. Η χρήση μιας τέτοιας μεθόδου η οποία χειρίζεται τις υποκατεστημένες ελλείπουσες τιμές ως παρατηρηθείσες αναμένεται να υποεκτιμά συστηματικά τις διασπορές των εκτιμητριών.

Κατά την σύγκλιση του JMRE μοντέλου ο πίνακας διασποράς συνδιασποράς των σταθερών επιδράσεων $\beta^{(joint)}$ είναι:

$$Var(\beta^{(joint)}) = (X^{(joint)}V^{(joint)-1}X^{(joint)}) \quad (4.4.1)$$

όπου $V^{(joint)} = Var(Y_i^{(joint)})$ και βασίζεται σύμφωνα με τον αλγόριθμο εκτίμησης, στα πλήρη δεδομένα μετά την αντικατάσταση των ελλειπουσών τιμών των περικομμένων χρόνων απόσυρσης από τις υπό συνθήκη αναμενόμενες τιμές τους (εξισώσεις 5.3.5 και 5.3.6).

Ο πίνακας πληροφορίας, $I(\theta|Y_0)$, προσαρμοσμένος για τις ελλείπουσες τιμές μπορεί να εκτιμηθεί διαφορίζοντας δύο φορές το λογάριθμο της συνάρτησης πιθανοφάνειας των παρατηρηθέντων δεδομένων $l(\theta|Y_0)$. Η διαδικασία διαφορίσης εμπεριέχει ολοκλήρωση πάνω σε ελλείπουσες τιμές και δεν είναι εύκολη. Λαμβάνοντας υπόψη τις υπολογιστικές δυσκολίες που υπάρχουν για την αναλυτική λύση του συγκεκριμένου προβλήματος οι Touloumi et al (1999 και 2003) αξιολόγησαν τρεις εναλλακτικές λύσεις:

- i. Την γενική μέθοδο που έχει προταθεί από τον Louis (1982)
- ii. Την μέθοδο πολλαπλών υποκαταστάσεων (multiple imputations)
- iii. Απλούστερη παραλλαγή της μεθόδου πολλαπλών υποκαταστάσεων που πρότειναν οι ίδιοι συγγραφείς.

Από την σύγκριση αυτή προέκυψε ότι η παραλλαγή της μεθόδου πολλαπλών υποκαταστάσεων είναι ισοδύναμη της γενικής μεθόδου του Louis και ταυτόχρονα υπολογιστικά απλούστερη. Η μέθοδος αυτή χρησιμοποιείται κατά την εφαρμογή του JMRE μοντέλου και περιγράφεται στη συνέχεια.

Η μέθοδος πολλαπλών υποκαταστάσεων βασίζεται στο γεγονός ότι κάθε ελλείπουσα τιμή αντικαθίσταται με περισσότερο από μια τιμές και προτάθηκε αρχικά από τον Rubin (1978). Οι πολλαπλές τιμές που αντικαθιστούν τις ελλείπουσες τιμές προέρχονται από επαναλαμβανόμενες δειγματοληψίες από την εκ των υστέρων (posterior) κατανομή των ελλειπουσών τιμών. Στην περίπτωση όπου οι δειγματοληψίες γίνονται από ένα μόνο μοντέλο μη απόκρισης, ο υπολογισμός της διασποράς των παραμέτρων του μοντέλου μπορεί να γίνει συνδυάζοντας τις διασπορές εντός και μεταξύ των συμπληρωμένων συνόλων δεδομένων.

Με M υποκαταστάσεις των ελλειπουσών τιμών δημιουργούνται M πλήρη σύνολα δεδομένων τα οποία μπορούν να αναλυθούν με χρήση μεθόδων για πλήρη δεδομένα. Έστω $\hat{\theta}_l$, $\widehat{W}_l$ με $l=1,2,\dots,M$ οι εκτιμήσεις και οι αντίστοιχοι πίνακες διασποράς συνδιασποράς που προκύπτουν από την ανάλυση των M πλήρων σύνολο δεδομένων. Η συνδυασμένη εκτιμήτρια είναι $\bar{\theta}_M = \sum_{l=1}^M \hat{\theta}_l / M$ και η συνολική διασπορά είναι το σταθμισμένο άθροισμα $T_M = \overline{W}_M + \frac{M+1}{M} B_M$, όπου $\overline{W}_M = \sum_{l=1}^M \widehat{W}_l / M$ είναι η μέση «εντός των υποκαταστάσεων διασπορά» και $B_M = \sum_{l=1}^M (\hat{\theta}_l - \bar{\theta}_M)(\hat{\theta}_l - \bar{\theta}_M)^T / M - 1$ είναι η μεταξύ των υποκαταστάσεων διασπορά. Τέλος, ο όρος $\frac{M+1}{M}$ είναι η διόρθωση για τον πεπερασμένο αριθμό υποκαταστάσεων. Οι μέσες τιμές των εκτιμητριών από τα πλήρη δεδομένα είναι συνεπείς (consistent) μόνο για τις παραμέτρους σταθερών επιδράσεων και όχι για τις παραμέτρους τυχαίων επιδράσεων γιατί οι παράμετροι αυτές δεν εισέρχονται ως γραμμικές συναρτήσεις των δεδομένων στη συνάρτηση μέγιστης πιθανοφάνειας. Οι Touloumi et al (1999) για να επιλύσουν αυτό το πρόβλημα πρότειναν δυο αλλαγές στην μέθοδο των πολλαπλών υποκαταστάσεων:

- Η «εντός των υποκαταστάσεων» μέση διασπορά αντικαθίσταται από τον πίνακα διασποράς συνδιασποράς κατά την σύγκλιση του μοντέλου.
- Το μέρος της «μεταξύ των υποκαταστάσεων διασποράς» που αναφέρεται στις παραμέτρους σταθερών επιδράσεων υπολογίζεται γύρω από την εκτίμηση μέγιστης πιθανοφάνειας κατά την σύγκλιση του μοντέλου.

Δηλαδή αντικαθίσταται το $B_M = \sum_{l=1}^M (\hat{\theta}_l - \bar{\theta}_M)(\hat{\theta}_l - \bar{\theta}_M)^T / M - 1$ με

$\mathbf{B}_M^{mod} = \sum_{l=1}^M (\hat{\beta}_l - \hat{\beta}^{(joint)})(\hat{\beta}_l - \hat{\beta}^{(joint)})^T / M - 1$. Η τιμή του M που χρησιμοποιείται είναι 5 όπως προτείνει ο Rubin (1978) και χρησιμοποιήθηκε κατά την εφαρμογή του JMRE μοντέλου. Πρέπει να σημειωθεί ότι οι διορθώσεις που προκύπτουν είναι σχετικά περιορισμένες και εντοπίζονται κυρίως στις παραμέτρους του μοντέλου επιβίωσης, πράγμα το οποίο είναι αναμενόμενο διότι ο EM αλγόριθμος εφαρμόζεται για περικομμένες τιμές του χρόνου απόσυρσης.

Κεφάλαιο 5: Διερεύνηση διαχρονικών δεδομένων – Η μελέτη AMACS

5.1 Εισαγωγή

Η ανάλυση διαχρονικών δεδομένων, όπως και κάθε άλλη στατιστική ανάλυση, αποτελείται από δυο συνιστώσες, την διερευνητική (exploratory) και την επιβεβαιωτική (confirmatory) ανάλυση. Η επιβεβαιωτική ανάλυση είναι κριτική δουλειά και ζυγίζει τις ενδείξεις των δεδομένων υπέρ ή κατά των υποθέσεων. Η διερευνητική ανάλυση είναι δουλειά ντετέκτιβ, στην ουσία συμπεριλαμβάνει τεχνικές έτσι ώστε να οπτικοποιηθούν τα μοτίβα των δεδομένων. Σκοπός της διερευνητικής ανάλυσης είναι:

- i. Η βαθύτερη κατανόηση των δεδομένων
- ii. Αποκάλυψη λανθάνουσων δομών και σχέσεων
- iii. Εντοπισμός σημαντικών μεταβλητών σε σχέση με το πρόβλημα και έκτοπες παρατηρήσεις
- iv. Κατασκευή/έλεγχος κατάλληλων υποθέσεων και δημιουργία φειδωλών μοντέλων

Ειδικότερα στην ανάλυση των διαχρονικών δεδομένων υπάρχει η ανάγκη διερεύνησης των δεδομένων τόσο σε ατομικό όσο σε πληθυσμιακό επίπεδο. Επιπρόσθετα, λόγω της παρουσίας ελλειπουσών τιμών με την μορφή του πρόωρου τερματισμού των μετρήσεων το οποίο αποτελεί σημαντικό χαρακτηριστικό των δεδομένων χρειάζεται ιδιαίτερη προσοχή στην διερευνητική ανάλυση επαναλαμβανόμενων μετρήσεων διότι αυτό σχετίζεται άμεσα με τις μεθόδους που θα χρησιμοποιηθούν στην περαιτέρω ανάλυση.

Τα κύρια εργαλεία που χρησιμοποιούνται στην διερευνητική ανάλυση είναι πίνακες περιγραφικών μέτρων και γραφικές αναπαραστάσεις των δεδομένων. Με τις γραφικές αναπαραστάσεις μπορούν να οπτικοποιηθούν οι μεταβολές στα επίπεδα των μεταβλητών κύριου ενδιαφέροντος με την πάροδο του χρόνου αλλά και να αποκαλυφθούν πιθανές διαφοροποιήσεις ή γενικά μη αναμενόμενα μοτίβα. Είναι γνωστό ότι δεν υπάρχει μόνο μια μέθοδος προσέγγισης για αποτελεσματικές γραφικές μεθόδους στο πλαίσιο διαχρονικών δεδομένων αλλά ο Diggle et al (1994) προτείνουν τα εξής:

- έμφαση όσο το δυνατόν μεγαλύτερη στα ακατέργαστα δεδομένα παρά σε σύνοψη των δεδομένων
- έμφαση σε μοτίβα που προκύπτουν από σύνοψη των δεδομένων με πιθανό επιστημονικό ενδιαφέρον
- εντοπισμό μοτίβων τόσο σε συγχρονικό όσο και σε διαχρονικό επίπεδο
- διευκόλυνση του εντοπισμού των ατόμων με ασυνήθιστη διαχρονική πορεία ή μεμονωμένων έκτροπων παρατηρήσεων.

Σε αυτό το κεφάλαιο θα επικεντρωθούμε στην διερευνητική ανάλυση διαχρονικών δεδομένων με εφαρμογή των μεθόδων σε ένα υποσύνολο των δεδομένων της μελέτης AMACS.

5.2 Μελέτη AMACS

Η μελέτη AMACS είναι μια επιδημιολογική προοπτική μελέτη HIV-1 οροθετικών ατόμων η οποία ξεκίνησε το 1996 με αρχικό σκοπό την αξιολόγηση της προγνωστικής αξίας ιολογικών και ανοσολογικών δεικτών στην ανταπόκριση στην αντιρετροϊκή θεραπεία και γενικότερα στην εξέλιξη της HIV λοίμωξης. Η μελέτη της φυσικής ιστορίας της HIV-1 λοίμωξης στη νέα περίοδο θεραπείας είναι ιδιαίτερα σημαντική τόσο από επιστημονικής απόψεως όσο και από την πλευρά της δημόσιας υγείας. Προάγει την κατανόηση της λειτουργίας του HIV-1 και της αλληλεπίδρασης του με τον ανθρώπινο οργανισμό και την αντιρετροϊκή θεραπεία. Ταυτόχρονα παρέχει ανανεωμένες εκτιμήσεις του χρόνου επώασης της νόσου, πληροφορίες απαραίτητες για την εκτίμηση της επίπτωσης της λοίμωξης στον εγχώριο πληθυσμό και τον αντίστοιχο προγραμματισμό προγραμμάτων παρέμβασης στην δημόσια υγεία. Για να απαντηθούν ερωτήματα σχετικά με την φυσική ιστορία της νόσου είναι απαραίτητο να υπάρχουν διαθέσιμες προοπτικές μελέτες. Οι προοπτικές μελέτες λειτουργούν συμπληρωματικά στις κλινικές δοκιμές ενώ ταυτόχρονα μπορούν να παρέχουν πληροφορίες αναφορικά με την εξέλιξη της φυσικής ιστορίας της HIV-1 λοίμωξης.

Οι κύριοι σκοποί της AMACS είναι:

- I. Η εκτίμηση του προσδόκιμου χρόνου επιβίωσης των οροθετικών ατόμων και οι διαχρονικές μεταβολές αυτού, ιδιαίτερα μετά την εισαγωγή της αντιρετροϊκής θεραπείας υψηλής δραστηριότητας.
- II. Η αξιολόγηση της μακροχρόνιας αποτελεσματικότητας των διαφορετικών θεραπευτικών σχημάτων/στρατηγικών τόσο στην ιολογική και ανοσολογική ανταπόκριση των ασθενών όσο και στην γενική θνησιμότητα.
- III. Η διερεύνηση των μεταβολών στην συχνότητα εμφάνισης συμπτωμάτων AIDS στη μετά-HAART εποχή.
- IV. Η εκτίμηση της αναμενόμενης εξέλιξης των ασθενών δοθείσης της κλινικής και θεραπευτικής ιστορίας τους.
- V. Η καταγραφή ανεπιθύμητων ενεργειών καθώς και η αξιολόγηση της συχνότητας εμφάνισης τους ανάλογα με τα θεραπευτικά σχήματα.
- VI. Η καταγραφή του ποσοστού συμμόρφωσης των ασθενών στα προτεινόμενα θεραπευτικά σχήματα καθώς και η διερεύνηση των παραγόντων (δημογραφικών και άλλων) που πιθανώς σχετίζονται με την συμμόρφωση.
- VII. Η διαχρονική παρακολούθηση και αξιολόγηση της ανάπτυξης ανοχής και οι επιπτώσεις στην εξέλιξη της HIV-1 λοίμωξης.

Τα κριτήρια εισόδου ασθενών στη βάση δεδομένων είναι να υπάρχει επιβεβαιωμένο HIV θετικό τεστ και να υπάρχει διάρκεια παρακολούθησης από την κλινική για τουλάχιστον 12 μήνες ή να εξακολουθούν να παρακολουθούνται. Επιπλέον, όλοι οι ασθενείς που βρίσκονταν υπό παρακολούθηση στις συνεργαζόμενες κλινικές και ανεξαρτήτως θεραπευτικής αγωγής ή τελικής έκβασης την 1/1/1996 ή που προσήλθαν στην κλινική μετά την 1/1/1996 μπορούν να ενταχθούν στη βάση. Τα δεδομένα που καταγράφονται συλλέγονται στα πλαίσια της καθημερινής φροντίδας των ασθενών και είναι ανώνυμα. Τα στοιχεία που συλλέγονται αφορούν σε:

- i. Δημογραφικά στοιχεία
- ii. Κλινική κατάσταση
- iii. Ατομικό ιατρικό ιστορικό
- iv. Ανοσολογικές εξετάσεις
- v. Ιολογικές εξετάσεις

- vi. Βιοχημικές εξετάσεις
- vii. Ορολογικές εξετάσεις
- viii. Τρέχουσα κατάσταση ασθενούς
- ix. Συλλοιμώσεις

Για τα στοιχεία που συλλέγονται γίνεται ποιοτικός έλεγχος με δυο τρόπους: 1) με ανεύρεση πιθανών λαθών καταγραφής μέσω λογικών ελέγχων και 2) με επιτόπιο έλεγχο των ιατρικών φακέλων για ένα μέρος των ασθενών(περίπου 10%).

Για την στατιστική ανάλυση δεδομένων χρησιμοποιούνται μέθοδοι ανάλογα με το υπό μελέτη ερώτημα όπως t-test, X² και μη παραμετρικά τεστ για την μονοπαραγοντική ανάλυση, λογιστική παλινδρόμηση και ανάλυση επιβίωσης για την πολυπαραγοντική ανάλυση. Η ανάλυση των τάσεων σε αναπληρωτές δείκτες όπως τα CD4 λεμφοκύτταρα και το ιικό φορτίο πριν ή μετά την έναρξη αντιρετροϊκής θεραπείας απαιτεί την εφαρμογή μεθόδων ανάλυσης πολλαπλών στο χρόνο μετρήσεων (longitudinal data analysis). Κατά την ανάλυση τέτοιων δεδομένων υπάρχει ο κίνδυνος εσφαλμένων συμπερασμάτων λόγω πληροφοριακής αποκοπής (informative censoring or informative drop-out). Για παράδειγμα, όταν μελετάται ο ρυθμός μεταβολής των CD4 λεμφοκυττάρων κατά την διακοπή της αντιρετροϊκής θεραπείας, άτομα με ταχύτερους ρυθμούς πτώσης είναι πιθανότερο να ξεκινήσουν νέα θεραπευτική αγωγή. Επομένως, άτομα με μικρότερο αριθμό μετρήσεων, και άρα με περισσότερο ασταθή εκτίμηση ρυθμού πτώσης (μεγαλύτερο πιθανό σφάλμα) είναι πιθανό να αποτελούν ειδική υπό-ομάδα. Άρα η έλλειψη μετρήσεων παρέχει πληροφορία. Για την ανάλυση διαχρονικών τάσεων χρησιμοποιούνται ευρέως μοντέλα τυχαίων επιδράσεων (random effects models) τα οποία λαμβάνουν υπόψη τη συσχέτιση των μετρήσεων στο ίδιο άτομο. Όμως αυτά τα μοντέλα οδηγούν σε εσφαλμένες εκτιμήσεις όταν υπάρχει πληροφοριακή αποκοπή.

5.3 Περιγραφή χρόνου, συχνότητας και αριθμού μετρήσεων CD4 λεμφοκυττάρων

Για την παρούσα ανάλυση θα χρησιμοποιηθούν δεδομένα από την βάση AMACS και για να εισαχθεί ένα άτομο στη μελέτη θα πρέπει να έχει καταγεγραμμένη τουλάχιστον μια μέτρηση CD4 και η ημερομηνία ορομετατροπής και θανάτου να είναι μεταγενέστερες από το 1996 που είναι η χρονιά έναρξης της AMACS.

Πίνακας 5.1: Ροή ατόμων στην μελέτη.

Άτομα που έχουν τουλάχιστον 1 μέτρηση CD4:	4768
Έχουν δημογραφικά στοιχεία	4751
Ημερομηνία θανάτου > 1996:	4745
Ημερομηνία ορομετατροπής γνωστή:	388
Ημερομηνία ορομετατροπής >1996:	271
Μετρήσεις CD4 έως το 1ο AIDS συμβάν/ θάνατο/ έναρξη HAART θεραπείας	253
Τελικός αριθμός ασθενών στην παρούσα ανάλυση:	253

Αυτή τη στιγμή στη βάση AMACS υπάρχουν 4768 ασθενείς για τους οποίους υπάρχει διαθέσιμη τουλάχιστον μια μέτρηση CD4 λεμφοκυττάρων εκ των οποίων μόνο οι 388 έχουν γνωστή ημερομηνία ορομετατροπής. Το ενδιαφέρον της ανάλυσης επικεντρώνεται σε ασθενείς με ημερομηνία ορομετατροπής μετά από το 1996 και για αυτό καταλήγουμε εν τέλει σε 253 ασθενείς για τους οποίους υπάρχει τουλάχιστον μια διαθέσιμη μέτρηση CD4 λεμφοκυττάρων μέχρι την εμφάνιση AIDS ή θανάτου ή έναρξη θεραπείας.

Πίνακας 5.2: Περιγραφικά χαρακτηριστικά δείγματος.

	N (%)
Φύλο	
<i>Αντρας</i>	241 (95,3)
<i>Γυναίκα</i>	12 (4,7)
Ηλικία στην ορομετατροπή	
<i><40</i>	203 (80,2)
<i>≥40</i>	50 (19,8)
Χώρα προέλευσης	
<i>Ελλάδα</i>	230 (90,9)
<i>Άλλη</i>	23 (9,1)
Τρόπος μετάδοσης	
<i>Ομοφυλόφιλος</i>	184 (80,0)
<i>Χρήστης ενδοφλέβιων ναρκωτικών ουσιών (XEN)</i>	13 (5,6)
<i>Ετεροφυλόφιλος</i>	31 (13,5)
<i>Άλλο</i>	2 (0,9)
	Διάμεσος (Ενδοτεταρτομοριακό εύρος)
Ηλικία (χρόνια)	30,9 (26,7 - 38,0)

Στον υπό μελέτη πληθυσμό έχουμε κυρίως άντρες (95,3%) και είναι στην πλειοψηφία τους έλληνες πολίτες (90,9%). Το 80% των ασθενών είναι ηλικίας κάτω των 40 ετών, η διάμεση ηλικία είναι 31 έτη ενώ το 50% των ατόμων του δείγματος είναι μεταξύ 26 και 38 ετών. Η μετάδοση γίνεται κυρίως μέσω σεξουαλικής επαφής μεταξύ ομοφυλόφιλων ατόμων.

Πίνακας 5.3: Περιγραφή χρόνου, αριθμού και συχνότητας των μετρήσεων CD4 λεμφοκυττάρων.

	AIDS/ Όχι Διάμεσος (E.T.E)	Θάνατος / Ναι Διάμεσος (E.T.E)	έναρξη HAART Σύνολο Διάμεσος (E.T.E)	p-value
Αριθμός μετρήσεων	4,0 (2,0-10,0)	3 (1,0-5,0)	3,0 (1,0-6,0)	0,029
Χρόνος πρώτης μέτρησης από ΟΜ* (έτη)	0,54 (0,27-1,10)	0,70 (0,34-1,43)	0,63 (0,33-1,31)	0,645
Χρόνος τελευταίας μέτρησης από ΟΜ* (έτη)	2,07 (0,90-3,87)	1,58 (0,72-2,82)	1,63 (0,77-3,05)	0,148
Χρόνος μεταξύ μετρήσεων (μήνες)	4,54 (2,80-5,93)	3,07 (1,30-4,28)	3,56 (1,83-4,81)	0,200

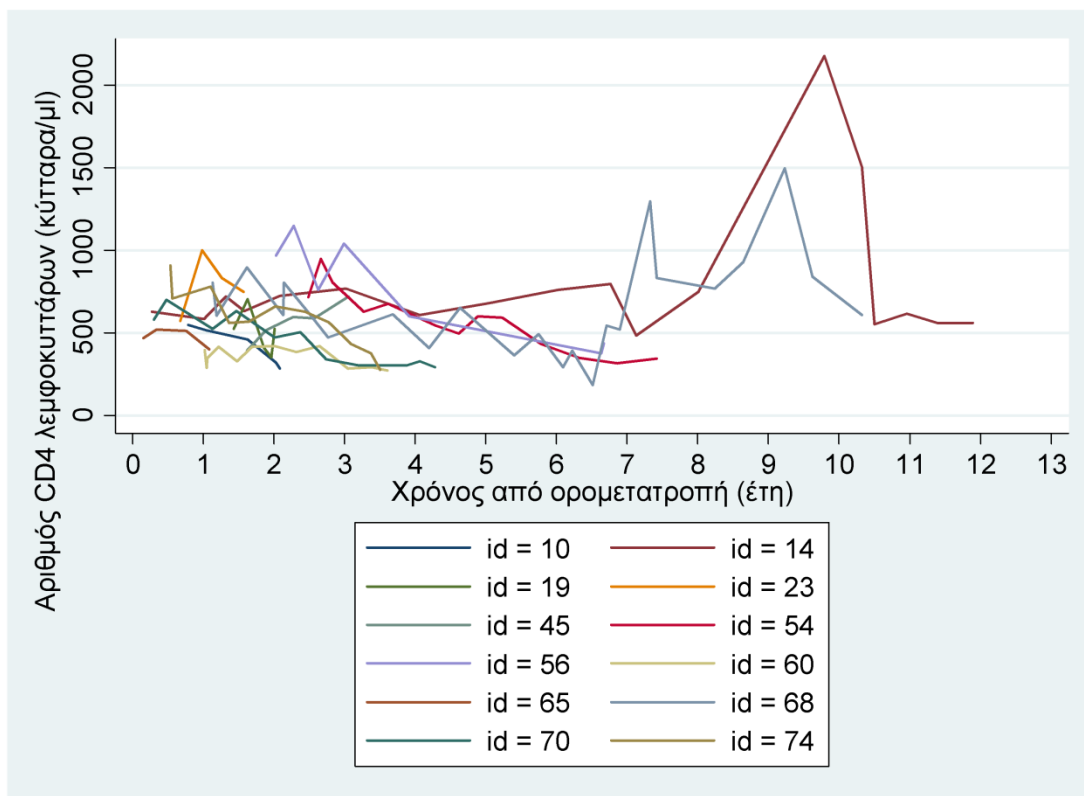
*: Ορομετατροπή

Από τον πίνακα 5,3 είναι εμφανές ότι για την ομάδα ατόμων που δεν έχουν το συμβάν (κάποια νόσο καθοριστική του AIDS ή το συμβάν θανάτου ή έναρξη θεραπείας) ο διάμεσος αριθμός μετρήσεων CD4 λεμφοκυττάρων είναι 4 και διαφέρει στατιστικά σημαντικά από τον αντίστοιχο αριθμό μετρήσεων της ομάδας που έχει εμφανίσει το συμβάν (p-value=0,029). Ο χρόνος μεταξύ των μετρήσεων είναι 4,5 μήνες για τους ασθενείς που ανήκουν στην ομάδα που δεν εμφάνισε το συμβάν ενώ στην ομάδα ατόμων με το συμβάν ο χρόνος μεταξύ των μετρήσεων είναι περίπου 3 μήνες. Τα άτομα της πρώτης ομάδας ίσως να αισθάνονται αρκετά καλύτερα σε σχέση με τα άτομα της άλλης ομάδας και γι αυτό να επισκέπτονται την κλινική για να κάνουν ανοσολογικές εξετάσεις σε πιο αραιά χρονικά διαστήματα. Επιπρόσθετα, οι ασθενείς που εμφανίζουν τελικά το συμβάν έχουν κατά μέσο όρο μεγαλύτερο χρόνο πρώτης μέτρησης και μικρότερο χρόνο τελευταίας μέτρησης σε σύγκριση με τους HIV οροθετικούς ασθενείς που δεν εμφανίζουν το συμβάν. Σύμφωνα με τα ανωτέρω στοιχεία φαίνεται ότι οι ασθενείς του δείγματος έρχονται για την πρώτη μέτρηση CD4 λεμφοκυττάρων περίπου στους πρώτους έξι μήνες και ανά 3,5 μήνες για παρακολούθηση της κατάστασής τους κάνοντας ανοσολογικές εξετάσεις.

Οι μελέτες επαναλαμβανόμενων δεικτών εστιάζουν κυρίως στον ρυθμό μεταβολής των δεικτών στον χρόνο. Το στικτόγραμμα, του οποίου τα στοιχεία που αντιστοιχούν στο ίδιο άτομο ενώνονται με συνεχείς γραμμές, των τιμών των CD4 λεμφοκυττάρων ως προς το χρόνο αποτελεί ένα χρήσιμο εποπτικό εργαλείο. Ωστόσο, όσο αυξάνεται ο αριθμός των ατόμων στη μελέτη το γράφημα αυτό γίνεται πιο πολύπλοκο και δυσνόητο. Για αυτό το λόγο είναι σύνηθες να επιλεχτεί μια υποομάδα και για αυτή δημιουργείται στικτόγραμμα το οποίο απεικονίζει την διαχρονική εξέλιξη του δείκτη ανά άτομο. Στα δεδομένα της παρούσας εργασίας χρησιμοποιείται ένα δείγμα 12 ατόμων οι οποίοι έχουν τουλάχιστον 4 μετρήσεις CD4 λεμφοκυττάρων.

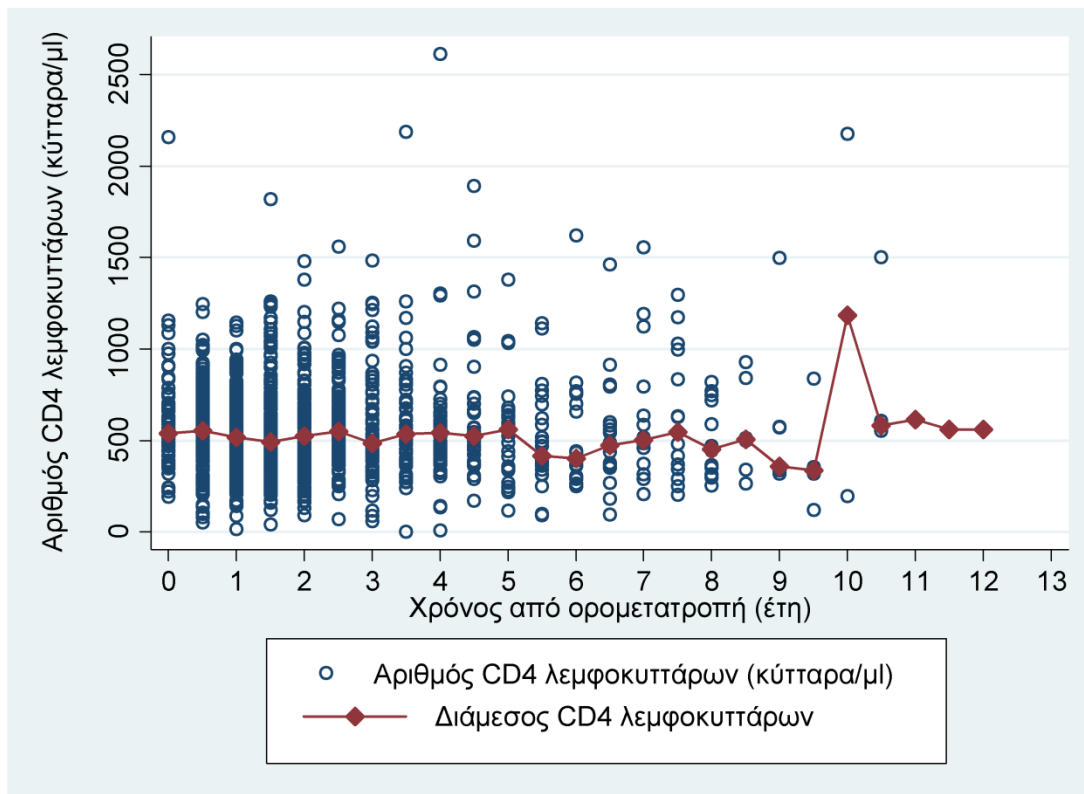
Στο γράφημα 5.1 φαίνονται οι ατομικές πορείες 12 ατόμων και είναι εμφανές ότι υπάρχει πτωτική τάση CD4 λεμφοκυττάρων με την πάροδο του χρόνου.

Γράφημα 5.1: Διαχρονική εξέλιξη CD4 λεμφοκυττάρων (cells/ml) σε δείγμα 12 ατόμων που έχουν τουλάχιστον 4 μετρήσεις CD4 λεμφοκυττάρων ο καθένας.



Επιπρόσθετα, εκτός από το ανωτέρω γράφημα ένα μέρος της περιγραφικής διερεύνησης σε διαχρονικές μελέτες αποτελούν τα γραφήματα που παρουσιάζουν τιμές κάποιου επιλεγμένου μέτρου θέσης (π.χ. μέση τιμή, διάμεσος) για την κατανομή του δείκτη ανά τακτά χρονικά διαστήματα. Τα γραφήματα αυτά δίνουν μια αδρή εικόνα για την διαχρονική εξέλιξη του δείκτη στον πληθυσμό. Αξίζει να σημειωθεί ότι τέτοιου είδους γραφήματα λόγω της συγχρονικής τους φύσης μπορεί να διαφέρουν από την μέση ατομική εξέλιξη.

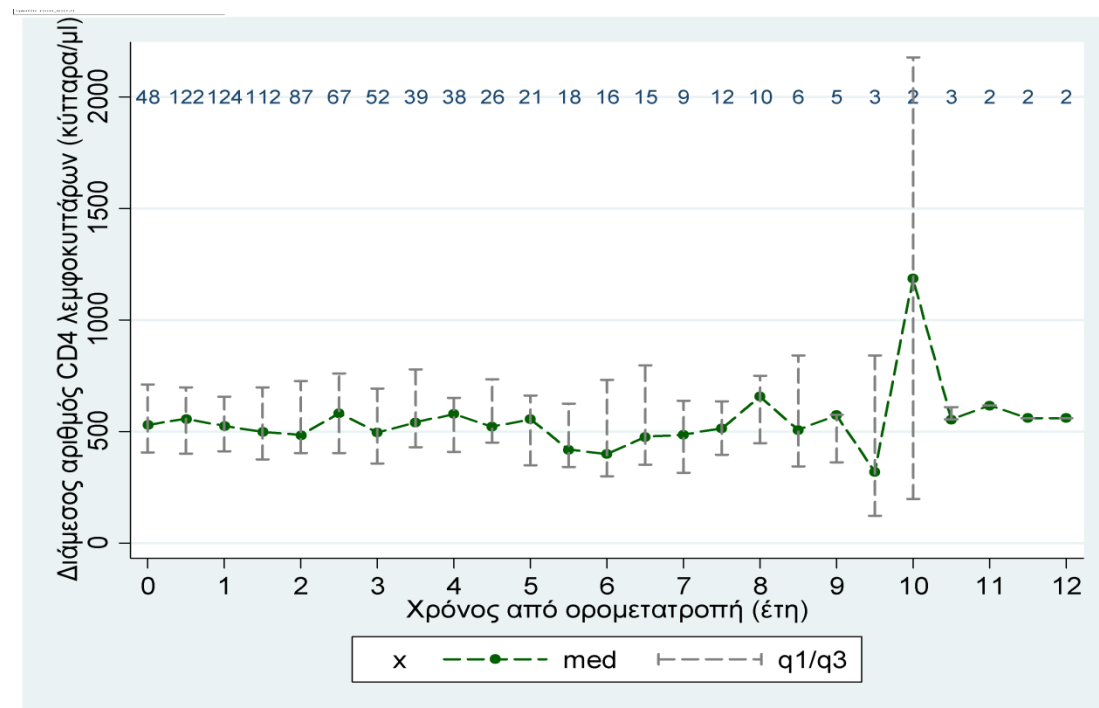
Γράφημα 5.2: Πληθυσμιακοί διάμεσοι CD4 λεμφοκυττάρων ανά εξάμηνο από την ημερομηνία ορομετατροπής έως την εμφάνιση κάποιου συμβάντος.



Το γράφημα 5.2 παρουσιάζει τις διάμεσες μέσες τιμές και τα CD4 λεμφοκύτταρα ανά εξάμηνο από την ημερομηνία ορομετατροπής. Συνήθως προτιμάται ο διάμεσος σε σχέση με την μέση τιμή διότι η δεύτερη παρουσιάζει μεγαλύτερη ευαισθησία στην ύπαρξη ακραίων τιμών και τα CD4 λεμφοκύτταρα έχουν ασύμμετρη κατανομή. Επιπλέον, πρέπει να αναφερθεί ότι εξαιτίας της μη ύπαρξης κοινών χρόνων μετρήσεων, εφαρμόστηκε μια

μορφή στρογγυλοποίησης στους χρόνους μέτρησης. Θεωρώντας ότι οι μετρήσεις CD4 λεμφοκυττάρων γίνονται ανά εξάμηνο, για κάθε άτομο και χρονική στιγμή t (σε μήνες) εντοπίστηκε η μέτρηση εκείνη που ήταν στο διάστημα $[t-3, t+3)$ και είχε την μικρότερη απόλυτη χρονική απόσταση από τη χρονική στιγμή t . Η συγκεκριμένη μέτρηση θεωρήθηκε ότι πραγματοποιήθηκε την χρονική στιγμή t και έτσι καταλήγουμε σε ένα σύνολο δεδομένων όπου τα άτομα έχουν κοινούς χρόνους μέτρησης.

Γράφημα 5.3: Πληθυσμιακοί διάμεσοι CD4 λεμφοκυττάρων ανά εξάμηνο από την ημερομηνία ορομετατροπής έως την εμφάνιση κάποιου συμβάντος.



Είναι σημαντικό να υπογραμμιστεί ότι τα συγκεκριμένα γραφήματα είναι πολύ πιθανό να δίνουν μια μεροληπτική εικόνα για τον ρυθμό πτώσης των CD4 λεμφοκυττάρων. Σε αυτά τα δεδομένα είναι σύνηθες το φαινόμενο της προοδευτικής περικοπής μετρήσεων εξαιτίας θανάτου ή ανάπτυξης κλινικού AIDS αλλά και λόγω έναρξης αντιρετροϊκής θεραπείας σε άτομα με χαμηλά CD4 λεμφοκύτταρα. Αντί για γραφήματα που παρουσιάζουν μέσες ή διάμεσες τιμές των δεικτών ανά χρονικό σημείο θα μπορούσαν να χρησιμοποιηθούν γραφήματα όπου παρουσιάζονται αλλαγές σε σχέση με κάποια αρχική τιμή. Τέτοιου είδους γραφήματα είναι περισσότερο πληροφοριακά όσο αφορά στις μεταβολές των δεικτών και λιγότερο μεροληπτικά. Ωστόσο, οι χρόνοι έναρξης των

μετρήσεων διαφέρουν σημαντικά μεταξύ των ατόμων της μελέτης και η έννοια μιας κοινής αρχικής μέτρησης δεν υφίσταται σε αυτή την περίπτωση για αυτό και δεν είναι δυνατό να προκύψουν τέτοια γραφήματα. Να σημειωθεί ότι στην περίπτωση όπου ο ρυθμός μεταβολής των δεικτών συνδέεται με τον χρόνο επιβίωσης, τα γραφήματα τέτοιου τύπου υπόκεινται επίσης σε μεροληψία.

5.4 Μοντέλο μικτών επιδράσεων

Είναι γνωστό ότι η κατανομή των CD4 λεμφοκυττάρων είναι δεξιά λοξή και για αυτό είθισται να χρησιμοποιείται η τετραγωνική ρίζα των CD4 σε αναλύσεις του μετρούμενου αυτού δείκτη. Στην συνέχεια της παρούσας εργασίας θα χρησιμοποιούμε την τετραγωνική ρίζα των CD4 λεμφοκυττάρων αντί για τα CD4 λεμφοκύτταρα.

Στόχος δεν είναι η εύρεση του καλύτερου μοντέλου αλλά η κατανόηση του ρυθμού πτώσης των CD4 λεμφοκυττάρων με την πάροδο του χρόνου. Για αυτό το λόγο στα δεδομένα εφαρμόζεται μοντέλο μικτών επιδράσεων με τυχαία σταθερά και κλίση.

Το μοντέλο που εφαρμόζεται είναι της μορφής: $Y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) * time_{ij} + e_{ij}$

Πίνακας 5.4: Μοντέλο μικτών επιδράσεων με τυχαία σταθερά και κλίση, εφαρμοσμένο στα δεδομένα της AMACS.

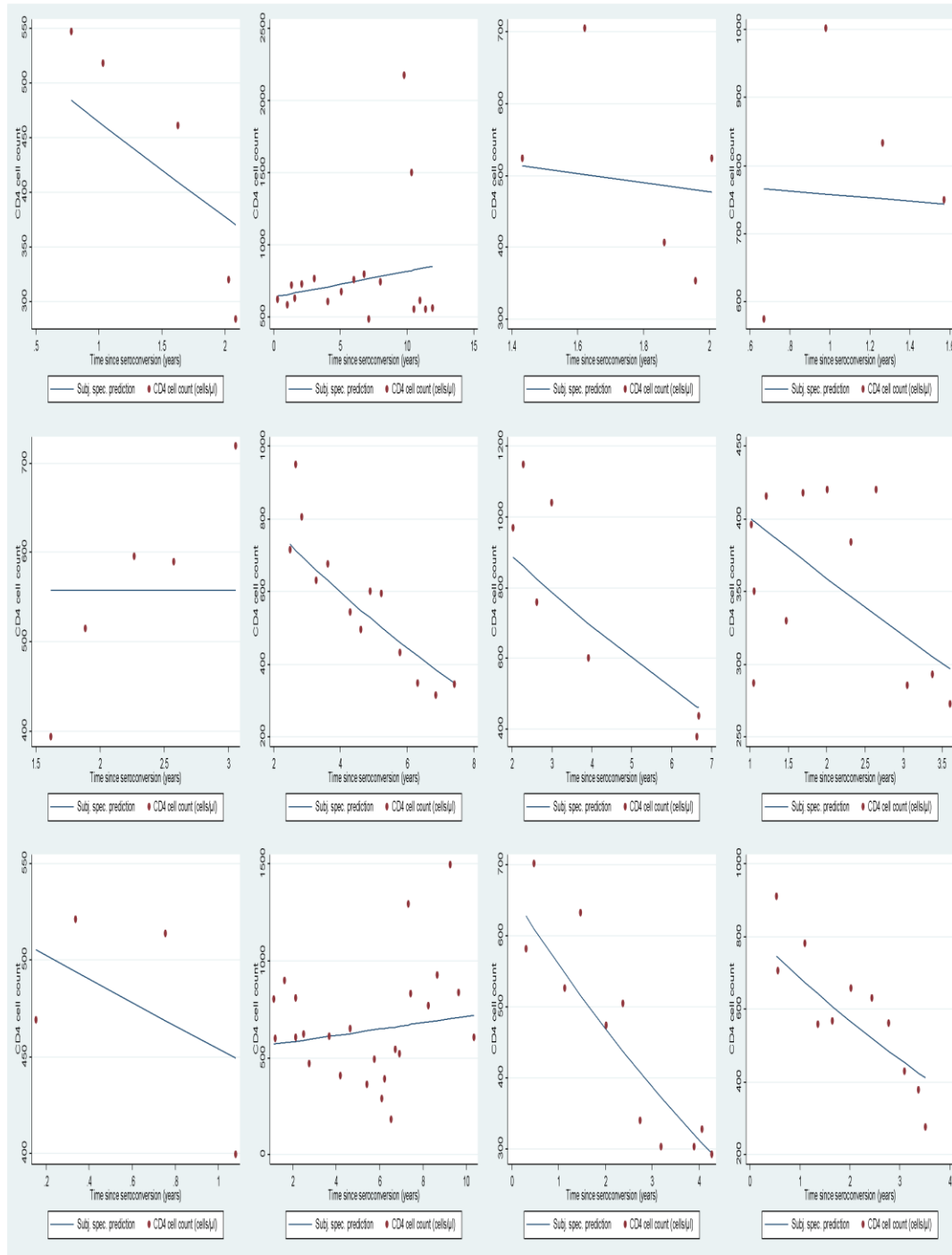
	Συντελεστής	S.E	z	P> z	[95% Δ.Ε]	
	συντελεστή					
	Σταθερές Επιδράσεις					
Σταθερά ($\sqrt{CD4}$ (κύτταρα/ μl))	23,617	0,370	63,76	<0,001	22,892	24,344
Κλίση ανά έτος	-1,288	0,183	-7,05	<0,001	-1,647	-0,930
	Τυχαίες Επιδράσεις					
Τυπική απόκλιση (σταθεράς)	4,550	0,302			3,996	5,181
Τυπική απόκλιση (κλίσης)	1,483	0,201			1,137	1,936

Το μοντέλο μπορεί να γραφεί ως:

$$Y_{ij} = (23,6 + b_{0i}) + (-1,29 + b_{1i}) * time_{ij} + e_{ij} \quad (5.4.1)$$

Είναι εμφανές από το μοντέλο 5.4.1, ότι υπάρχει πτωτική τάση στα CD4 λεμφοκύτταρα με την πάροδο του χρόνου ωστόσο αυτό εξαρτάται και από το ατομικό προφίλ του κάθε ατόμου γιατί για κάθε άτομο η πρώτη μέτρηση CD4 λεμφοκυττάρων και ο ρυθμός πτώσης τους διαφέρει. Στο γράφημα 5.4 φαίνονται τα subject specific models για την διαχρονική εξέλιξη των CD4 λεμφοκυττάρων σε δείγμα 12 ατόμων. Είναι φανερό ότι τα CD4 λεμφοκύτταρα μειώνονται με το χρόνο, ωστόσο για μερικά άτομα παρατηρείται ανοδική τάση η οποία οφείλεται κυρίως στις ατομικές τυχαίες επιδράσεις.

Γράφημα 5.4: Subject specific models για την διαχρονική εξέλιξη CD4 λεμφοκυττάρων (cells/ml) σε δείγμα 12 ατόμων.



5.5 Λογιστική παλινδρόμηση σε επαναλαμβανόμενα δεδομένα

Στα δεδομένα που αναλύουμε συνήθως παρατηρείται περικοπή μετρήσεων λόγω ανάπτυξης κλινικού AIDS, θανάτου ή έναρξης αντιρετροϊκής θεραπείας. Θα ήταν χρήσιμο να διερευνηθεί από ποιους παράγοντες επηρεάζεται η πιθανότητα έναρξης αντιρετροϊκής θεραπείας. Αν αγνοηθούν οι άλλες δυο αιτίες και θεωρηθεί ως συμβάν η έναρξη θεραπείας τότε για κάθε άτομο έχουμε μετρήσεις CD4 λεμφοκυττάρων από την ημερομηνία ορομετατροπής έως την έναρξη HAART² και μια δίτιμη μεταβλητή που δηλώνει την έναρξη θεραπείας η οποία θα αποτελείται από μηδενικά για τις μετρήσεις μέχρι την έναρξη θεραπείας και θα παίρνει την τιμή 1 όταν κάποιος ξεκινάει αντιρετροϊκή θεραπεία.

Απαραίτητη προϋπόθεση για την κατασκευή του μοντέλου είναι η διερεύνηση της σχέσης μεταξύ της μεταβλητής απόκρισης (έναρξη αντιρετροϊκής θεραπείας) και των ανεξάρτητων μεταβλητών. Χρησιμοποιούμε μονοπαραγοντικά μοντέλα λογιστικής παλινδρόμησης με εξαρτημένη μεταβλητή την έναρξη θεραπείας και ανεξάρτητες τις διαθέσιμες μεταβλητές στο υπό μελέτη δείγμα.

Πίνακας 5.5: Μονοπαραγοντικά μοντέλα λογιστικής παλινδρόμησης με εξαρτημένη μεταβλητή την έναρξη θεραπείας.

	OR	95% CI	p-value
Φύλο			0,677
Αντρας	1		
Γυναίκα	1,165	(0,567-2,394)	
Ηλικία (έτη)			<0,001
1-19	1		
20-29	3,657	(0,365-36,640)	0,270
30-39	4,800	(0,480-47,987)	0,182
40+	9,335	(0,931-93,571)	0,057
Ηλικία (έτη)	1,043	(1,027-1,059)	<0,001

² HAART: Highly Active Antiretroviral Therapy

Ηλικία (έτη)			<0,001
<40	1		
≥40	2,291	(1,639-3,204)	
Κατανάλωση αλκοόλ			0,581
Όχι	1		
Ναι	1,139	(0,716-1,812)	
Επίπεδο εκπαίδευσης			0,455
Δημοτικό	1		
Γυμνάσιο	2,375	(0,750-7,513)	0,141
Λύκειο	1,544	(0,691-3,449)	0,290
Πανεπιστήμιο ή Μεταπτυχιακό	1,303	(0,585-2,905)	0,517
Φυλή			0,410
Λευκός	1		
Μαύρος	0,556	(0,137-2,250)	
Εθνικότητα			0,430
Ελληνική	1		
Άλλη	1,264	(0,706-0,263)	
Πηγή μόλυνσης			0,0148
ΑΣΑ/ΑΣΓ*	1		
ΧΕΝ**	1,548	(0,795-3,017)	0,199
Ετεροφυλόφιλος	1,842	(1,119-3,031)	0,016
Άλλο	2,566	(1,092-6,030)	0,031
Χρόνος ορομετατροπής			0,068
>2005	1		
≤2005	1,376	(0,977-1,939)	
CD4 λεμφοκύτταρα (κύτταρα/μl)	0,999	(0,998-0,999)	0,005
CD4 λεμφοκύτταρα (κύτταρα/μl)			<0,001
>350	1		
≤350	1,963	(1,419-2,717)	

*: Άντρες με Σεξουαλική επαφή με Άντρες (ΑΣΑ), Άντρες με Σεξουαλική επαφή με Γυναίκες (ΑΣΓ)

** : Χρήστες Ενδοφλέβιων Ναρκωτικών ουσιών

Στον πίνακα 5.5 φαίνονται τα μονοπαραγοντικά μοντέλα λογιστικής παλινδρόμησης για την πιθανότητα έναρξης θεραπείας σε σχέση με κάθε μεταβλητή που υπάρχει στο δείγμα, συμπεριλαμβανομένων εκείνων που δημιουργήθηκαν από ήδη υπάρχουσες μεταβλητές όπως η ηλικία (κατηγορική). Για κάθε μεταβλητή αποφασίστηκε η μορφή με την οποία (συνεχής, δίτιμη κλπ) θα χρησιμοποιηθεί στο μοντέλο βάση του Likelihood Ratio Test θεωρώντας ως πλήρες μοντέλο εκείνο όπου η μεταβλητή είναι συνεχής (ή έχει τις περισσότερες κατηγορίες). Καταλήγουμε λοιπόν ότι η ηλικία θα χρησιμοποιηθεί ως δίτιμη (<40 & >40 ετών), η πηγή μόλυνσης θα έχει 4 κατηγορίες (ομοφυλόφιλος, XEN, ετεροφυλόφιλος, άλλο) και ο αριθμός CD4 λεμφοκυττάρων θα χρησιμοποιηθεί ως δίτιμη μεταβλητή (<350 κύτταρα/μl, >350 κύτταρα/μl).

Η επιλογή του καλύτερου μοντέλου βασίστηκε στα βήματα του Collet (Collet, 1994). Αυτή η προσέγγιση υποθέτει ότι όλες οι μεταβλητές θεωρούνται ισότιμες και δεν υπάρχει κάποιος εκ των προτέρων λόγος να συμπεριληφθεί κάποια από αυτές οπωσδήποτε στο μοντέλο. Τα βήματα της διαδικασίας του Collet (Collet, 1994) που εφαρμόστηκαν στο συγκεκριμένο δείγμα είναι:

1. Εφαρμόζουμε μονοπαραγοντικά μοντέλα και κρατάμε μόνο εκείνες τις μεταβλητές που είναι στατιστικά σημαντικές σε επίπεδο σημαντικότητας $p_1=0.20$.
2. Εφαρμόζουμε πολυπαραγοντικό μοντέλο με όλες τις στατιστικά σημαντικές μεταβλητές από το βήμα 1 και χρησιμοποιούμε την τυποποιημένη διαδικασία backward selection για να βγουν από το μοντέλο οι μεταβλητές που δεν είναι στατιστικά σημαντικές σε επίπεδο σημαντικότητας $p_2=0.10$.
3. Κρατώντας το μοντέλο που προέκυψε στο βήμα 2, χρησιμοποιούμε forward selection με επίπεδο στατιστικής σημαντικότητας $p_3=0.05$ για να διερευνήσουμε ποιες μεταβλητές που δεν ήταν στατιστικά σημαντικές στο βήμα 1 μπορούν να μπουν στο μοντέλο.
4. Εφαρμόζουμε stepwise regression με επίπεδο στατιστικής σημαντικότητας $\alpha=0.05$ για να αποφασίσουμε ποιες μεταβλητές θα παραμείνουν στο μοντέλο. Επιπλέον,

σε αυτό το βήμα διερευνούμε και αλληλεπιδράσεις μεταξύ των κυρίων επιδράσεων που υπάρχουν στο μοντέλο.

Το μοντέλο στο οποίο καταλήγουμε είναι της μορφής:

$$\log\left(\frac{\pi}{1-\pi}\right) = -2,09 + 0,78 * age_{>40} + 0,63 * CD4_{<350}$$

Πίνακας 5.6: Πολυμεταβλητό μοντέλο λογιστικής παλινδρόμησης με πολλαπλές μετρήσεις για κάθε άτομο.

	OR	95% CI	p-value
Age			
<40	1		
≥40	2,191	(1,603-2,997)	<0,001
CD4			
>350	1		
≤350	1,873	(1,359-2,583)	<0,001

Τα άτομα που έχουν CD4 λεμφοκύτταρα κάτω από 350 κύτταρα/μλ είναι 87% πιο πιθανό να πάρουν θεραπεία σε σχέση με τα άτομα που έχουν CD4 λεμφοκύτταρα πάνω από 350 κύτταρα/μλ, σταθμίζοντας για ηλικία. Αυτό είναι αναμενόμενο και συμβαδίζει με τις διεθνείς οδηγίες για την έναρξη αντιρετροϊκής θεραπείας, σύμφωνα με τις οποίες τα HIV οροθετικά άτομα με απόλυτο αριθμό CD4 λεμφοκυττάρων μικρότερο ή ίσο από 350 κύτταρα/μλ πρέπει να ενταχθούν σε πρόγραμμα χορήγησης αντιρετροϊκής θεραπείας.

Τα άτομα ηλικίας άνω των 40 ετών είναι 2 φορές πιο πιθανό να πάρουν αντιρετροϊκή θεραπεία σε σχέση με τα άτομα ηλικίας κάτω των 40 ετών, σταθμίζοντας για CD4 λεμφοκύτταρα.

5.6 Ανάλυση Επιβίωσης

Σε επαναλαμβανόμενα δεδομένα εκτός από τις μετρήσεις του μετρούμενου δείκτη υπάρχει πληροφορία και για το χρόνο μέχρι το συμβάν. Εδώ συμβάν θεωρούμε την ανάπτυξη κλινικού AIDS, τον θάνατο ή την έναρξη θεραπείας. Εφαρμόζουμε λογαριθμοκανονικό μοντέλο ανάλυσης επιβίωσης και για την επιλογή του καλύτερου μοντέλου ακολουθούμε τα βήματα της διαδικασίας του Collet που αναφέρθηκε παραπάνω.

Στο λογαριθμοκανονικό μοντέλο (Log normal) η συνάρτηση επιβίωσης είναι:

$$S(t) = 1 - \Phi \left\{ \frac{\ln(t) - \mu}{\sigma} \right\} \quad (5.6.1)$$

όπου Φ είναι η αθροιστική συνάρτηση κατανομής της τυποποιημένης κανονικής κατανομής και $\mu = X\beta$. Η συνάρτηση πυκνότητας είναι:

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp \left[-\frac{1}{2\sigma^2} \{\ln(t) - \mu\}^2 \right] \quad (5.6.2)$$

και το hazard rate είναι:

$$h(t) = \frac{\frac{1}{t\sigma\sqrt{2\pi}} \exp \left[-\frac{1}{2\sigma^2} \{\ln(t) - \mu\}^2 \right]}{1 - \Phi \left\{ \frac{\ln(t) - \mu}{\sigma} \right\}} \quad (5.6.3)$$

Η σχέση μπορεί να γραφτεί ως: $\log(T_i) = \beta_{AFT} \mathbf{K}_i + \sigma \varepsilon$

όπου ε είναι μια τυχαία μεταβλητή από την κανονική κατανομή, σ είναι scale factor, Y είναι ο λογάριθμος της μεταβλητής του χρόνου επιβίωσης και $\beta_{AFT} = -\sigma\beta_e$ όπου β_e προέρχεται από το hazard $h = \exp(\beta \mathbf{K})$.

Πίνακας 5.7: Αποτελέσματα ανάλυσης επιβίωσης εφαρμόζοντας το λογαριθμοκανονικό μοντέλο για δεδομένα της AMACS.

Μεταβλητή	Συντελεστής	S.E συντελεστή	z	P> z	[95% Διάστημα Εμπιστοσύνης]	
Ηλικία						
<40 κατ. αναφοράς						
≥40	-0,433	0,191	-2,27	0,023	-0,808	-0,059
Σταθερά	0,882	0,090	9,77	0,000	0,705	1,058
σ	1,169	0,063			1,051	1,299

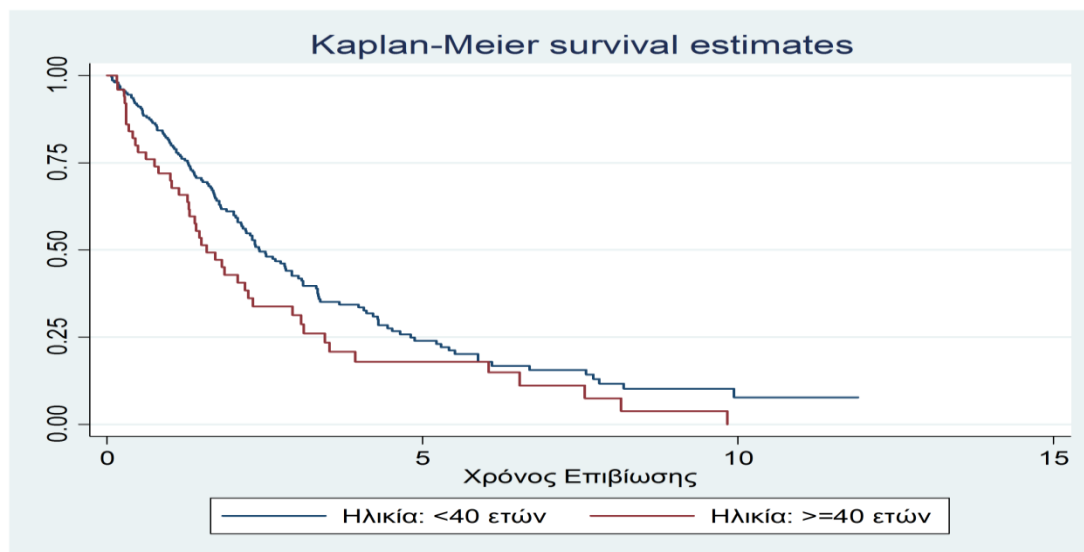
Το μοντέλο μπορεί να γραφτεί ως:

$$\ln(t) = 0,882 - 0,433 * age_{40} + 1,169 * \varepsilon$$

Με βάση τα αποτελέσματα του μοντέλου η ηλικία αποτελεί στατιστικά σημαντικό προγνωστικό παράγοντα και συγκεκριμένα για τα άτομα ηλικίας άνω των 40 ετών ο μέσος χρόνος επιβίωσης είναι το 0,65 του μέσου χρόνου επιβίωσης των ατόμων ηλικίας κάτω των 40 ετών.

Για να γίνει πιο κατανοητή η διαφορά στο χρόνο επιβίωσης παραθέτουμε την καμπύλη επιβίωσης Kaplan-Meier.

Γράφημα 5.5: Καμπύλη επιβίωσης Kaplan Meier, ανάλογα με την ηλικιακή ομάδα.



Είναι εμφανές ότι τα άτομα ηλικίας κάτω των 40 ετών έχουν καλύτερη επιβίωση σε σχέση με τα άτομα ηλικίας άνω των 40 ετών.

5.7 JMRE μοντέλο σε πραγματικά δεδομένα

Εάν θεωρήσουμε ότι η αποκοπή παρατηρήσεων λόγω έναρξης θεραπείας οδηγεί σε πληροφοριακή περικοπή τότε θα ήταν προτιμότερο να εφαρμοστούν Joint μοντέλα που λαμβάνουν υπόψη την διαχρονική εξέλιξη του μελετώμενου δείκτη αλλά και το χρόνο μέχρι το συμβάν (Pantazis & Touloumi (2010) και Crowther et al (2012)).

Εφαρμόζουμε JMRE μοντέλο το οποίο αποτελείται από ένα μοντέλο τυχαίων επιδράσεων για τον δείκτη και ένα λογαριθμοκανονικό μοντέλο επιβίωσης για τον χρόνο μέχρι το συμβάν. Το μοντέλο του δείκτη είναι ένα μοντέλο τυχαίων επιδράσεων με τυχαία σταθερά και κλίση, γιατί όπως είδαμε και παραπάνω κάθε άτομο έχει αρχικές τιμές και ρυθμό πτώσης του δείκτη που αποκλίνουν από τις μέσες τιμές του συνόλου των ατόμων. Το μοντέλο για τον χρόνο επιβίωσης είναι ένα μοντέλο λογαριθμοκανονικής ανάλυσης επιβίωσης το οποίο επιτρέπει ο γεωμετρικός μέσος του χρόνου επιβίωσης (δηλαδή μέχρι το θάνατο ή ανάπτυξη AIDS ή έναρξη θεραπείας) να διαφέρει ανάλογα με την ηλικιακή ομάδα στην ορομετατροπή.

Τα αποτελέσματα του jmre μοντέλου φαίνονται στον πίνακα 5.8.

Πίνακας 5.8: Αποτελέσματα μοντέλου JMRE εφαρμοσμένα στα δεδομένα της AMACS.

Σταθερές Επιδράσεις						
	Συντελεστής	S.E	z	P> z	[95% Δ.Ε]	
συντελεστή						
A. CD4 μοντέλο						
Σταθερά ($\sqrt{\text{κύτταρα}/\mu\text{l}}$)	25,006	0,366	68,30	<0,001	24,288	25,724
Κλίση ανά έτος	-3,024	0,215	-14,04	<0,001	-3,446	-2,602
Ηλικία						
<40						
≥40	-0,087	0,887	-0,10	0,922	-1,826	1,652
Ηλικία * χρόνος	-0,938	0,516	-1,82	0,069	-1,948	0,0729
B. Μοντέλο χρόνου επιβίωσης						
Σταθερά (χρόνια)	0,886	0,085	10,39	<0,001	0,719	1,053
Ηλικία						
<40						
≥40	-0,442	0,189	-2,34	0,019	-0,813	-0,0715
Τυχαιές Επιδράσεις (Πίνακας Διασποράς Συνδιασποράς)						
	$\sqrt{CD4}$ σταθερά	$\sqrt{CD4}$ κλίση	Υπόλοιπα μοντέλου επιβίωσης			
$\sqrt{CD4}$ σταθερά	1,417					
$\sqrt{CD4}$ κλίση	1,341	17,319				
Υπόλοιπα μοντέλου επιβίωσης	2,819	1,805	6,234			
Διακύμανση 1 ^{ου} επιπέδου ($\sqrt{CD4}$): 9,324						

Σημαντική πληροφορία δίνεται και από τον πίνακα συσχετίσεων των τυχαίων επιδράσεων του μοντέλου του δείκτη με τα κατάλοιπα του μοντέλου επιβίωσης.

Πίνακας 5.9: Πίνακας συσχετίσεων των τυχαίων επιδράσεων με τα κατάλοιπα του χρόνου επιβίωσης.

	$\sqrt{CD4}$ σταθερά	$\sqrt{CD4}$ κλίση	Υπόλοιπα μοντέλου επιβίωσης
$\sqrt{CD4}$ σταθερά	1		
$\sqrt{CD4}$ κλίση	0,271	1	
Υπόλοιπα μοντέλου επιβίωσης	0,949	0,174	1

Από τον πίνακα 5.9 φαίνεται ότι τα κατάλοιπα του μοντέλου επιβίωσης έχουν υψηλή συσχέτιση με τις τυχαίες επιδράσεις του μοντέλου του μετρούμενου δείκτη ($\text{Corr}(\sqrt{CD4} \text{ σταθερά, υπόλοιπα μοντέλου επιβίωσης})=0,948$). Αυτό συνεπάγεται ότι τα άτομα που έχουν υψηλότερες τιμές στην κλίση (άρα λιγότερο απότομη η πτώση των CD4 λεμφοκυττάρων) τείνουν να παίρνουν αντιρετροϊκή θεραπεία ή να αναπτύσσουν AIDS ή να πεθάνουν αργότερα. Η συσχέτιση με τις αρχικές τιμές του δείκτη είναι επίσης θετική [$\text{Corr}(\sqrt{CD4} \text{ σταθερά, } \sqrt{CD4} \text{ κλίση})=0,271$] αλλά λιγότερο ισχυρή.

Εφαρμόζουμε ξανά JMRE μοντέλο αλλά με την διαφορά ότι θέτουμε τις συσχετίσεις αυτές να είναι μηδέν. Συγκρίνοντας αυτά τα δυο μοντέλα χρησιμοποιώντας Likelihood Ratio Test μπορούμε να αξιολογήσουμε την στατιστική σημαντικότητα των ανωτέρω συσχετίσεων. Προκύπτει ότι οι συσχετίσεις διαφέρουν στατιστικά σημαντικά από το μηδέν (LR test $X^2=169,78$, $df=2$, $p\text{-value}<0,001$). Επομένως, συμπεραίνουμε ότι η πρόωρη διακοπή των μετρήσεων του δείκτη εξαιτίας της έναρξης αντιρετροϊκής θεραπείας ή ανάπτυξης AIDS ή θανάτου, οδηγεί σε πληροφοριακή αποκοπή.

Συγκρίνοντας το μοντέλο μικτών επιδράσεων σε σχέση με το JMRE μοντέλο είναι φανερό ότι η εκτίμηση της κλίσης που εκφράζει στην ουσία τον ρυθμό πτώσης των CD4 λεμφοκυττάρων, στο δεύτερο μοντέλο είναι πολύ πιο απότομη (-3,02) σε σχέση με την εκτίμηση του πρώτου (-1,29) μοντέλου. Στόχος είναι να ερμηνεύσουμε τον ρυθμό πτώσης των CD4 λεμφοκυττάρων για αυτό το λόγο πρέπει να αποφασίσουμε ποια αποτελέσματα θα κρατήσουμε ως τελικά, αυτά που δίνονται από το μοντέλο τυχαίων επιδράσεων ή αυτά που δίνονται από το JMRE μοντέλο.

Στη συνέχεια παρουσιάζονται αποτελέσματα προσομοιώσεων για την σύγκριση των μοντέλων μικτών επιδράσεων και του JMRE μοντέλου υπό διαφορετικά σενάρια θεραπείας. Η απόφαση για το ποιο μοντέλο είναι καταλληλότερο θα βασιστεί στα αποτελέσματα των προσομοιώσεων.

Κεφάλαιο 6 : Προσομοιώσεις

Προσομοιωμένα δεδομένα

Σε αυτό το κεφάλαιο γίνεται διερεύνηση των μοντέλων μικτών επιδράσεων και των joint μοντέλων χρησιμοποιώντας προσομοιωμένα δεδομένα. Στόχος είναι να διερευνηθεί η σχετική απόδοση των δυο αυτών τύπων μοντέλων όταν εφαρμόζονται σε επαναλαμβανόμενες μετρήσεις CD4 κυττάρων οι οποίες περικόπτονται λόγω της έναρξης θεραπείας, με την περικοπή να προκύπτει από διάφορους τυχαίους μηχανισμούς (MCAR, MAR, MNAR).

Οι παράμετροι που χρησιμοποιήθηκαν για την προσομοίωση των δεδομένων (σταθερές επιδράσεις, πίνακας διακύμανσης-συνδιακύμανσης) βασίστηκαν στα αποτελέσματα ενός μικτού γραμμικού μοντέλου με τυχαία σταθερά και κλίση που είχε εφαρμοστεί σε πραγματικά δεδομένα της μελέτης CASCADE. Η CASCADE είναι μια μελέτη στην οποία συμμετέχουν 23 μελέτες κοορτής από 15 Ευρωπαϊκές χώρες, την Αυστραλία και τον Καναδά και σε αυτή περιλαμβάνονται άτομα που έχουν καλά εκτιμημένη ημερομηνία ορομετατροπής. Χρησιμοποιήσαμε παραμέτρους από την μελέτη CASCADE επειδή ο αριθμός των ατόμων είναι πολύ μεγαλύτερος σε σύγκριση με την μελέτη AMACS, και κατά συνέπεια οι εκτιμήσεις αυτών είναι πιο έγκυρες. Η διάρκεια παρακολούθησης ορίστηκε στα 5 χρόνια με μετρήσεις των CD4 λεμφοκυττάρων κάθε τρεις μήνες και κάποιες προαιρετικές επισκέψεις για μέτρηση CD4 λεμφοκυττάρων στο μεσοδιάστημα των τριών μηνών των υποχρεωτικών μετρήσεων. Στόχος είναι η εκτίμηση του ρυθμού μεταβολής των CD4 λεμφοκυττάρων από την ημερομηνία ορομετατροπής μέχρι την έναρξη θεραπείας. Ο μέσος αριθμός της τετραγωνικής ρίζας CD4 λεμφοκυττάρων θεωρήθηκε ότι μειώνεται γραμμικά κατά $1,59 \sqrt{\text{κύτταρα}}/\mu\text{l}$ το χρόνο, ενώ η μέση πληθυσμιακή σταθερά τέθηκε $24,2 \sqrt{\text{κύτταρα}}/\mu\text{l}$, που αντιστοιχεί σε 585 κύτταρα/ μl κατά την ορομετατροπή. Ο πίνακας διακύμανσης-συνδιακύμανσης των τυχαίων επιδράσεων αποτελείται από την διακύμανση της σταθεράς που ορίστηκε να είναι 29,3, η διακύμανση της κλίσης είναι 2,8 με την συνδιακύμανση τους ίση με -3,1. Η τιμή των υπολοίπων επιπέδου 1 τέθηκε 5,5.

Προσομοιώνουμε τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων λόγω έναρξης θεραπείας.

1. Η απόσυρση από την μελέτη να συμβαίνει εντελώς τυχαία (**MCAR**): Σε αυτή την περίπτωση οι περικομμένες παρατηρήσεις είναι εντελώς τυχαίες και συμβαίνουν λόγω σχεδιασμού της μελέτης. Στο σενάριο αυτό η απόσυρση από την μελέτη θεωρείται ανεξάρτητη από τις τιμές των CD4 λεμφοκυττάρων. Οι χρόνοι απόσυρσης ακολουθούν εκθετική κατανομή (ως ενδιάμεσοι χρόνοι μιας Poisson διαδικασίας) με μέσο ρυθμό 5% το χρόνο. Άρα οι χρόνοι απόσυρσης C_i παράγονται από την σχέση $C_i \approx -(1/0,05) * \ln(U)$, όπου $U \sim U(0,1)$.
2. Η απόσυρση από την μελέτη να συμβαίνει τυχαία (**MAR**): Στο σενάριο αυτό εκτός από τον εντελώς τυχαίο μηχανισμό του σεναρίου 1 θεωρήθηκε ότι τα άτομα της μελέτης αποσύρονται όταν τα επίπεδα CD4 λεμφοκυττάρων πέσουν κάτω από 350 κύτταρα/μλ. Αυτό το σενάριο βασίζεται στις διεθνείς οδηγίες για έναρξη αντιρετροϊκής θεραπείας όταν για πρώτη φορά παρατηρούνται CD4 λεμφοκύτταρα κάτω από 350 κύτταρα/μλ.
3. Η απόσυρση από την μελέτη να συμβαίνει με τρόπο μη αγνοήσιμο (**MNAR 1**): Στο σενάριο αυτό εκτός από τον εντελώς τυχαίο μηχανισμό του σεναρίου 1 θεωρήθηκε ότι τα άτομα της μελέτης αποσύρονται όταν τα επίπεδα CD4 λεμφοκυττάρων πέσουν κάτω από 350 κύτταρα/μλ ή αν τα επίπεδα είναι πάνω από αυτή την ποσότητα, το άτομα αποσύρονται με βάση τα πραγματικά CD4 (παρατηρηθέντα CD4 - κατάλοιπα).
4. Η απόσυρση από την μελέτη να συμβαίνει με τρόπο μη αγνοήσιμο (**MNAR 2**): Είναι ακριβώς ίδιο με το σενάριο MNAR 1 με την διαφορά ότι εδώ προστίθεται και ένας μη παρατηρηθείς όρος frailty, που θεωρητικά είναι μια ποσότητα που δεν παρατηρείται και θα μπορούσε να είναι η ψυχολογία του ασθενούς. Για τους σκοπούς της ανάλυσης έχει θεωρηθεί ότι ο όρος frailty ακολουθεί την τυποποιημένη κανονική κατανομή.

Αρχικά θεωρήσαμε ότι ο όρος frailty είναι ασυσχέτιστος με τα random effects. Στη συνέχεια επαναλάβαμε τις προσομοιώσεις θεωρώντας ότι ο όρος frailty είναι αρνητικά/θετικά συσχετισμένος με τα random effects.

Σε όλα τα ανωτέρω σενάρια θεωρείται ότι για κάθε ασθενή υπάρχουν 4 επισκέψεις τον χρόνο (μια ανά τρίμηνο) οι οποίες είναι υποχρεωτικές και υπάρχουν κάποιες προαιρετικές επισκέψεις που γίνονται με πιθανότητα 10% (1^ο σενάριο προαιρετικών επισκέψεων).

Τα ίδια σενάρια εφαρμόστηκαν ξανά χρησιμοποιώντας δυο διαφορετικά ενδεχόμενα για την πιθανότητα προαιρετικής επίσκεψης.

- ❖ Η πιθανότητα να πραγματοποιηθεί μια προαιρετική επίσκεψη εξαρτάται από τα observed CD4 στην προηγούμενη διαθέσιμη επίσκεψη (2^ο σενάριο προαιρετικών επισκέψεων).
- ❖ Η πιθανότητα να πραγματοποιηθεί μια προαιρετική επίσκεψη εξαρτάται από τα observed CD4 στην προηγούμενη διαθέσιμη επίσκεψη αλλά και έναν όρο frailty (3^ο σενάριο προαιρετικών επισκέψεων).

Τέλος, το ποσοστό σχετικής μεροληψίας (% Relative Bias) και η πιθανότητα κάλυψης (Coverage Probability) είναι τα δυο μέτρα που χρησιμοποιήθηκαν για να αποτυπωθεί η μεροληψία των εκτιμήσεων στα προσομοιωμένα δεδομένα. Το ποσοστό σχετικής μεροληψίας είναι η μεροληψία προς την πραγματική τιμή πολλαπλασιασμένη με 100%. Ενώ η πιθανότητα κάλυψης είναι η συχνότητα με την οποία περιλαμβάνεται η πραγματική τιμή στο διάστημα εμπιστοσύνης αν επαναλάβουμε το ίδιο πείραμα πολλές φορές.

6.1 Σύγκριση μοντέλου μικτών επιδράσεων σε σχέση με το JMRE μοντέλο για τα 4 διαφορετικά σενάρια θεραπείας και με πιθανότητα προαιρετικής επίσκεψης 10%.

6.1.1 Αποτελέσματα προσομοιώσεων

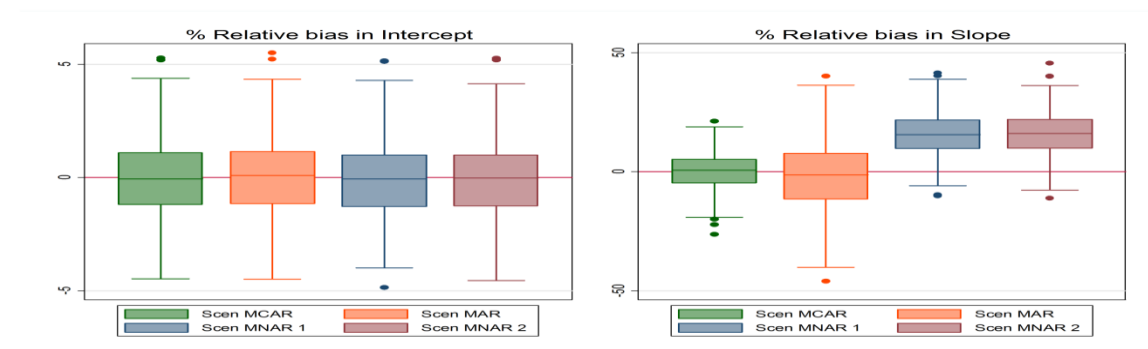
Πίνακας 6.1: Αποτελέσματα προσομοιώσεων (500 επαναλήψεις με 200 άτομα/επανάληψη) για τα 4 διαφορετικά σενάρια θεραπείας. Μέγιστος χρόνος παρακολούθησης 5 χρόνια με 4 υποχρεωτικές επισκέψεις ανά χρόνο για κάθε ασθενή και οι προαιρετικές επισκέψεις πραγματοποιούνται με πιθανότητα 10%. Οι αληθινές (True) παράμετροι προέρχονται από πραγματικά δεδομένα.

Mixed model							
Scenario	Parameter	TRUE	Mean Estim.	% Rel. Bias	SD(estim.)	Mean SE(estim.)	95% Cov. Prob
MCAR	intercept	24,236	24,232	-0,02	0,386	0,39	96,00 (94,28-97,72)
	slope	-1,589	-1,583	0,38	0,125	0,124	94,20 (92,14-96,26)
MAR	intercept	24,236	24,247	0,04	0,395	0,397	96,00 (94,28-97,72)
	slope	-1,589	-1,612	-1,43	0,228	0,165	86,20 (83,17- 89,23)
MNAR 1	intercept	24,236	24,216	-0,09	0,386	0,389	95,40 (93,56-97,24)
	slope	-1,589	-1,339	15,71	0,14	0,137	55,40 (51,03- 59,77)
MNAR 2	intercept	24,236	24,223	-0,05	0,386	0,389	95,60 (93,80- 97,40)
	slope	-1,589	-1,335	15,98	0,145	0,139	57,20 (52,85-61,55)
JMRE model							
Scenario	Parameter	TRUE	Mean Estim.	% Rel. Bias	SD(estim.)	Mean SE(estim.)	95% Cov. Prob
MCAR	intercept	24,236	24,242	0,02	0,386	0,389	96,00 (94,28-97,72)
	slope	-1,589	-1,617	-1,77	0,129	0,124	92,60 (90,30-94,90)
MAR	intercept	24,236	24,325	0,37	0,397	0,401	95,40 (93,56-97,24)
	slope	-1,589	-3,363	-111,64	0,312	0,173	0,00 (0,00-0,00)
MNAR 1	intercept	24,236	24,232	-0,02	0,389	0,387	95,80 (94,04- 97,56)
	slope	-1,589	-1,555	2,14	0,16	0,135	89,40 (86,69-92,11)
MNAR 2	intercept	24,236	24,238	0,01	0,389	0,388	95,40 (93,56-97,24)
	slope	-1,589	-1,562	1,68	0,166	0,137	89,00 (86,25-91,75)

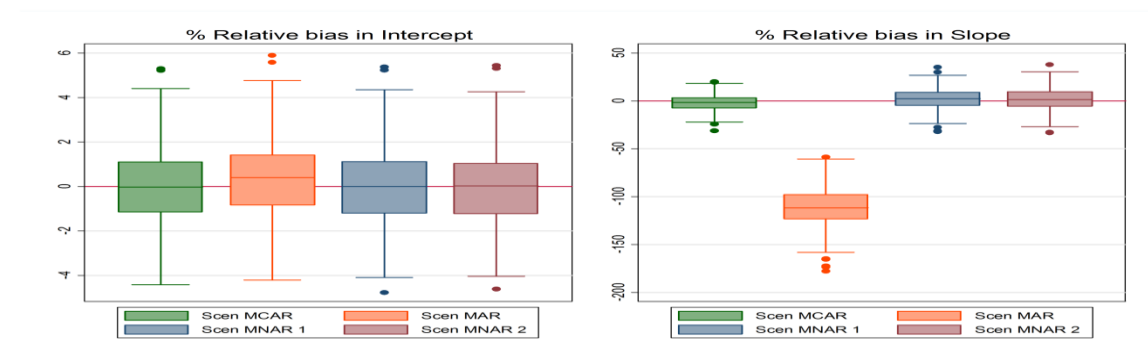
Είναι εμφανές ότι το μοντέλο τυχαίων επιδράσεων και το JMRE μοντέλο δίνουν αμερόληπτες εκτιμήσεις για την σταθερά σε όλα τα διαφορετικά σενάρια θεραπείας. Όπως αναμενόταν το μοντέλο τυχαίων επιδράσεων σε περιπτώσεις μη τυχαίων ελλειπουσών τιμών (MNAR) υποεκτιμά τον ρυθμό πτώσης των CD4 λεμφοκυττάρων δίνοντας εκτιμήσεις για την κλίση οι οποίες έχουν μεροληψία της τάξης του 15%. Αντίθετα, παρατηρείται ότι σε MAR ελλείπουσες τιμές το JMRE μοντέλο υπερεκτιμά τον ρυθμό πτώσης των CD4 λεμφοκυττάρων (Πίνακας 6.1). Ως συνέπεια του εισαγόμενου σφάλματος, τα 95% Coverage Probability είναι χαμηλά στην περίπτωση MNAR ελλειπουσών τιμών όταν εφαρμόζουμε μοντέλα τυχαίων επιδράσεων ή ιδιαίτερα χαμηλά όταν εφαρμόζουμε JMRE μοντέλα στην περίπτωση MAR ελλειπουσών τιμών. Εποπτική παρουσίαση των αποτελεσμάτων με το σφάλμα δίνεται στα γραφήματα 6.1.1 έως 6.1.4.

6.1.2 Box plots για την μεροληψία της σταθεράς και της κλίσης των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων

Γράφημα 6.1.1: Box plot για την σχετική μεροληψία της σταθεράς και της κλίσης των μοντέλων μικτών επιδράσεων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.

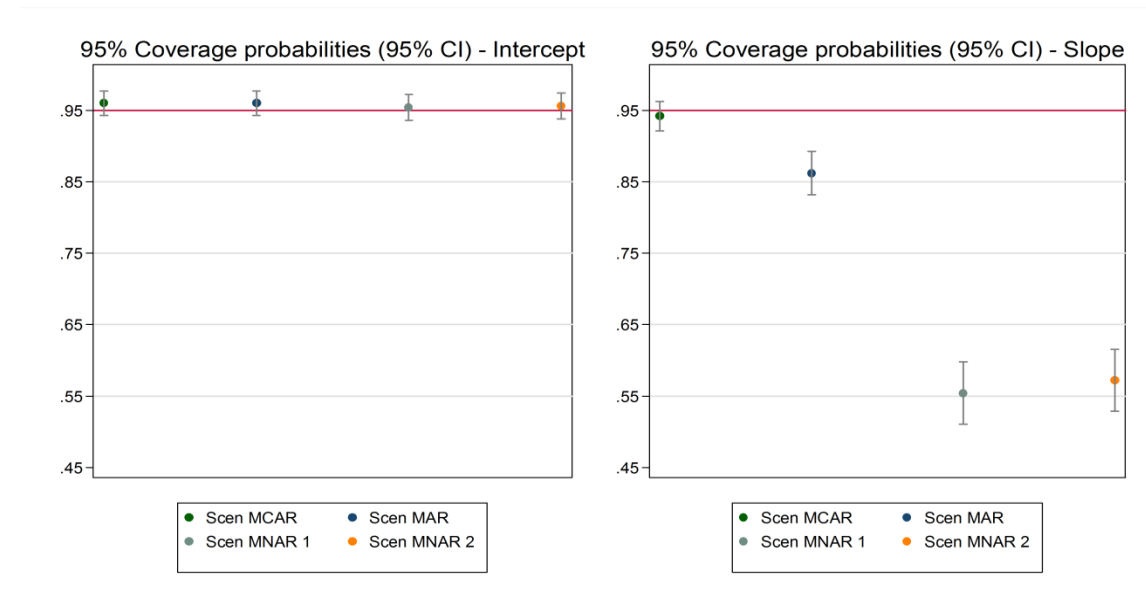


Γράφημα 6.1.2: Box plot για την σχετική μεροληψία της σταθεράς και της κλίσης των JMRE μοντέλων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.

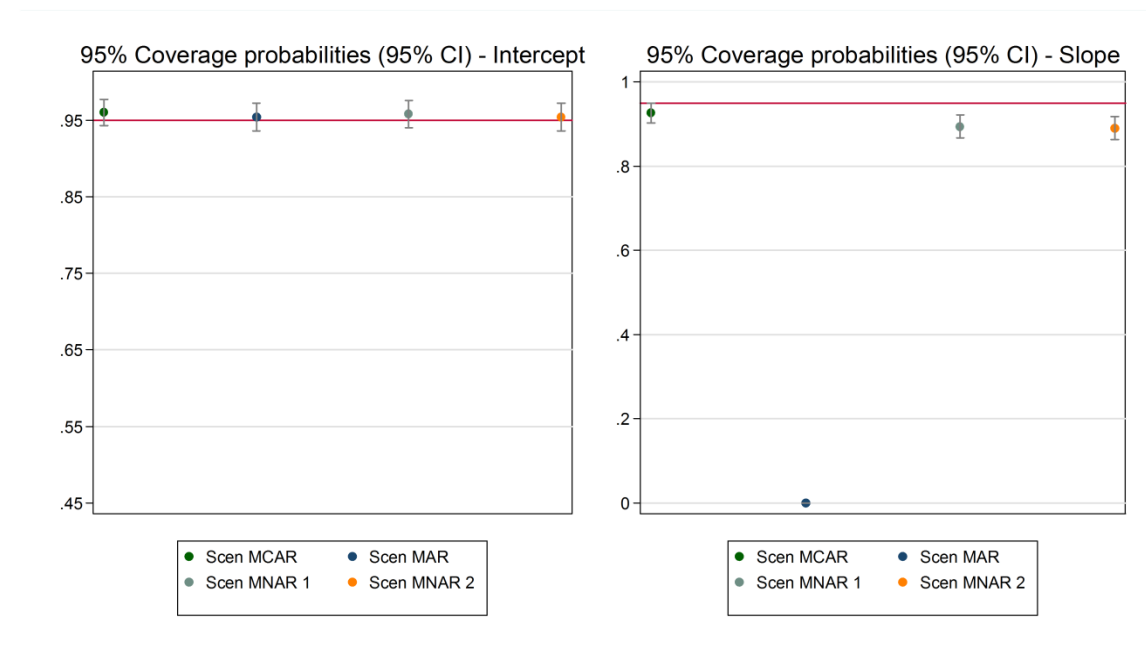


6.1.3 95% διάστημα εμπιστοσύνης για την σταθερά και την κλίση των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων.

Γράφημα 6.1.3: 95% Διάστημα Εμπιστοσύνης για την σταθερά και την κλίση των μοντέλων μικτών επιδράσεων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.



Γράφημα 6.1.4: 95% Διάστημα Εμπιστοσύνης για την σταθερά και την κλίση των JMRE μοντέλων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.



6.2 Σύγκριση μοντέλου μικτών επιδράσεων σε σχέση με το JMRE μοντέλο για τα 4 διαφορετικά σενάρια θεραπείας και με πιθανότητα προαιρετικής επίσκεψης που εξαρτάται από τα παρατηρηθέντα CD4 στην προηγούμενη επίσκεψη.

6.2.1 Αποτελέσματα προσομοιώσεων

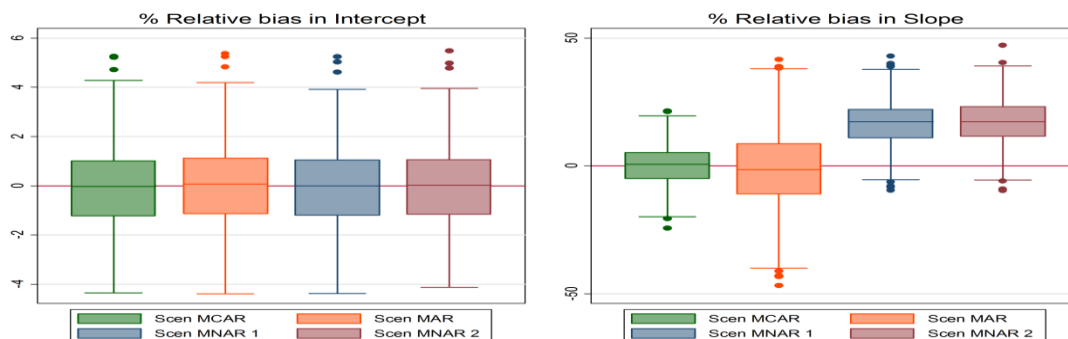
Πίνακας 6.2: Αποτελέσματα προσομοιώσεων (500 επαναλήψεις με 200 άτομα/επανάληψη) για τα 4 διαφορετικά σενάρια θεραπείας. Μέγιστος χρόνος παρακολούθησης 5 χρόνια με 4 υποχρεωτικές επισκέψεις ανά χρόνο για κάθε ασθενή και η πιθανότητα πραγματοποίησης μιας προαιρετικής επίσκεψης εξαρτάται από τα παρατηρηθέντα CD4 στην προηγούμενη διαθέσιμη επίσκεψη. Οι αληθινές (True) παράμετροι προέρχονται από πραγματικά δεδομένα.

Mixed Model							
Scenario	Parameter	TRUE	Mean Estim.	% Rel. Bias	SD(estim.)	Mean SE(estim.)	95% Cov. Prob
MCAR	intercept	24,236	24,233	-0,01	0,384	0,389	95,60 (93,80- 97,40)
	slope	-1,589	-1,584	0,27	0,124	0,124	93,80 (91,68- 95,92)
MAR	intercept	24,236	24,245	0,03	0,395	0,396	96,20 (94,52-97,88)
	slope	-1,589	-1,609	-1,27	0,237	0,167	82,00 (78,62-85,38)
MNAR 1	intercept	24,236	24,236	0	0,379	0,386	96,00 (94,28- 97,72)
	slope	-1,589	-1,319	16,95	0,143	0,138	50,40 (46,00-54,80)
MNAR 2	intercept	24,236	24,244	0,03	0,377	0,386	96,20 (94,52- 97,88)
	slope	-1,589	-1,311	17,47	0,147	0,141	51,00 (46,60-55,40)
JMRE model							
Scenario	Parameter	TRUE	Mean Estim.	% Rel. Bias	SD(estim.)	Mean SE(estim.)	95% Cov. Prob
MCAR	intercept	24,236	24,242	0,02	0,384	0,388	95,60 (93,80-97,40)
	slope	-1,589	-1,616	-1,7	0,127	0,123	93,20 (90,99- 95,41)
MAR	intercept	24,236	24,325	0,37	0,394	0,398	95,40 (93,56-97,24)
	slope	-1,589	-3,479	-118,98	0,316	0,173	0,00 (0,00-0,00)
MNAR 1	intercept	24,236	24,285	0,2	0,384	0,384	95,40 (93,56- 97,24)
	slope	-1,589	-1,724	-8,53	0,177	0,136	77,40 (73,72-81,08)
MNAR 2	intercept	24,236	24,304	0,28	0,373	0,384	96,20 (94,52- 97,88)
	slope	-1,589	-1,746	-9,87	0,187	0,138	74,60 (70,77- 78,43)

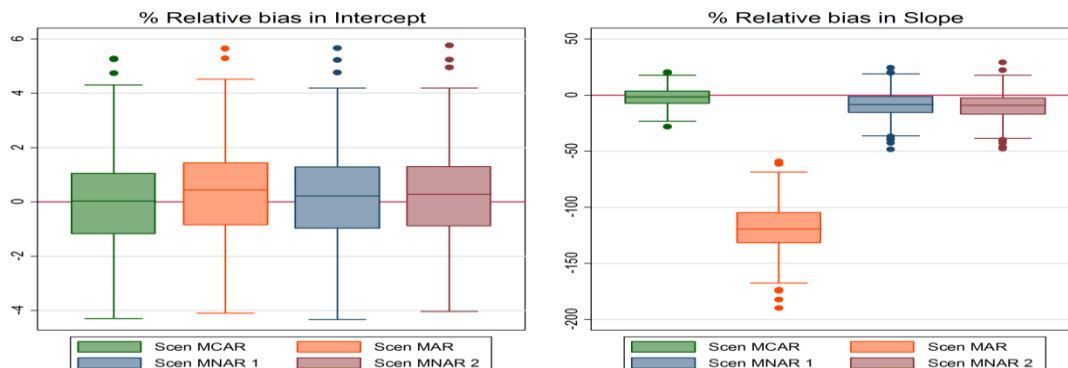
Στην περίπτωση όπου η πιθανότητα προαιρετικών επισκέψεων εξαρτάται από τα παρατηρηθέντα CD4 λεμφοκύτταρα στην προηγούμενη επίσκεψη, τα αποτελέσματα των προσομοιώσεων είναι παρόμοια με αυτά που παρουσιάστηκαν στο κεφάλαιο 6.1. Ωστόσο, εδώ παρατηρείται μεγαλύτερο σφάλμα όταν εφαρμόζονται μοντέλα τυχαίων επιδράσεων και JMRE μοντέλα σε MNAR ελλείπουσες τιμές καθώς εισάγεται και μηχανισμός MAR στον οποίο έχει ιδιαίτερη άσχημη εφαρμογή το JMRE μοντέλο. Τα αποτελέσματα του πίνακα 6.2 παρουσιάζονται και στα γραφήματα 6.2.2 έως 6.2.4, όπου παρατηρείται ότι υπάρχει υψηλό bias όταν εφαρμόζουμε μοντέλα τυχαίων επιδράσεων σε MNAR ελλείπουσες τιμές και JMRE μοντέλα σε MAR ελλείπουσες τιμές.

6.2.2 Box plots για την μεροληψία της σταθεράς και της κλίσης των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων.

Γράφημα 6.2.1: Box plot για την σχετική μεροληψία της σταθεράς και της κλίσης των μοντέλων μικτών επιδράσεων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.

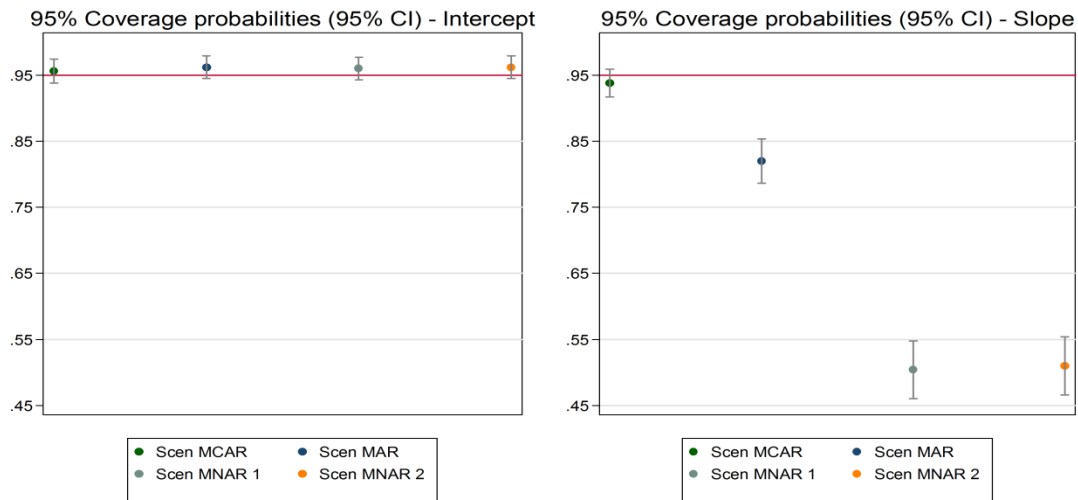


Γράφημα 6.2.2: Box plot για την σχετική μεροληψία της σταθεράς και της κλίσης των JMRE μοντέλων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.

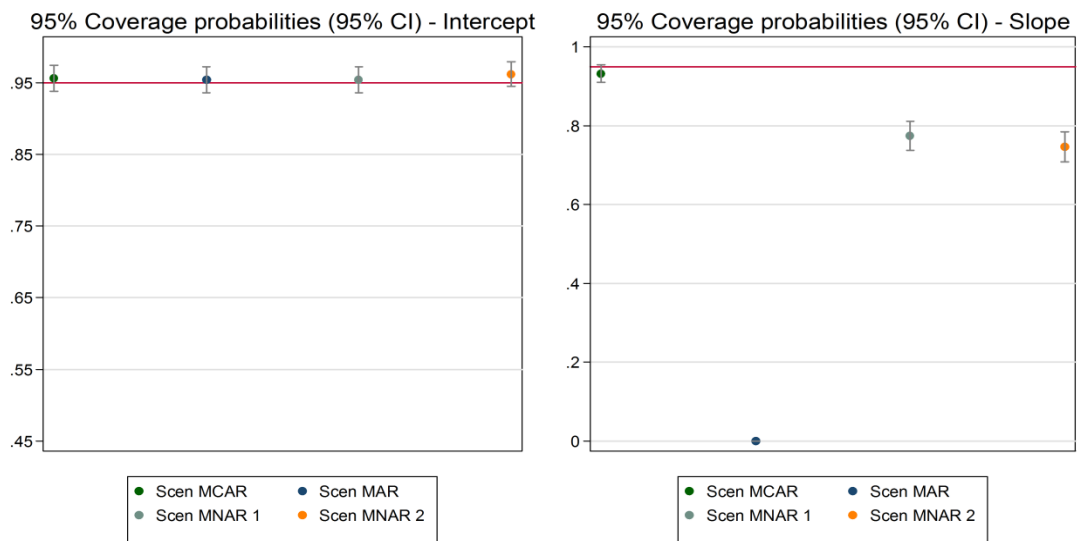


6.2.3 95% διάστημα εμπιστοσύνης για την σταθερά και την κλίση των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων.

Γράφημα 6.2.3: 95% Διάστημα Εμπιστοσύνης για την σταθερά και την κλίση των μοντέλων μικτών επιδράσεων, για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.



Γράφημα 6.2.4: Διάστημα Εμπιστοσύνης για την σταθερά και την κλίση των JMRE μοντέλων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.



6.3 Σύγκριση μοντέλου μικτών επιδράσεων σε σχέση με το JMRE μοντέλο για τα 4 διαφορετικά σενάρια θεραπείας και με πιθανότητα προαιρετικής επίσκεψης που εξαρτάται από τα παρατηρηθέντα CD4 στην προηγούμενη επίσκεψη και έναν όρο frailty.

6.3.1 Αποτελέσματα προσομοιώσεων

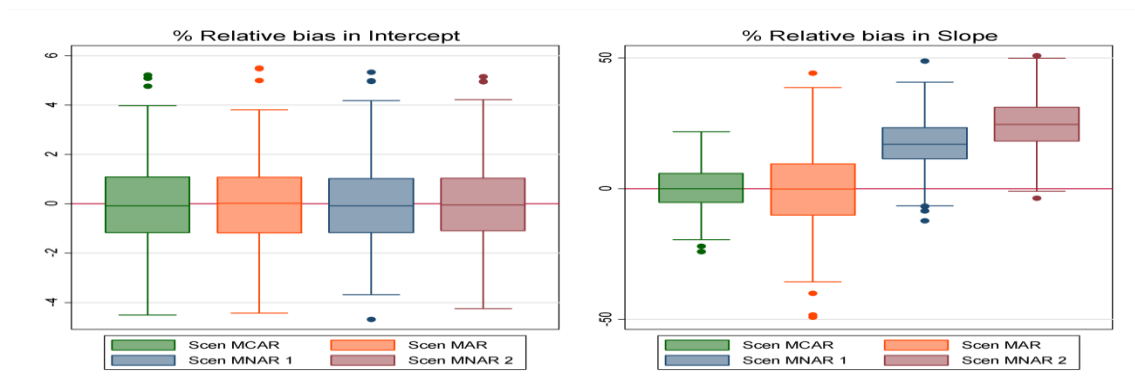
Πίνακας 6.3: Αποτελέσματα προσομοιώσεων (500 επαναλήψεις με 200 άτομα/επανάληψη) για τα 4 διαφορετικά σενάρια θεραπείας. Μέγιστος χρόνος παρακολούθησης 5 χρόνια με 4 υποχρεωτικές επισκέψεις ανά χρόνο για κάθε ασθενή και η πιθανότητα πραγματοποίησης μιας προαιρετικής επίσκεψης εξαρτάται από τα παρατηρηθέντα CD4 στην προηγούμενη διαθέσιμη επίσκεψη. Οι αληθινές (True) παράμετροι προέρχονται από πραγματικά δεδομένα.

Mixed Model							
Scenario	Parameter	TRUE	Mean Estim.	% Rel. Bias	SD(estim.)	Mean SE(estim.)	95% Cov. Prob
MCAR	intercept	24,236	24,231	-0,02	0,384	0,388	96,00 (94,28-97,72)
	slope	-1,589	-1,588	0,05	0,127	0,123	94,60 (92,61-96,59)
MAR	intercept	24,236	24,238	0,01	0,394	0,397	95,40 (93,56- 97,24)
	slope	-1,589	-1,593	-0,24	0,241	0,167	81,60 (78,19-85,01)
MNAR 1	intercept	24,236	24,226	-0,04	0,383	0,386	96,00 (94,28-97,72)
	slope	-1,589	-1,315	17,25	0,147	0,138	49,80 (45,40- 54,20)
MNAR 2	intercept	24,236	24,235	-0,01	0,38	0,386	96,40 (94,76- 98,04)
	slope	-1,589	-1,201	24, 42	0,149	0,139	21,60 (17,98-25,22)
JMRE model							
Scenario	Parameter	TRUE	Mean Estim.	% Rel. Bias	SD(estim.)	Mean SE(estim.)	95% Cov. Prob
MCAR	intercept	24,236	24,241	0,02	0,384	0,387	96,00 (94,28- 97,72)
	slope	-1,589	-1,62	-1,97	0,132	0,123	92,00 (89,61-94,39)
MAR	intercept	24,236	24,322	0,35	0,396	0,399	94,40 (92,38- 96,42)
	slope	-1,589	-3,458	-117,68	0,317	0,172	0,00 (0,00- 0,00)
MNAR 1	intercept	24,236	24,273	0,15	0,383	0,384	95,40 (93,56- 97,24)
	slope	-1,589	-1,719	-8,2	0,178	0,135	76,20 (72,45- 79,95)
MNAR 2	intercept	24,236	24,294	0,24	0,386	0,384	95,20 (93,32-97,08)
	slope	-1,589	-1,678	-5,63	0,193	0,136	76,40 (72,67- 80,13)

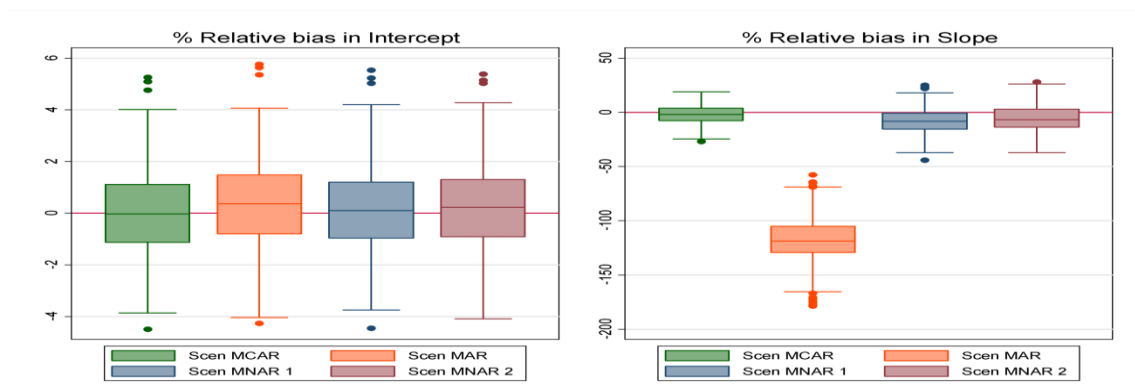
Στην περίπτωση όπου η πιθανότητα προαιρετικών επισκέψεων εξαρτάται από τα παρατηρηθέντα CD4 λεμφοκύτταρα και έναν όρο frailty παρατηρείται εντονότερο σφάλμα στις εκτιμήσεις όταν εφαρμόζουμε μοντέλα τυχαίων επιδράσεων σε MNAR ελλείπουσες τιμές. Είναι εμφανές ότι όσο πιο έντονος είναι ο μηχανισμός δημιουργίας ελλειπουσών τιμών MAR τόσο πιο έντονο είναι το σφάλμα. Ως αποτέλεσμα του εισαγόμενου σφάλματος, τα 95% Coverage Probability είναι ιδιαίτερα χαμηλά στην περίπτωση MNAR όταν εφαρμόζουμε μοντέλα τυχαίων επιδράσεων και στην περίπτωση MAR όταν εφαρμόζουμε JMRE μοντέλα. Τα αποτελέσματα των προσομοιώσεων παρατίθενται στον πίνακα 6.3 ενώ ακολουθεί εποπτική παρουσίασή τους στα γραφήματα 6.3.1 έως 6.3.2.

6.3.2 Box plots για την μεροληψία της σταθεράς και της κλίσης των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων.

Γράφημα 6.3.1: Box plot για την σχετική μεροληψία της σταθεράς και της κλίσης των μοντέλων μικτών επιδράσεων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.

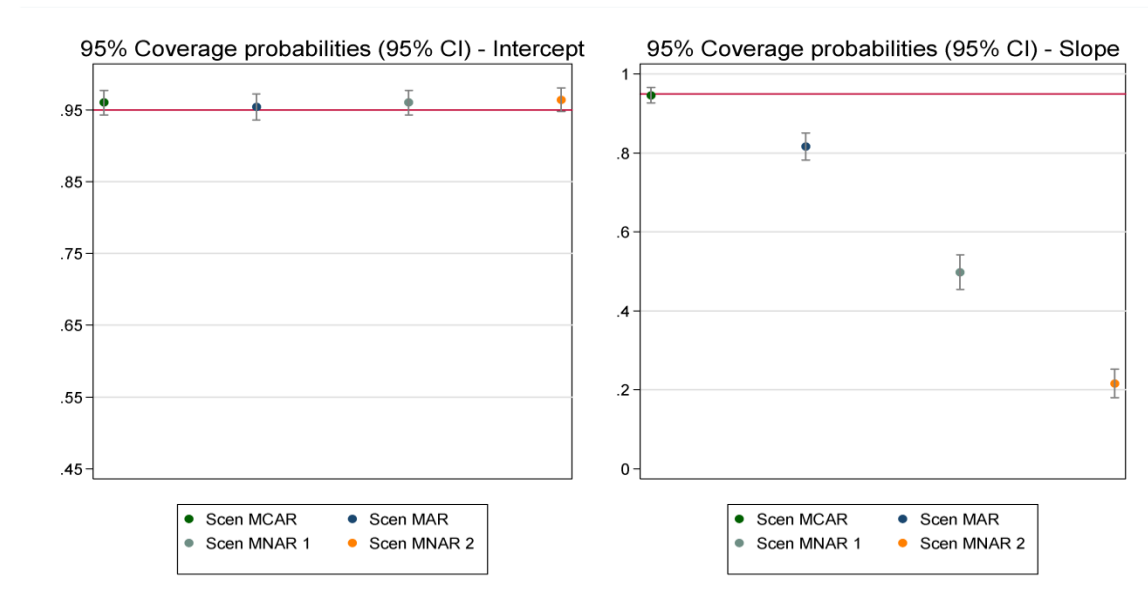


Γράφημα 6.3.2: Box plot για την σχετική μεροληψία της σταθεράς και της κλίσης των JMRE μοντέλων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.

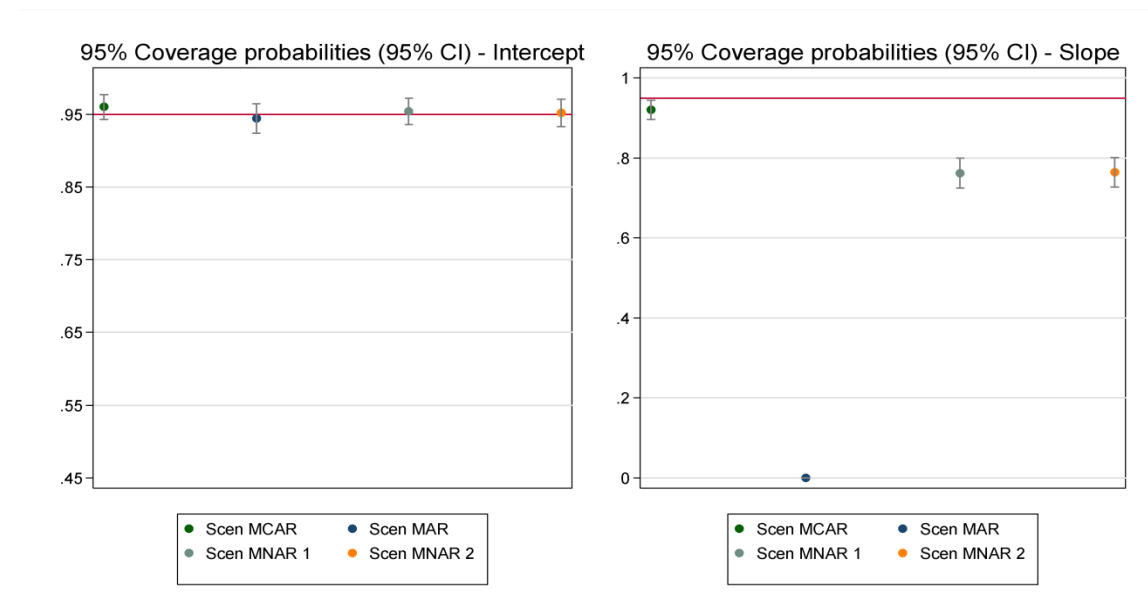


6.3.3 95% διάστημα εμπιστοσύνης για την σταθερά και την κλίση των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.

Γράφημα 6.3.3: 95% Διάστημα Εμπιστοσύνης για την σταθερά και την κλίση των μοντέλων μικτών επιδράσεων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.



Γράφημα 6.3.4: 95% Διάστημα Εμπιστοσύνης για την σταθερά και την κλίση των JMRE μοντέλων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.



Στην περίπτωση όπου η έναρξη θεραπείας οδηγεί σε MAR ελλείπουσες τιμές τότε τα JMRE μοντέλα υπερεκτιμούν τον ρυθμό πτώσης των CD4 λεμφοκυττάρων δίνοντας εκτιμήσεις για την κλίση οι οποίες έχουν μεροληψία 110%. Η υπερεκτίμηση της κλίσης σε περίπτωση MAR ελλειπουσών τιμών, συμβαίνει εξαιτίας της υπόθεσης που κάνει το μοντέλο JMRE, δηλαδή ότι οι τυχαίες επιδράσεις σχετίζονται με τα κατάλοιπα του μοντέλου επιβίωσης. Συγκεκριμένα, ύστερα από διερεύνηση των συσχετίσεων στα 500 σετ δεδομένων για όλα τα σενάρια επισκέψεων που αναφέρθηκαν προκύπτει ότι η μέση συσχέτιση της τυχαίας σταθεράς (random intercept) με τα κατάλοιπα του χρόνου μέχρι την αποκοπή είναι 0,8 ενώ η μέση συσχέτιση της τυχαίας κλίσης (random slope) με τα κατάλοιπα του χρόνου μέχρι την αποκοπή είναι σχεδόν 0,85. Και οι δυο συσχετίσεις είναι πολύ υψηλές και για αυτό το λόγο το JMRE μοντέλο διορθώνει για πληροφοριακή αποκοπή γιατί βλέπει αυτά τα δεδομένα ως MNAR ενώ στην πραγματικότητα είναι MAR (βλ. Πίνακα 6.3.1).

Πίνακας 6.3.1: Περιγραφικός πίνακας της μέσης συσχέτισης όπως αυτή εκτιμάται από το JMRE μοντέλο σε κάθε προσομοιωμένο dataset, για όλα τα διαφορετικά σενάρια επισκέψεων με την έναρξη της θεραπείας να δημιουργεί MAR ελλείπουσες τιμές.

Σενάριο επισκέψε ων	Συσχέτιση	Μέση τιμή	Τυπική απόκλιση	Διάμεσος	Ενδοτεταρτομοριακό εύρος (25 ^ο , 75 ^ο)
1	Corr ($\sqrt{CD4}$ σταθερά $\sqrt{CD4}$ κλίση)	0,798	0,033	0,799	(0,776-0,818)
	Corr($\sqrt{CD4}$ σταθερά υπόλοιπα μοντέλου επιβίωσης)	0,854	0,040	0,859	(0,832-0,882)
2	Corr ($\sqrt{CD4}$ σταθερά $\sqrt{CD4}$ κλίση)	0,811	0,031	0,813	(0,793-0,833)
	Corr($\sqrt{CD4}$ σταθερά υπόλοιπα μοντέλου επιβίωσης)	0,861	0,038	0,867	(0,841-0,888)
3	Corr ($\sqrt{CD4}$ σταθερά $\sqrt{CD4}$ κλίση)	0,812	0,030	0,813	(0,792-0,832)
	Corr($\sqrt{CD4}$ σταθερά υπόλοιπα μοντέλου επιβίωσης)	0,857	0,037	0,859	(0,836-0,882)

Πίνακας 6.3.2: Περιγραφικός πίνακας της μέσης συσχέτισης όπως αυτή εκτιμάται από το JMRE μοντέλο σε κάθε προσομοιωμένο dataset, για όλα τα διαφορετικά σενάρια επισκέψεων με την έναρξη της θεραπείας να δημιουργεί MNAR ελλείπουσες τιμές.

Σενάριο επισκέψεων	Συσχέτιση	Μέση τιμή	Τυπική απόκλιση	Διάμεσος	Ενδοτεταρτομοριακό εύρος (25°, 75°)
1	Corr($\sqrt{CD4}$ σταθερά $\sqrt{CD4}$ κλίση)	0,351	0,063	0,351	(0,306-0,395)
	Corr($\sqrt{CD4}$ σταθερά υπόλοιπα μοντέλου επιβίωσης)	0,179	0,111	0,182	(0,111- 0,255)
2	Corr($\sqrt{CD4}$ σταθερά $\sqrt{CD4}$ κλίση)	0,425	0,061	0,425	(0,381-0,469)
	Corr($\sqrt{CD4}$ σταθερά υπόλοιπα μοντέλου επιβίωσης)	0,346	0,116	0,346	(0,264-0,422)
3	Corr($\sqrt{CD4}$ σταθερά $\sqrt{CD4}$ κλίση)	0,425	0,068	0,422	(0,377-0,473)
	Corr($\sqrt{CD4}$ σταθερά υπόλοιπα μοντέλου επιβίωσης)	0,343	0,122	0,352	(0,251-0,450)

Είναι εμφανές ότι σε MNAR ελλείπουσες τιμές η μέση συσχέτιση της τυχαίας κλίσης (random slope) με τα κατάλοιπα του χρόνου μέχρι την αποκοπή κυμαίνεται από 0,1 έως 0,3. Σε αυτή την περίπτωση το JMRE μοντέλο δίνει αμερόληπτες εκτιμήσεις ενώ σε περίπτωση MAR ελλειπουσών τιμών δίνει μεροληπτικές εκτιμήσεις (Πίνακας 6.3.2).

Σε όλα τα ανωτέρω σενάρια θεωρούμε ότι ο όρος frailty είναι ανεξάρτητος από τα random effects. Στο επόμενο βήμα εφαρμόζουμε τα ίδια σενάρια για την έναρξη θεραπείας και για τις επισκέψεις αλλά υποθέτουμε ότι ο όρος frailty είναι αρνητικά συσχετισμένος με τα random effects και η συσχέτιση αυτή θεωρούμε ότι είναι της τάξης του -0,5. Υποθέτουμε ότι ο όρος frailty εκφράζει τον φόβο ενός ασθενή για την εξέλιξη της νόσου του, άρα όσο πιο πολύ φοβάται τόσο πιο πιθανό είναι να έρχεται στις προαιρετικές επισκέψεις αλλά αυτό επηρεάζει και την έναρξη θεραπείας. Επομένως, όσο πιο πολύ φοβάται κάποιος θεωρούμε ότι θα πάρει θεραπεία πιο νωρίς. Στις επόμενες ενότητες δίνονται τα αποτελέσματα από τα προσομοιωμένα δεδομένα που περιγράψαμε πιο πάνω.

6.4 Σύγκριση μοντέλου μικτών επιδράσεων σε σχέση με το JMRE μοντέλο για τα 4 διαφορετικά σενάρια θεραπείας, με πιθανότητα προαιρετικής επίσκεψης 10%, και με αρνητική συσχέτιση μεταξύ frailty και random effects.

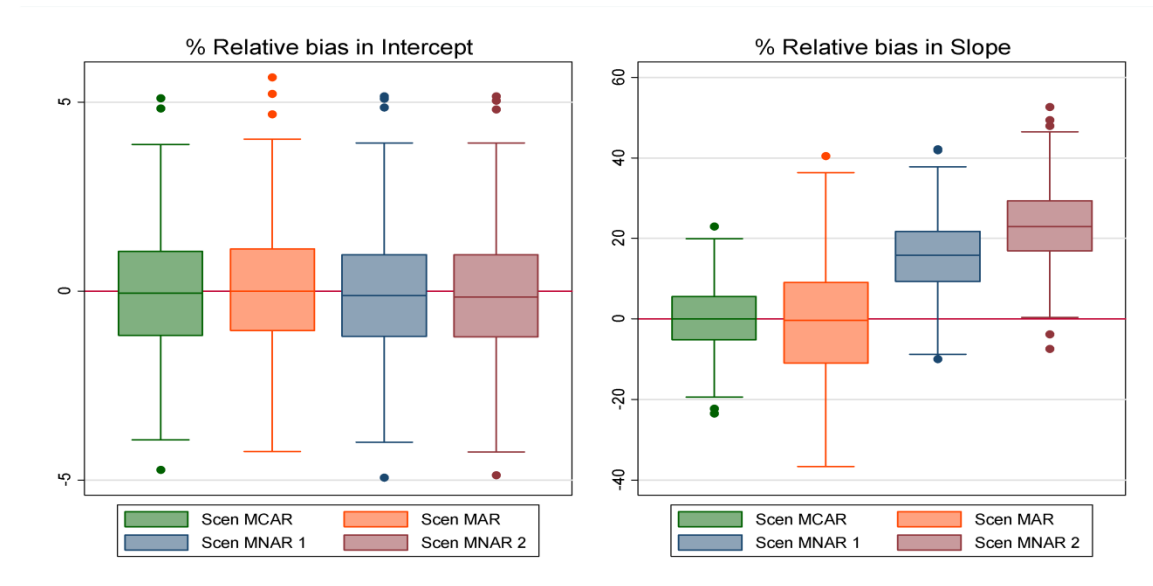
6.4.1 Αποτελέσματα προσομοιώσεων

Πίνακας 6.4: Αποτελέσματα προσομοιώσεων (500 επαναλήψεις με 200 άτομα/επανάληψη) για τα 4 διαφορετικά σενάρια θεραπείας. Μέγιστος χρόνος παρακολούθησης 5 χρόνια με 4 υποχρεωτικές επισκέψεις ανά χρόνο για κάθε ασθενή και οι προαιρετικές επισκέψεις πραγματοποιούνται με πιθανότητα 10%. Οι αληθινές (True) παράμετροι προέρχονται από πραγματικά δεδομένα.

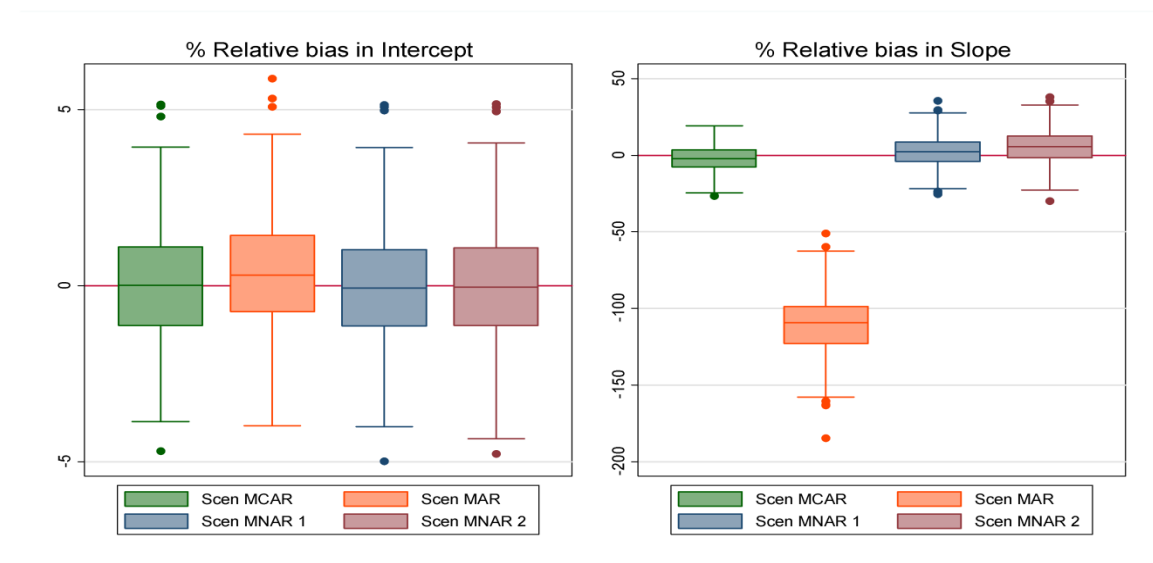
Mixed Model							
Scenario	Parameter	TRUE	Mean Estim.	% Rel. Bias	SD(estim.)	Mean SE(estim.)	95% Cov. Prob
MCAR	intercept	24,236	24,23	-0,03	0,388	0,39	95,80 (94,04-97,56)
	slope	-1,589	-1,587	0,12	0,129	0,124	94,00 (91,91-96,09)
MAR	intercept	24,236	24,239	0,01	0,395	0,397	95,20 (93,32-97,08)
	slope	-1,589	-1,603	-0,9	0,224	0,165	84,20 (80,99-87,41)
MNAR 1	intercept	24,236	24,211	-0,11	0,386	0,389	95,20 (93,3- 97,08)
	slope	-1,589	-1,341	15,59	0,141	0,137	53,80 (49,42- 58,18)
MNAR 2	intercept	24,236	24,208	-0,12	0,387	0,389	96,20 (94,52-97,88)
	slope	-1,589	-1,223	23,04	0,147	0,137	23,80 (20,05-27,55)
JMRE model							
Scenario	Parameter	TRUE	Mean Estim.	% Rel. Bias	SD(estim.)	Mean SE(estim.)	95% Cov. Prob
MCAR	intercept	24,236	24,24	0,02	0,388	0,389	95,80 (94,04-97,56)
	slope	-1,589	-1,621	-2,06	0,133	0,124	91,80 (89,39-94,21)
MAR	intercept	24,236	24,318	0,34	0,398	0,402	94,20 (92,4-96,26)
	slope	-1,589	-3,343	-110,41	0,302	0,172	0,00 (0,00-0,00)
MNAR 1	intercept	24,236	24,225	-0,05	0,39	0,388	95,40 (93,56-97,24)
	slope	-1,589	-1,548	2,55	0,16	0,135	89,40 (86,69-92,11)
MNAR 2	intercept	24,236	24,23	-0,02	0,391	0,388	95,40 (93,56-97,24)
	slope	-1,589	-1,499	5,64	0,171	0,135	84,20 (80,99-87,41)

6.4.2 Box plots για την μεροληψία της σταθεράς και της κλίσης των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.

Γράφημα 6.4.1: Box plot για την σχετική μεροληψία της σταθεράς και της κλίσης των μοντέλων μικτών επιδράσεων, για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.

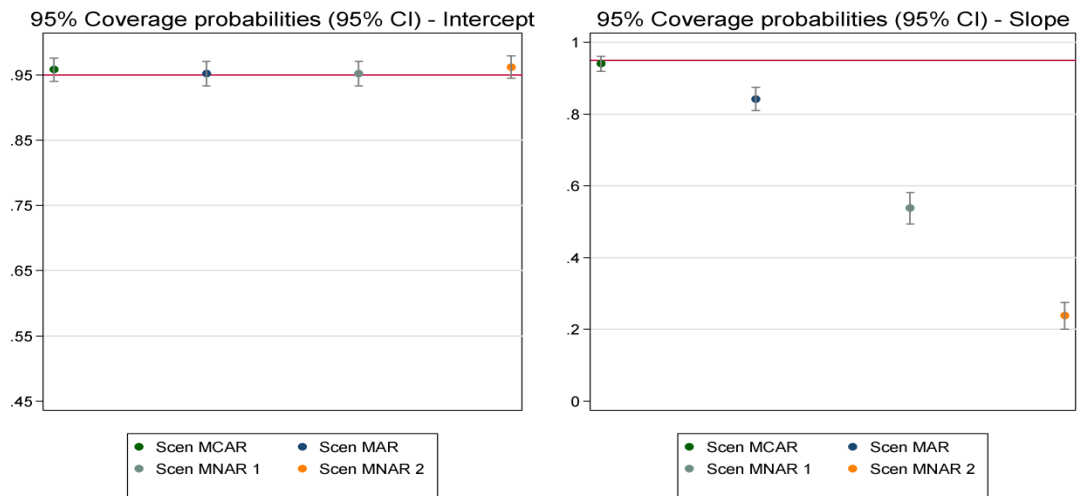


Γράφημα 6.4.2: Box plot για την σχετική μεροληψία της σταθεράς και της κλίσης των JMRE μοντέλων, για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.

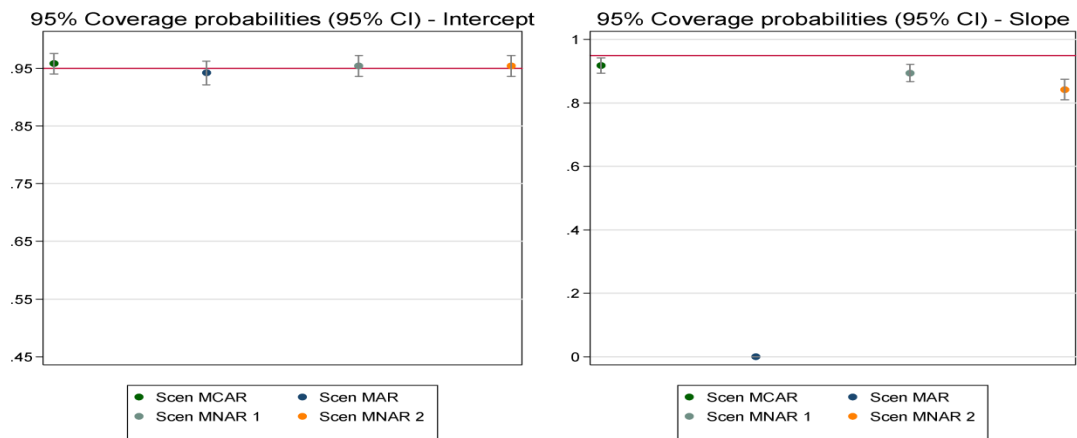


6.4.3 95% διάστημα εμπιστοσύνης για την σταθερά και την κλίση των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.

Γράφημα 6.4.3: 95% Διάστημα Εμπιστοσύνης για την σταθερά και την κλίση των μοντέλων μικτών επιδράσεων, για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.



Γράφημα 6.4.4: 95% Διάστημα Εμπιστοσύνης για την σταθερά και την κλίση των JMRE μοντέλων, για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.



6.5 Σύγκριση μοντέλου μικτών επιδράσεων σε σχέση με το JMRE μοντέλο για τα 4 διαφορετικά σενάρια θεραπείας, με πιθανότητα προαιρετικής επίσκεψης που εξαρτάται από τα παρατηρηθέντα CD4 στην προηγούμενη επίσκεψη και με αρνητική συσχέτιση μεταξύ frailty και random effects.

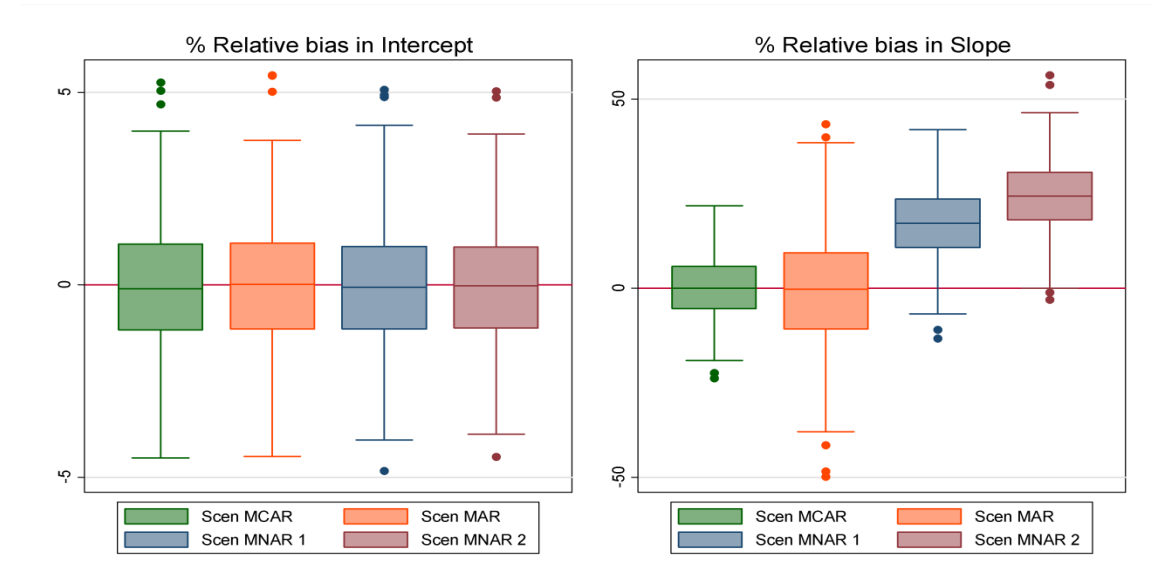
6.5.1 Αποτελέσματα προσομοιώσεων

Πίνακας 6.5: Αποτελέσματα προσομοιώσεων (500 επαναλήψεις με 200 άτομα/επανάληψη) για τα 4 διαφορετικά σενάρια θεραπείας. Μέγιστος χρόνος παρακολούθησης 5 χρόνια με 4 υποχρεωτικές επισκέψεις ανά χρόνο για κάθε ασθενή και η πιθανότητα πραγματοποίησης μιας προαιρετικής επίσκεψης εξαρτάται από τα παρατηρηθέντα CD4 στην προηγούμενη διαθέσιμη επίσκεψη. Οι αληθινές (True) παράμετροι προέρχονται από πραγματικά δεδομένα.

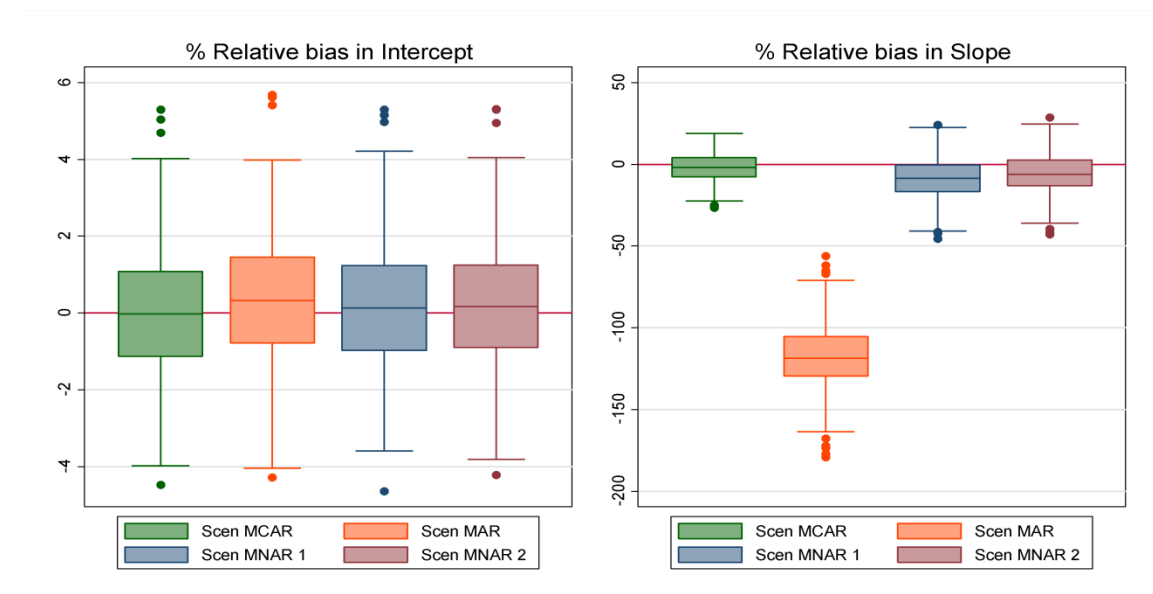
Mixed Model							
Scenario	Parameter	TRUE	Mean Estim.	% Rel. Bias	SD(estim.)	Mean SE(estim.)	95% Cov. Prob
MCAR	intercept	24,236	24,231	-0,02	0,383	0,388	96,20 (94,52-97,88)
	slope	-1,589	-1,587	0,09	0,127	0,123	94,40 (92,38-96,42)
MAR	intercept	24,236	24,238	0,01	0,393	0,397	95,40 (93,56-97,24)
	slope	-1,589	-1,602	-0,81	0,241	0,167	82,80 (79,48-86,12)
MNAR 1	intercept	24,236	24,228	-0,03	0,378	0,386	96,40 (94,76-98,04)
	slope	-1,589	-1,317	17,1	0,146	0,138	49,80 (45,40-54,20)
MNAR 2	intercept	24,236	24,233	-0,01	0,38	0,386	96,40 (94,76-98,04)
	slope	-1,589	-1,2	24,47	0,149	0,139	22,00 (18,36-25,64)
JMRE model							
Scenario	Parameter	TRUE	Mean Estim.	% Rel. Bias	SD(estim.)	Mean SE(estim.)	95% Cov. Prob
MCAR	intercept	24,236	24,24	0,02	0,383	0,387	96,20 (94,52-97,88)
	slope	-1,589	-1,619	-1,91	0,132	0,123	92,00 (89,61-94,39)
MAR	intercept	24,236	24,32	0,35	0,396	0,399	94,40 (92,38- 96,42)
	slope	-1,589	-3,464	-118,05	0,317	0,173	0,00 (0,00- 0,00)
MNAR 1	intercept	24,236	24,281	0,19	0,38	0,384	95,40 (93,56-97,24)
	slope	-1,589	-1,723	-8,48	0,188	0,136	74,40 (70,56-78,24)
MNAR 2	intercept	24,236	24,286	0,21	0,38	0,384	95,60 (93,80-97,40)
	slope	-1,589	-1,675	-5,44	0,19	0,136	79,60 (76,06-83,14)

6.5.2 Box plots για την μεροληψία της σταθεράς και της κλίσης των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.

Γράφημα 6.5.1: Box plot για την σχετική μεροληψία της σταθεράς και της κλίσης των μοντέλων μικτών επιδράσεων, για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.

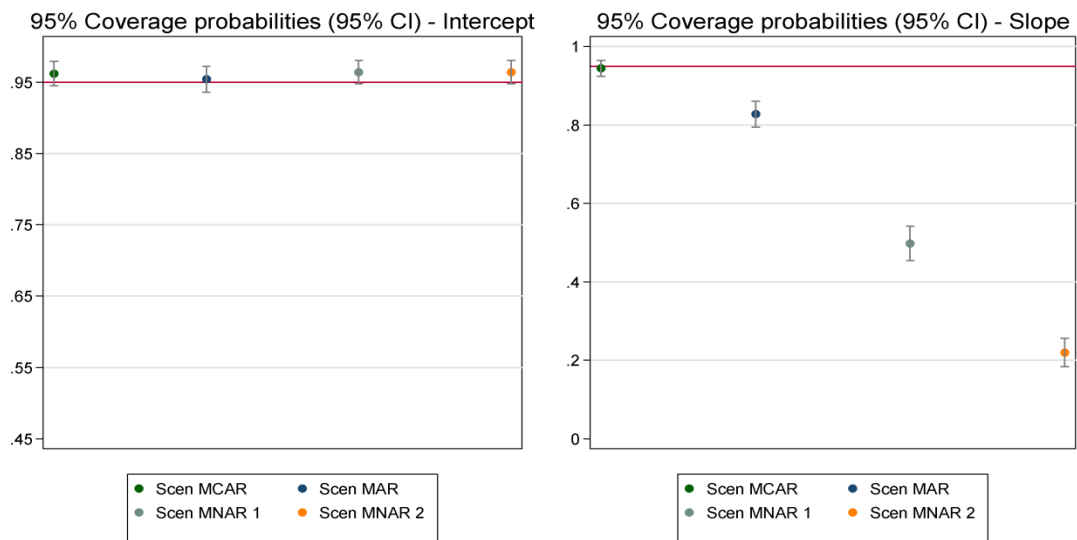


Γράφημα 6.5.2: Box plot για την σχετική μεροληψία της σταθεράς και της κλίσης των JMRE μοντέλων, για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.

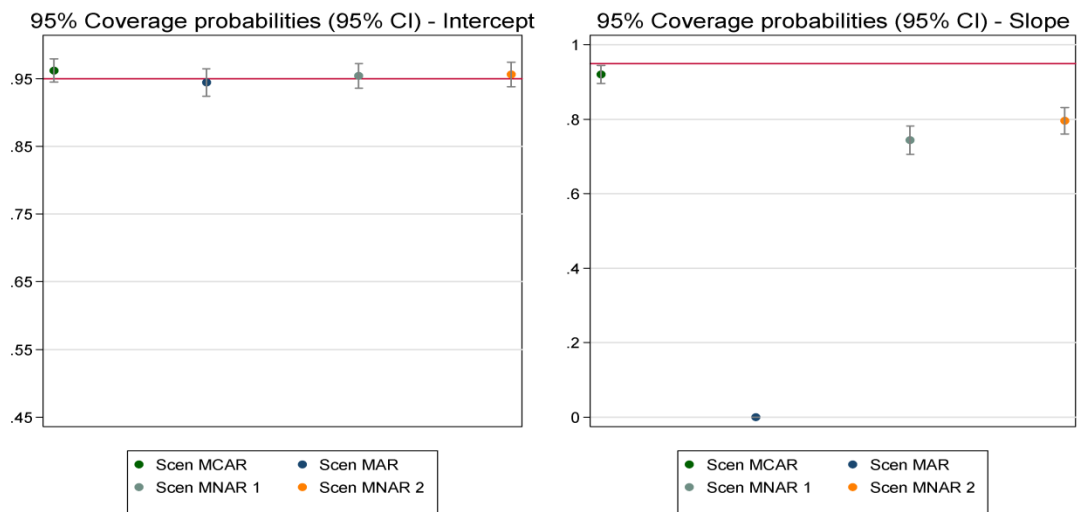


6.5.3 95% διάστημα εμπιστοσύνης για την σταθερά και την κλίση των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.

Γράφημα 6.5.3: 95% Διάστημα Εμπιστοσύνης για την σταθερά και την κλίση των μοντέλων μικτών επιδράσεων, για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.



Γράφημα 6.5.4: 95% Διάστημα Εμπιστοσύνης για την σταθερά και την κλίση των JMRE μοντέλων, για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.



6.6 Σύγκριση μοντέλου μικτών επιδράσεων σε σχέση με το JMRE μοντέλο για τα 4 διαφορετικά σενάρια θεραπείας, με πιθανότητα προαιρετικής επίσκεψης που εξαρτάται από τα παρατηρηθέντα CD4 στην προηγούμενη επίσκεψη και από έναν όρο frailty, και με αρνητική συσχέτιση μεταξύ frailty και random effects.

6.6.1 Αποτελέσματα προσομοιώσεων

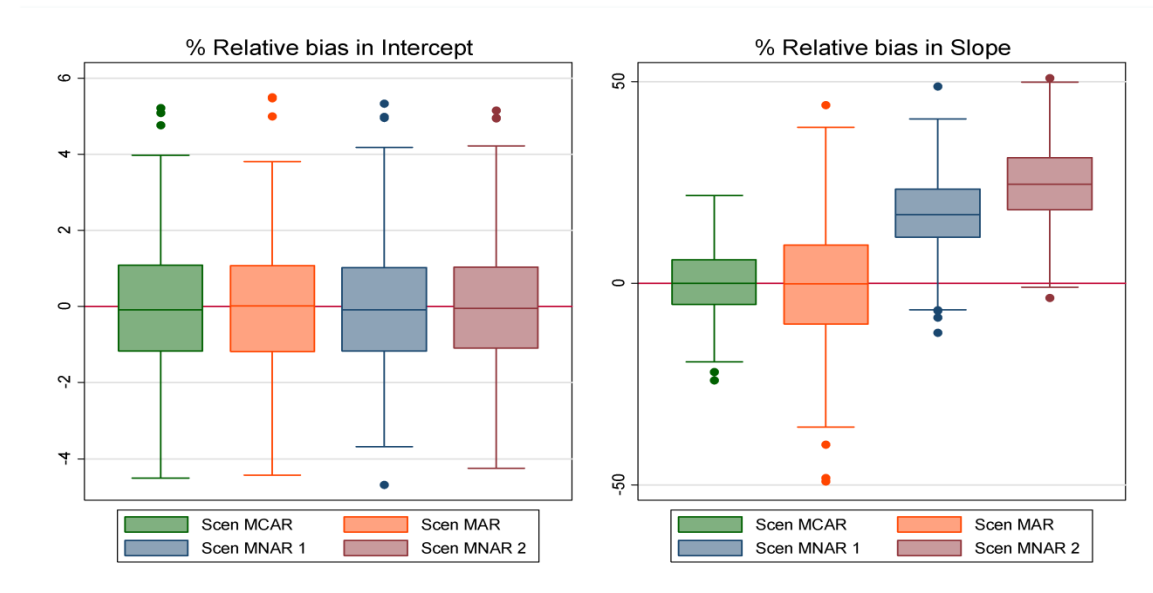
Πίνακας 6.6: Αποτελέσματα προσομοιώσεων (500 επαναλήψεις με 200 άτομα/επανάληψη) για τα 4 διαφορετικά σενάρια θεραπείας. Μέγιστος χρόνος παρακολούθησης 5 χρόνια με 4 υποχρεωτικές επισκέψεις ανά χρόνο για κάθε ασθενή και η πιθανότητα πραγματοποίησης μιας προαιρετικής επίσκεψης εξαρτάται από τα παρατηρηθέντα CD4 στην προηγούμενη διαθέσιμη επίσκεψη και έναν όρο frailty. Οι αληθινές (True) παράμετροι προέρχονται από πραγματικά δεδομένα.

Mixed Model							
Scenario	Parameter	TRUE	Mean Estim.	% Rel. Bias	SD(estim.)	Mean SE(estim.)	95% Cov. Prob
MCAR	intercept	24,236	24,231	-0,02	0,384	0,388	96,00 (94,28-97,72)
	slope	-1,589	-1,588	0,05	0,127	0,123	94,60 (92,61-96,59)
	intercept	24,236	24,238	0,01	0,394	0,397	95,40 (93,56-97,24)
MAR	slope	-1,589	-1,593	-0,24	0,241	0,167	81,60 (78,19-85,01)
	intercept	24,236	24,226	-0,04	0,383	0,386	96,00 (94,28-97,72)
MNAR 1	slope	-1,589	-1,315	17,25	0,147	0,138	49,80 (45,40-54,20)
	intercept	24,236	24,235	-0,01	0,38	0,386	96,40 (94,76-98,04)
MNAR 2	slope	-1,589	-1,201	24,42	0,149	0,139	21,60 (17,98-25,22)

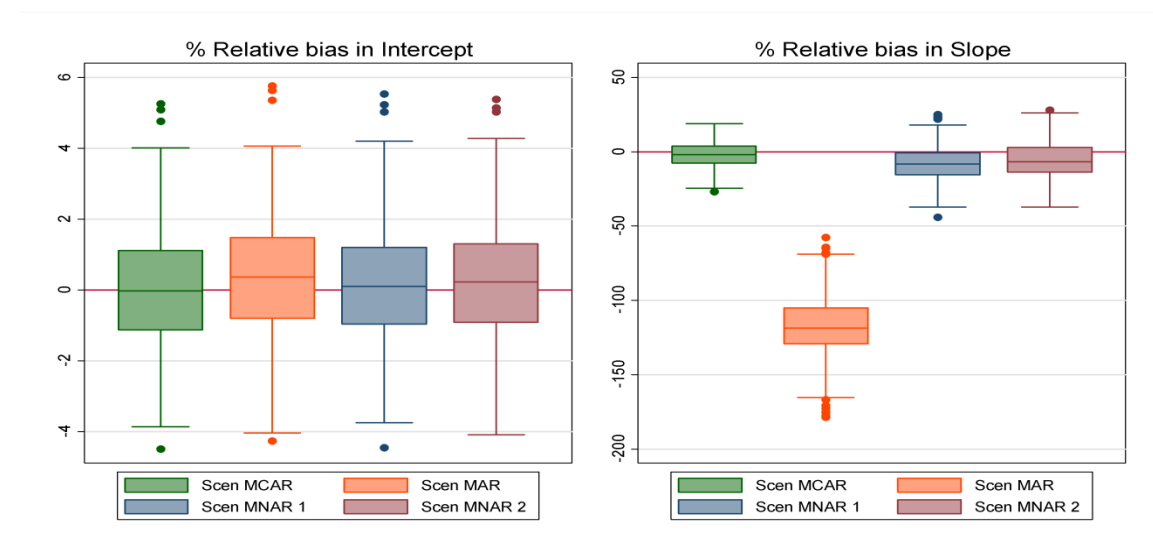
JMRE model							
Scenario	Parameter	TRUE	Mean Estim.	% Rel. Bias	SD(estim.)	Mean SE(estim.)	95% Cov. Prob
MCAR	intercept	24,236	24,241	0,02	0,384	0,387	96,00 (94,28-97,72)
	slope	-1,589	-1,62	-1,97	0,132	0,123	92,00 (89,61-94,39)
	intercept	24,236	24,322	0,35	0,396	0,399	94,40 (92,38-96,42)
MAR	slope	-1,589	-3,458	-117,68	0,317	0,172	0,00 (0,00-0,00)
	intercept	24,236	24,273	0,15	0,383	0,384	95,40 (93,56-97,24)
MNAR 1	slope	-1,589	-1,719	-8,2	0,178	0,135	76,20 (72,45-79,95)
	intercept	24,236	24,294	0,24	0,386	0,384	95,20 (93,32-97,08)
MNAR 2	slope	-1,589	-1,678	-5,63	0,193	0,136	76,40 (72,67-80,13)

6.6.2 Box plots για την μεροληψία της σταθεράς και της κλίσης των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.

Γράφημα 6.6.1: Box plot για την σχετική μεροληψία της σταθεράς και της κλίσης των μοντέλων μικτών επιδράσεων, για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.

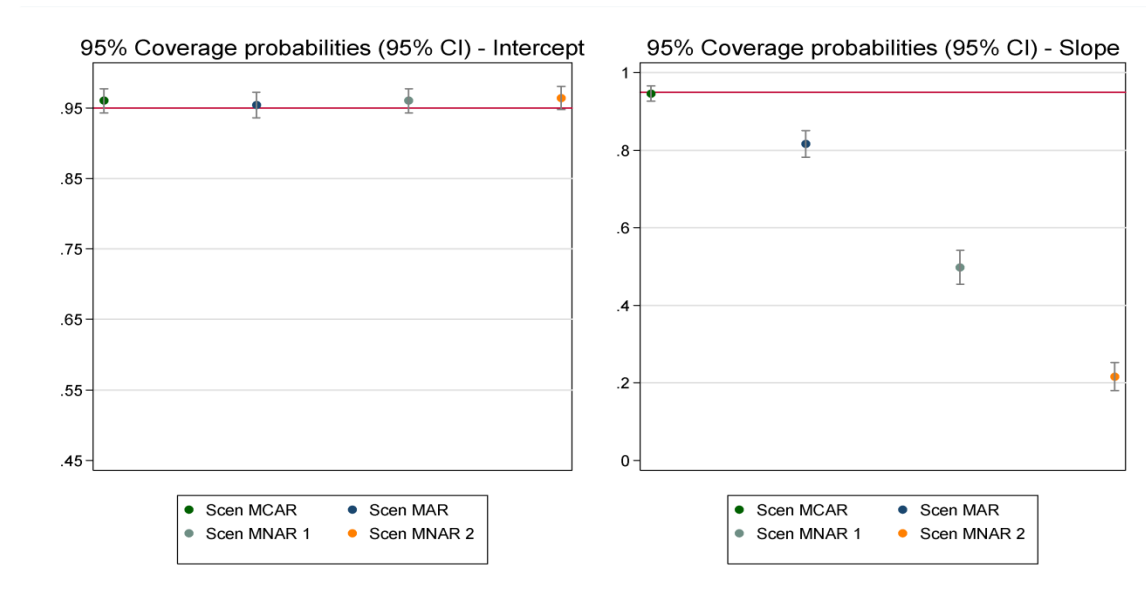


Γράφημα 6.6.2: Box plot για την σχετική μεροληψία της σταθεράς και της κλίσης των JMRE μοντέλων, για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.

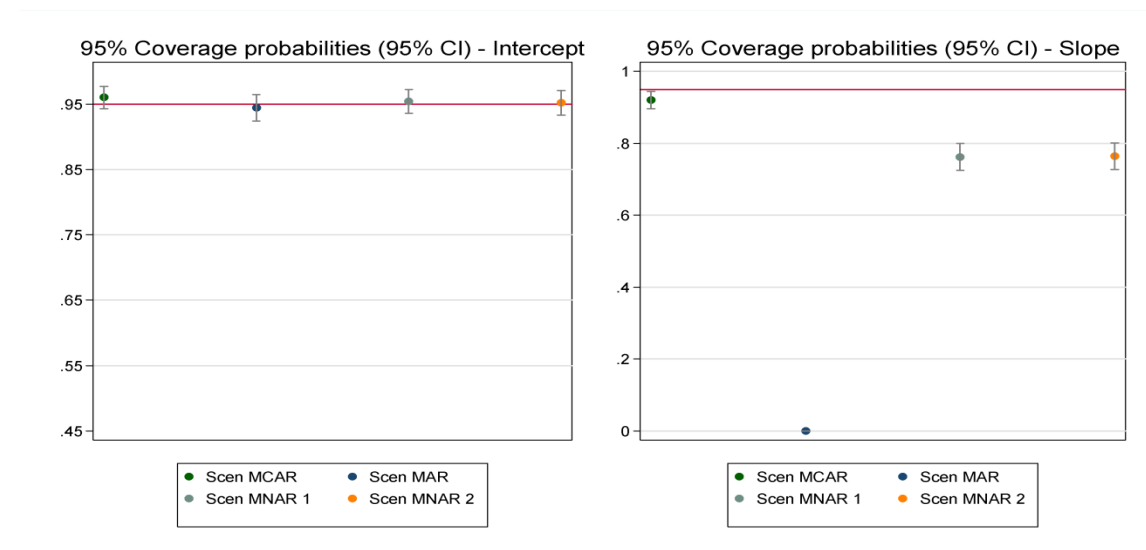


6.6.3 95% διάστημα εμπιστοσύνης για την σταθερά και την κλίση των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.

Γράφημα 6.6.3: 95% Διάστημα Εμπιστοσύνης για την σταθερά και την κλίση των μοντέλων μικτών επιδράσεων, για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.



Γράφημα 6.6.4: 95% Διάστημα Εμπιστοσύνης για την σταθερά και την κλίση των JMRE μοντέλων, για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.



Τα αποτελέσματα των κεφαλαίων 6.5 έως 6.6 συμβαδίζουν με τα αποτελέσματα των κεφαλαίων 6.1 έως 6.3. Παρατηρούμε ότι τόσο τα μοντέλα μικτών επιδράσεων όσο και τα JMRE μοντέλα δίνουν αμερόληπτες εκτιμήσεις για την σταθερά σε όλα τα διαφορετικά σενάρια θεραπείας και επισκέψεων. Όταν η έναρξη θεραπείας οδηγεί σε MCAR ή MAR ελλείπουσες τιμές τα μοντέλα μικτών επιδράσεων δίνουν αμερόληπτες εκτιμήσεις, όπως και πριν. Αλλά στην περίπτωση όπου η έναρξη θεραπείας οδηγεί σε MNAR ελλείπουσες τιμές τότε τα μοντέλα μικτών επιδράσεων υποεκτιμούν τον ρυθμό πτώσης των CD4 λεμφοκυττάρων ενώ τα JMRE μοντέλα δίνουν αμερόληπτες εκτιμήσεις. Ωστόσο, στην περίπτωση όπου η έναρξη θεραπείας οδηγεί σε MAR ελλείπουσες τιμές τότε τα JMRE μοντέλα υπερεκτιμούν τον ρυθμό πτώσης των CD4 λεμφοκυττάρων δίνοντας εκτιμήσεις για την κλίση οι οποίες έχουν μεροληψία 110%, όπως και στα προηγούμενα προσομοιωμένα δεδομένα. Και εδώ πραγματοποιήθηκε διερεύνηση των συσχετίσεων στα 500 datasets για όλα τα σενάρια επισκέψεων που αναφέρθηκαν και προκύπτει ότι η μέση συσχέτιση της τυχαίας σταθεράς (random intercept) με τα κατάλοιπα του χρόνου μέχρι την αποκοπή είναι 0,8 ενώ η μέση συσχέτιση της τυχαίας κλίσης (random slope) με τα κατάλοιπα του χρόνου μέχρι την αποκοπή είναι σχεδόν 0,85. Και οι δυο συσχετίσεις είναι πολύ υψηλές και για αυτό το λόγο το JMRE μοντέλο διορθώνει για πληροφοριακή αποκοπή γιατί βλέπει αυτά τα δεδομένα ως MNAR ενώ στην πραγματικότητα είναι MAR.

Καταλήγουμε στο ίδιο συμπέρασμα ακόμα και εάν υποθέσουμε ότι ο όρος frailty είναι θετικά συσχετισμένος με τα random effects και η συσχέτιση αυτή θεωρούμε ότι είναι της τάξης του 0,5. Τα αποτελέσματα των προσομοιώσεων δίνονται στα κεφάλαια 6.7 έως 6.9.

Επομένως, μπορούμε να πούμε ότι είτε ο όρος frailty είτε είναι ασυσχέτιστος με τις τυχαίες επιδράσεις είτε είναι αρνητικά/θετικά συσχετισμένος, τα αποτελέσματα δεν διαφέρουν στα διαφορετικά σενάρια περικοπής παρατηρήσεων. Τα μοντέλα μικτών επιδράσεων δίνουν αμερόληπτες εκτιμήσεις όταν ο μηχανισμός δημιουργίας ελλειπουσών τιμών είναι MCAR ή MAR ενώ τα JMRE μοντέλα δίνουν αμερόληπτες εκτιμήσεις σε περιπτώσεις όπου ο μηχανισμός δημιουργίας ελλειπουσών τιμών είναι MNAR. Για να αποφασίσουμε ποιο μοντέλο θα χρησιμοποιήσουμε σε κάθε περίπτωση

πρέπει να είμαστε απόλυτα σίγουροι για τον μηχανισμό δημιουργίας ελλειπυσών τιμών στα υπό μελέτη δεδομένα. Ο διαχωρισμός μεταξύ MAR και MNAR μηχανισμού δεν είναι μια εύκολη υπόθεση και για αυτό συνιστάται να εξετάζονται προσεκτικά οι επαναλαμβανόμενες μετρήσεις παρουσία ελλειπυσών τιμών και να πραγματοποιείται ανάλυση ευαισθησίας με σκοπό να αναδειχθεί το εύρος τιμών των παραμέτρων ενδιαφέροντος.

6.7 Σύγκριση μοντέλου μικτών επιδράσεων σε σχέση με το JMRE μοντέλο για τα 4 διαφορετικά σενάρια θεραπείας, με πιθανότητα προαιρετικής επίσκεψης 10%, και με θετική συσχέτιση μεταξύ frailty και random effects.

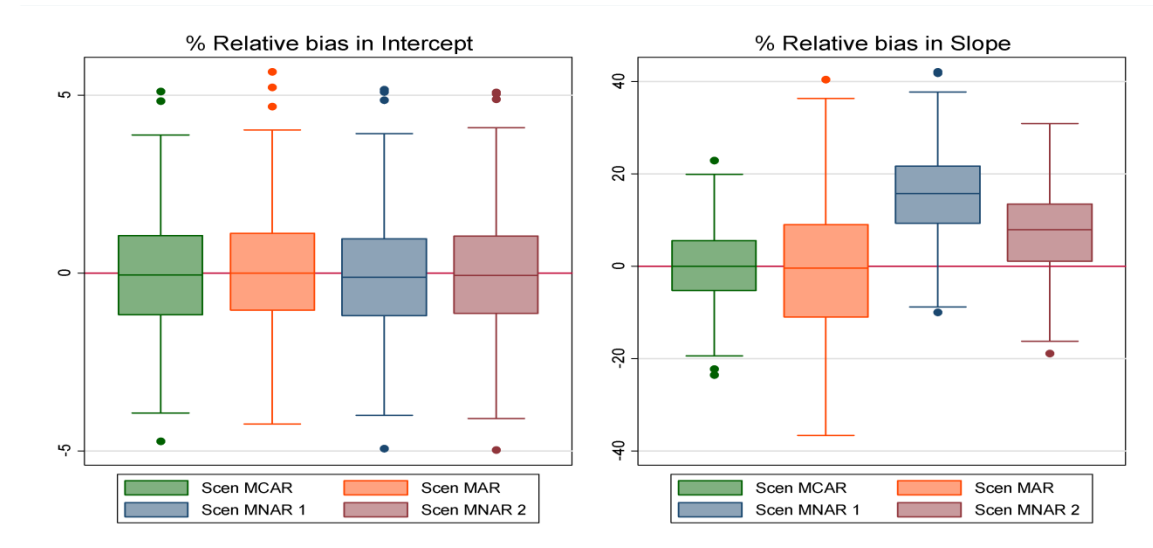
6.7.1 Αποτελέσματα προσομοιώσεων

Πίνακας 6.7: Αποτελέσματα προσομοιώσεων (500 επαναλήψεις με 200 άτομα/επανάληψη) για τα 4 διαφορετικά σενάρια θεραπείας. Μέγιστος χρόνος παρακολούθησης 5 χρόνια με 4 υποχρεωτικές επισκέψεις ανά χρόνο για κάθε ασθενή και η πιθανότητα πραγματοποίησης μιας προαιρετικής επίσκεψης εξαρτάται από τα παρατηρηθέντα CD4 στην προηγούμενη διαθέσιμη επίσκεψη και έναν όρο frailty. Οι αληθινές (True) παράμετροι προέρχονται από πραγματικά δεδομένα.

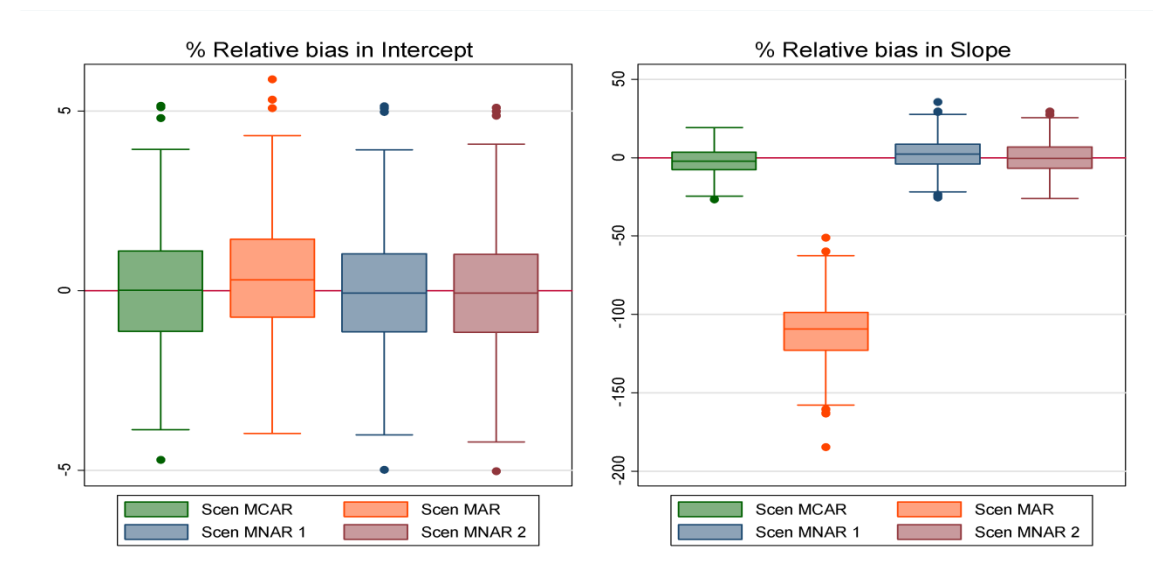
Mixed Model							
Scenario	Parameter	TRUE	Mean Estim.	% Rel. Bias	SD(estim.)	Mean SE(estim.)	95% Cov. Prob
MCAR	intercept	24,236	24,23	-0,03	0,388	0,39	95,80 (94,04-97,56)
	slope	-1,589	-1,587	0,12	0,129	0,124	94,00 (91,91-96,09)
MAR	intercept	24,236	24,239	0,01	0,395	0,397	95,20 (93,32-97,08)
	slope	-1,589	-1,603	-0,9	0,224	0,165	84,20 (80,99-87,41)
MNAR 1	intercept	24,236	24,211	-0,11	0,386	0,389	95,20 (93,32-97,08)
	slope	-1,589	-1,341	15,59	0,141	0,137	53,80 (49,42-58,18)
MNAR 2	intercept	24,236	24,227	-0,04	0,385	0,389	95,40 (93,56-97,24)
	slope	-1,589	-1,469	7,51	0,141	0,135	84,00 (80,78-87,22)
JMRE model							
Scenario	Parameter	TRUE	Mean Estim.	% Rel. Bias	SD(estim.)	Mean SE(estim.)	95% Cov. Prob
MCAR	intercept	24,236	24,24	0,02	0,388	0,389	95,80 (94,04-97,56)
	slope	-1,589	-1,621	-2,06	0,133	0,124	91,80 (89,39-94,21)
MAR	intercept	24,236	24,318	0,34	0,398	0,402	94,20 (92,14- 96,26)
	slope	-1,589	-3,343	-110,41	0,302	0,172	0,00 (0,00-0,00)
MNAR 1	intercept	24,236	24,225	-0,05	0,39	0,388	95,40 (93,56-97,24)
	slope	-1,589	-1,548	2,55	0,16	0,135	89,40 (86,69-92,11)
MNAR 2	intercept	24,236	24,225	-0,05	0,388	0,388	95,00 (93,08-96,92)
	slope	-1,589	-1,591	-0,15	0,153	0,134	91,80 (89,39-94,21)

6.7.2 Box plots για την μεροληψία της σταθεράς και της κλίσης των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.

Γράφημα 6.7.1: Box plot για την σχετική μεροληψία της σταθεράς και της κλίσης των μοντέλων μικτών επιδράσεων, για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.

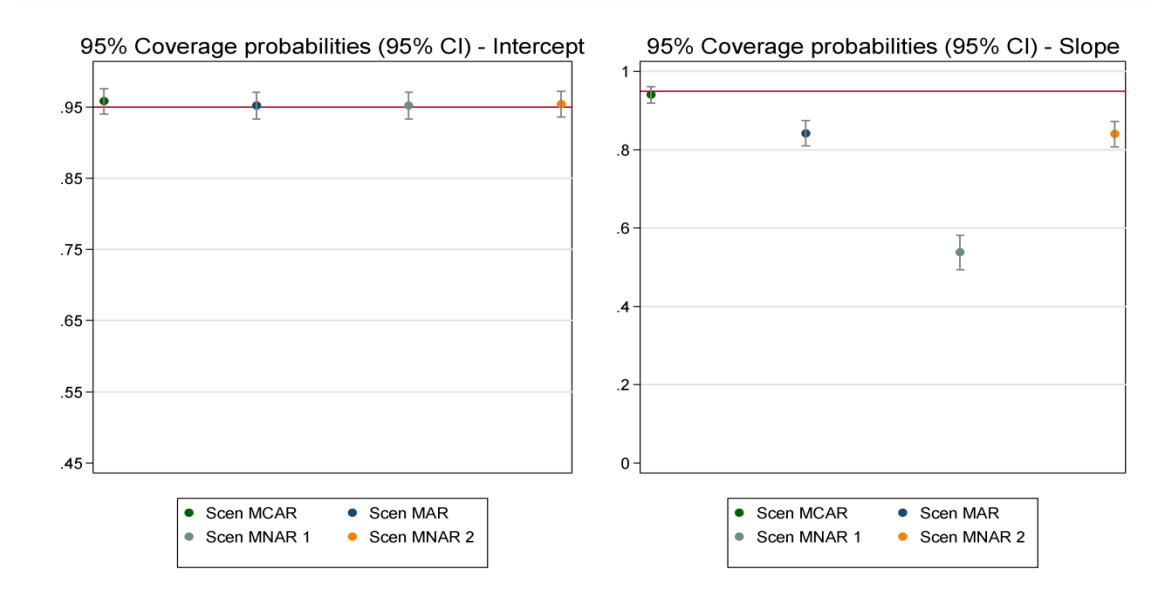


Γράφημα 6.7.2: Box plot για την σχετική μεροληψία της σταθεράς και της κλίσης των JMRE μοντέλων, για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.

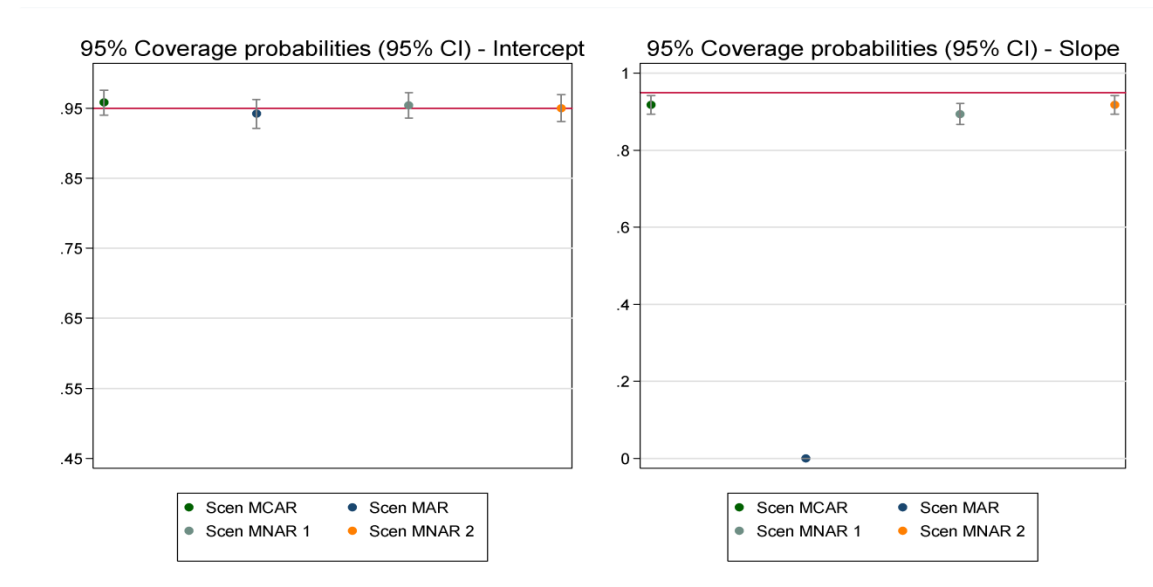


6.7.3 95% διάστημα εμπιστοσύνης για την σταθερά και την κλίση των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.

Γράφημα 6.7.3: 95% Διάστημα Εμπιστοσύνης για την σταθερά και την κλίση των μοντέλων μικτών επιδράσεων, για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.



Γράφημα 6.7.4: 95% Διάστημα Εμπιστοσύνης για την σταθερά και την κλίση των JMRE μοντέλων, για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.



6.8 Σύγκριση μοντέλου μικτών επιδράσεων σε σχέση με το JMRE μοντέλο για τα 4 διαφορετικά σενάρια θεραπείας, με πιθανότητα προαιρετικής επίσκεψης που εξαρτάται από τα παρατηρηθέντα CD4 στην προηγούμενη επίσκεψη και με θετική συσχέτιση μεταξύ frailty και random effects.

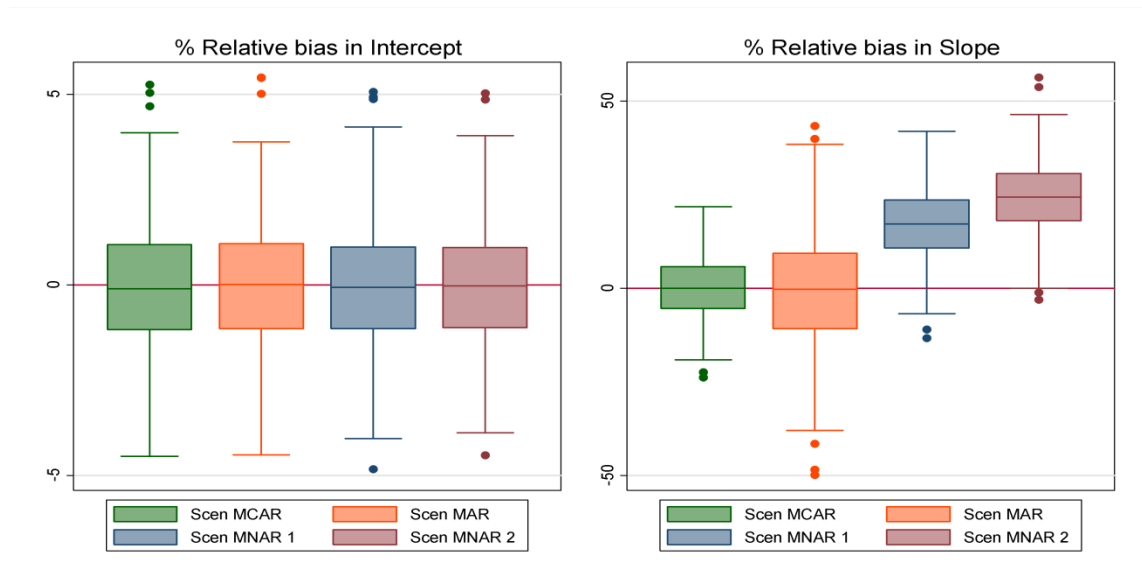
6.8.1 Αποτελέσματα προσομοιώσεων

Πίνακας 6.8: Αποτελέσματα προσομοιώσεων (500 επαναλήψεις με 200 άτομα/επανάληψη) για τα 4 διαφορετικά σενάρια θεραπείας. Μέγιστος χρόνος παρακολούθησης 5 χρόνια με 4 υποχρεωτικές επισκέψεις ανά χρόνο για κάθε ασθενή και η πιθανότητα πραγματοποίησης μιας προαιρετικής επίσκεψης εξαρτάται από τα παρατηρηθέντα CD4 στην προηγούμενη διαθέσιμη επίσκεψη και έναν όρο frailty. Οι αληθινές (True) παράμετροι προέρχονται από πραγματικά δεδομένα.

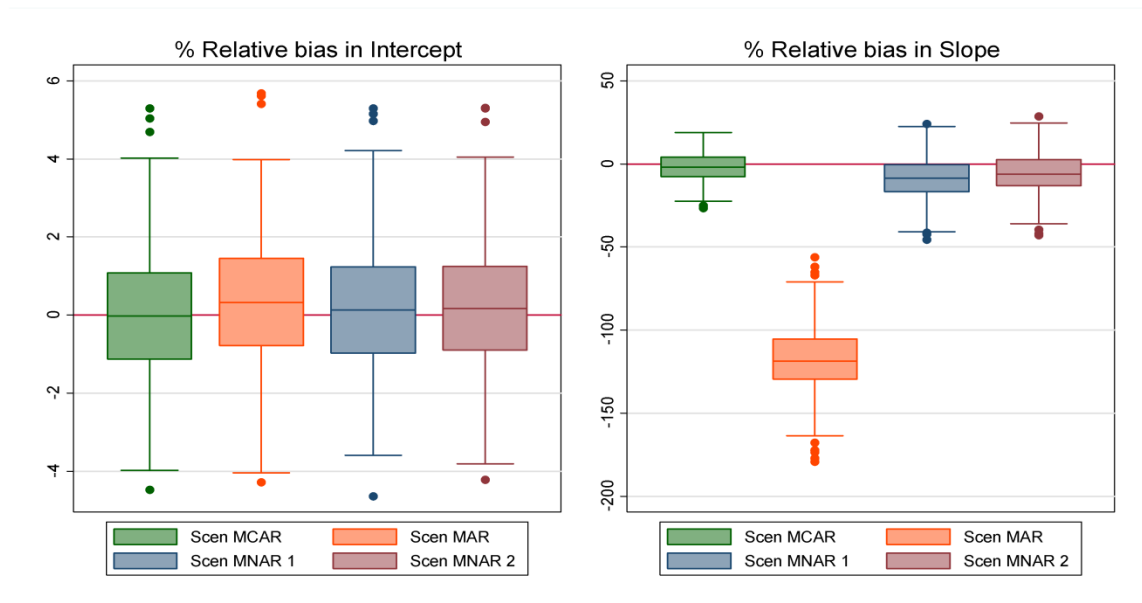
Mixed Model							
Scenario	Parameter	TRUE	Mean Estim.	% Rel. Bias	SD(estim.)	Mean SE(estim.)	95% Cov. Prob
MCAR	intercept	24,236	24,231	-0,02	0,383	0,388	96,20 (94,52-97,88)
	slope	-1,589	-1,587	0,09	0,127	0,123	94,40 (92,38-96,42)
MAR	intercept	24,236	24,238	0,01	0,393	0,397	95,40 (93,56-97,24)
	slope	-1,589	-1,602	-0,81	0,241	0,167	82,80 (79,48-86,12)
MNAR 1	intercept	24,236	24,228	-0,03	0,378	0,386	96,40 (94,76-98,04)
	slope	-1,589	-1,317	17,1	0,146	0,138	49,80 (45,40-54,20)
MNAR 2	intercept	24,236	24,243	0,03	0,382	0,386	96,60 (95,01-98,19)
	slope	-1,589	-1,464	7,86	0,144	0,136	84,20 (80,99-87,41)
JMRE model							
Scenario	Parameter	TRUE	Mean Estim.	% Rel. Bias	SD(estim.)	Mean SE(estim.)	95% Cov. Prob
MCAR	intercept	24,236	24,24	0,02	0,383	0,387	96,20 (94,52-97,88)
	slope	-1,589	-1,619	-1,91	0,132	0,123	92,00 (89,61-94,39)
MAR	intercept	24,236	24,32	0,35	0,396	0,399	94,40 (92,38-96,42)
	slope	-1,589	-3,464	-118,05	0,317	0,173	0,00 (0,00-0,00)
MNAR 1	intercept	24,236	24,281	0,19	0,38	0,384	95,40 (93,56-97,24)
	slope	-1,589	-1,723	-8,48	0,188	0,136	74,40 (70,56-78,24)
MNAR 2	intercept	24,236	24,269	0,13	0,386	0,384	95,60 (93,80-97,40)
	slope	-1,589	-1,735	-9,21	0,166	0,134	76,40 (72,67-80,13)

6.8.2 Box plots για την μεροληψία της σταθεράς και της κλίσης των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.

Γράφημα 6.8.1: Box plot για την σχετική μεροληψία της σταθεράς και της κλίσης των μοντέλων μικτών επιδράσεων, για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.

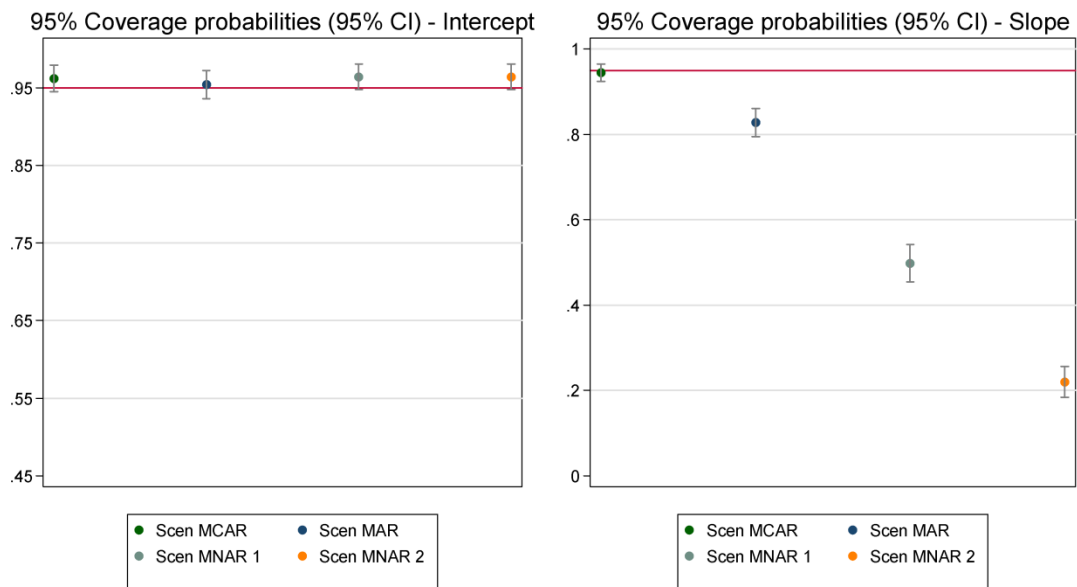


Γράφημα 6.8.2: Box plot για την σχετική μεροληψία της σταθεράς και της κλίσης των JMRE μοντέλων, για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.

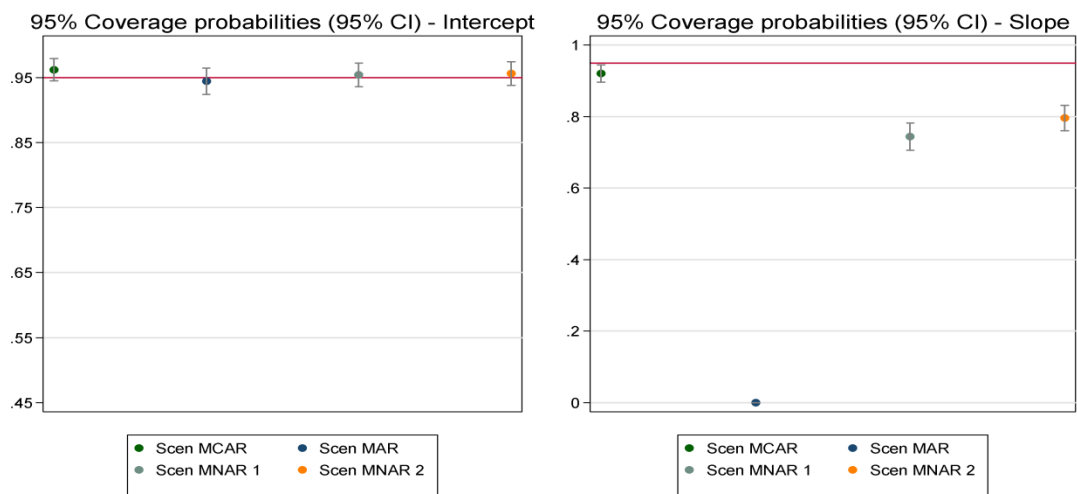


6.8.3 95% διάστημα εμπιστοσύνης για την σταθερά και την κλίση των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.

Γράφημα 6.8.3: 95% Διάστημα Εμπιστοσύνης για την σταθερά και την κλίση των μοντέλων μικτών επιδράσεων, για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.



Γράφημα 6.8.4: 95% Διάστημα Εμπιστοσύνης για την σταθερά και την κλίση των JMRE μοντέλων, για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.



6.9 Σύγκριση μοντέλου μικτών επιδράσεων σε σχέση με το JMRE μοντέλο για τα 4 διαφορετικά σενάρια θεραπείας, με πιθανότητα προαιρετικής επίσκεψης που εξαρτάται από τα παρατηρηθέντα CD4 στην προηγούμενη επίσκεψη και από έναν όρο frailty, και με αρνητική συσχέτιση μεταξύ frailty και random effects.

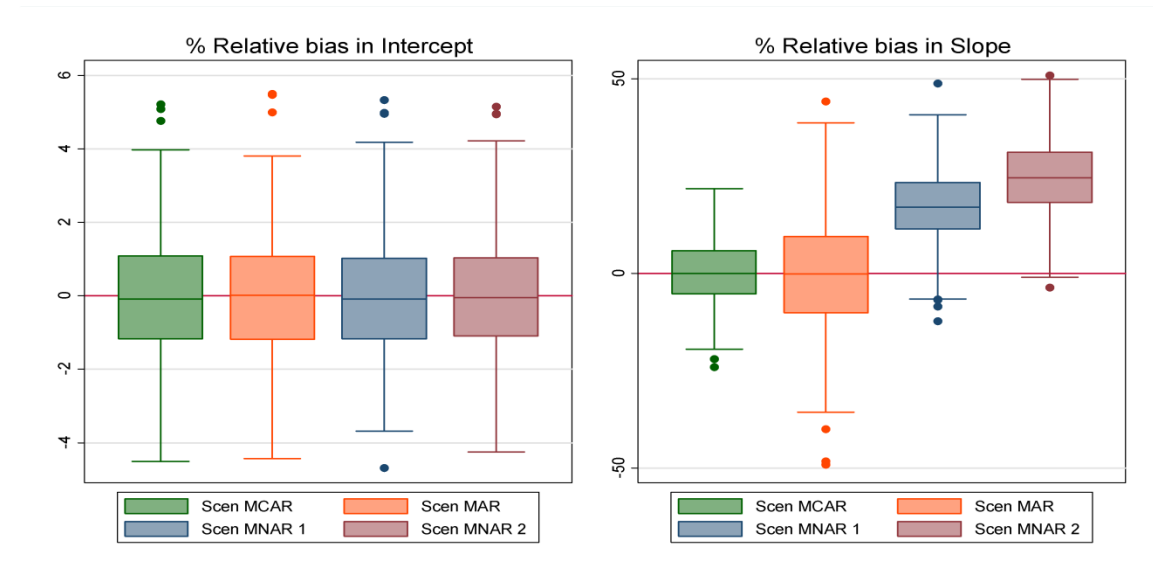
6.9.1 Αποτελέσματα προσομοιώσεων

Πίνακας 6.9: Αποτελέσματα προσομοιώσεων (500 επαναλήψεις με 200 άτομα/επανάληψη) για τα 4 διαφορετικά σενάρια θεραπείας. Μέγιστος χρόνος παρακολούθησης 5 χρόνια με 4 υποχρεωτικές επισκέψεις ανά χρόνο για κάθε ασθενή και η πιθανότητα πραγματοποίησης μιας προαιρετικής επίσκεψης εξαρτάται από τα παρατηρηθέντα CD4 στην προηγούμενη διαθέσιμη επίσκεψη και έναν όρο frailty. Οι αληθινές (True) παράμετροι προέρχονται από πραγματικά δεδομένα.

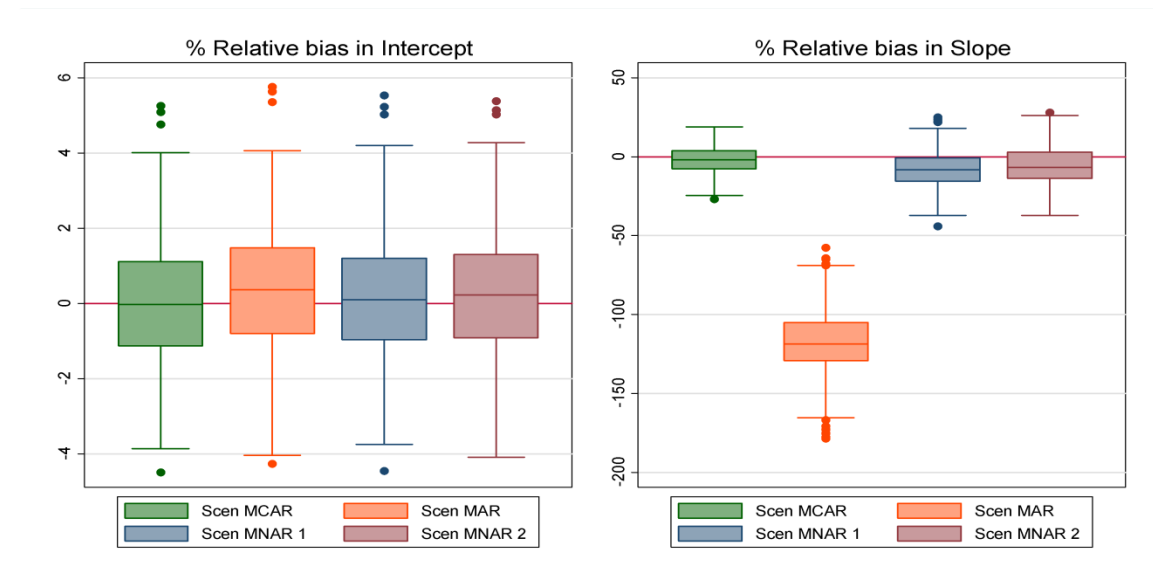
Mixed Model							
Scenario	Parameter	TRUE	Mean Estim.	% Rel. Bias	SD(estim.)	Mean SE(estim.)	95% Cov. Prob
MCAR	intercept	24,236	24,231	-0,02	0,384	0,388	96,00 (94,28-97,72)
	slope	-1,589	-1,588	0,05	0,127	0,123	94,60 (92,61-96,59)
MAR	intercept	24,236	24,238	0,01	0,394	0,397	95,40 (93,56-97,24)
	slope	-1,589	-1,593	-0,24	0,241	0,167	81,60 (78,19- 85,01)
MNAR 1	intercept	24,236	24,226	-0,04	0,383	0,386	96,00 (94,28-97,72)
	slope	-1,589	-1,315	17,25	0,147	0,138	49,80 (45,40-54,20)
MNAR 2	intercept	24,236	24,235	-0,01	0,38	0,386	96,40 (94,76-98,04)
	slope	-1,589	-1,201	24,42	0,149	0,139	21,60 (17,98-25,22)
JMRE model							
Scenario	Parameter	TRUE	Mean Estim.	% Rel. Bias	SD(estim.)	Mean SE(estim.)	95% Cov. Prob
MCAR	intercept	24,236	24,241	0,02	0,384	0,387	96,00 (94,28-97,72)
	slope	-1,589	-1,62	-1,97	0,132	0,123	92,00 (89,61-94,39)
MAR	intercept	24,236	24,322	0,35	0,396	0,399	94,40 (92,38-96,42)
	slope	-1,589	-3,458	-117,68	0,317	0,172	0,00 (0,00-0,00)
MNAR 1	intercept	24,236	24,273	0,15	0,383	0,384	95,40 (93,56-97,24)
	slope	-1,589	-1,719	-8,2	0,178	0,135	76,20 (72,45-79,95)
MNAR 2	intercept	24,236	24,294	0,24	0,386	0,384	95,20 (93,32-97,08)
	slope	-1,589	-1,678	-5,63	0,193	0,136	76,0 (72,67- 80,13)

6.9.2 Box plots για την μεροληψία της σταθεράς και της κλίσης των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.

Γράφημα 6.9.1: Box plot για την σχετική μεροληψία της σταθεράς και της κλίσης των μοντέλων μικτών επιδράσεων, για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.

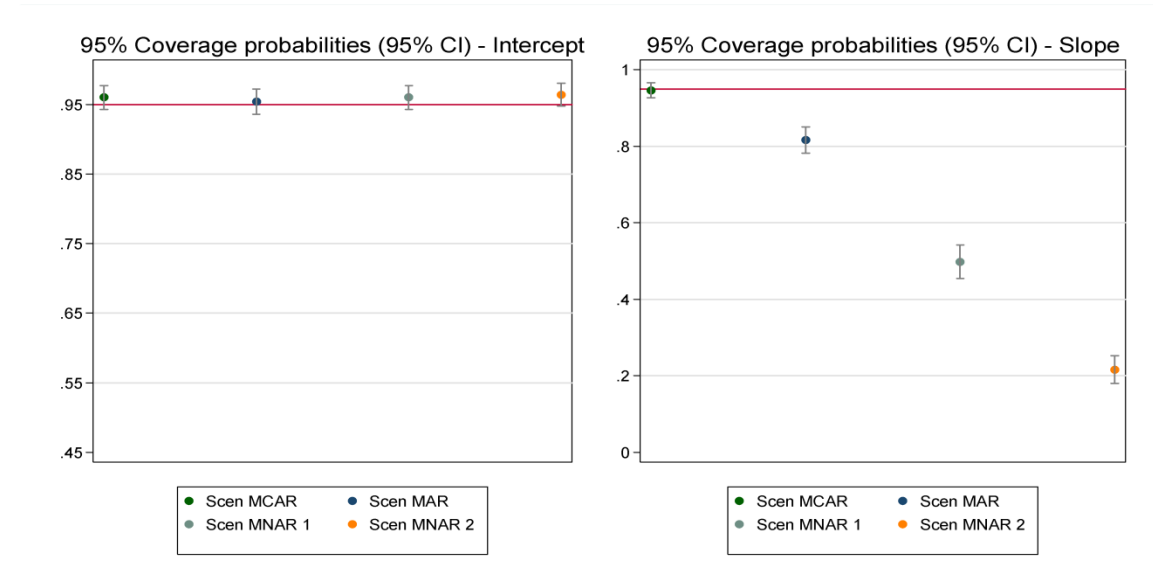


Γράφημα 6.9.2: Box plot για την σχετική μεροληψία της σταθεράς και της κλίσης των JMRE μοντέλων, για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.

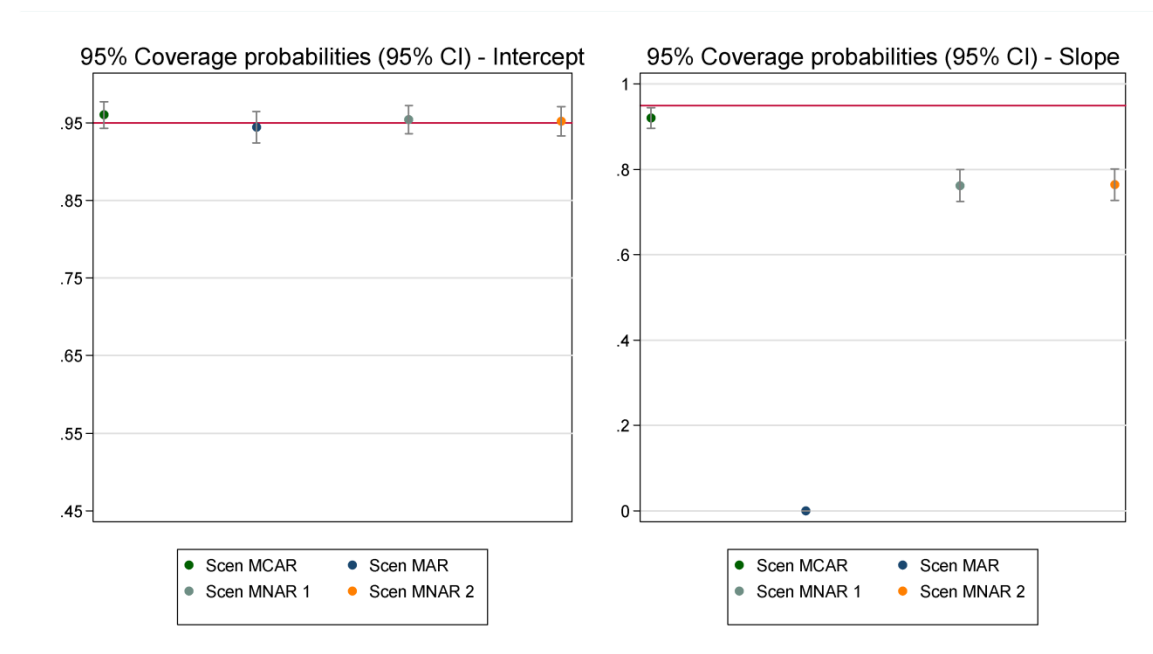


6.9.3 95% διάστημα εμπιστοσύνης για την σταθερά και την κλίση των μοντέλων μικτών επιδράσεων και των JMRE μοντέλων για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.

Γράφημα 6.9.3: 95% Διάστημα Εμπιστοσύνης για την σταθερά και την κλίση των μοντέλων μικτών επιδράσεων, για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.



Γράφημα 6.9.4: 95% Διάστημα Εμπιστοσύνης για την σταθερά και την κλίση των JMRE μοντέλων, για τα τέσσερα διαφορετικά σενάρια περικοπής παρατηρήσεων.



Συμπεράσματα – Συζήτηση

Οι διαχρονικές μελέτες χρησιμοποιούνται ευρέως στο χώρο της ιατρικής έρευνας. Τέτοιου είδους δεδομένα συλλέγονται σε επιδημιολογικές μελέτες παρατήρησης καθώς και στις κλινικές δοκιμές. Το κυρίως πρόβλημα από στατιστικής πλευράς που προκύπτει σε αυτές τις σειρές μετρήσεων είναι η συχνή εμφάνιση ελλειπουσών τιμών. Στο πλαίσιο αυτής της διπλωματικής διερευνήθηκε η απόδοση μοντέλων μικτών επιδράσεων και JMRE μοντέλων σε πραγματικά δεδομένα και κάτω από διαφορετικούς μηχανισμούς δημιουργίας ελλειπουσών τιμών μέσω προσομοιωμένων δεδομένων.

Τα δεδομένα που είναι διαθέσιμα από τις διαχρονικές μελέτες είναι αρκετά πολύπλοκα και σε πρώτο στάδιο είναι απαραίτητο να πραγματοποιείται περιγραφική ανάλυση για να κατανοήσουμε τις ατομικές καθώς και τις πληθυσμιακές τάσεις. Μια τέτοια ανάλυση περιλαμβάνει συνήθως γραφήματα της εξέλιξης πληθυσμιακών μέσων ή διάμεσων τιμών καθώς και γραφήματα ατομικών πορειών τα οποία αναδεικνύουν την διαφορετικότητα των τιμών του δείκτη για κάθε άτομο. Ωστόσο, τέτοιου είδους γραφήματα πρέπει να χρησιμοποιούνται με ιδιαίτερη προσοχή διότι συχνά οδηγούν σε λανθασμένα συμπεράσματα παρουσία ελλειπουσών τιμών.

Αφορμή για την παρούσα εργασία αποτέλεσε η ανάλυση διαχρονικών δεδομένων CD4 λεμφοκυττάρων από HIV οροθετικά άτομα που συμμετέχουν στην μελέτη AMACS. Οι εκτιμήσεις των σταθερών παραμέτρων σύμφωνα με το μοντέλο μικτών επιδράσεων διαφέρουν σημαντικά από τις αντίστοιχες εκτιμήσεις που δίνει το JMRE μοντέλο. Για να διερευνηθεί η συμπεριφορά τέτοιων μοντέλων υπό διαφορετικούς μηχανισμούς δημιουργίας ελλειπουσών τιμών, δημιουργήσαμε προσομοιωμένα δεδομένα και σε αυτά εφαρμόστηκαν μοντέλα τυχαίων επιδράσεων και JMRE μοντέλα.

Υπό την προϋπόθεση εντελώς τυχαίου μηχανισμού δημιουργίας ελλειπουσών τιμών (MCAR) και τα δυο μοντέλα έδωσαν αμερόληπτες εκτιμήσεις. Στην περίπτωση των τυχαίων ελλειπουσών τιμών (MAR) το μοντέλο μικτών επιδράσεων έδωσε αμερόληπτες εκτιμήσεις ενώ το JMRE μοντέλο έδωσε ισχυρά μεροληπτικές εκτιμήσεις, σε όλα τα σενάρια επισκέψεων. Το JMRE μοντέλο σε MAR ελλείπουσες τιμές υπερεκτιμούσε τον ρυθμό μεταβολής των CD4 λεμφοκυττάρων, δίνονται εκτιμήσεις σχεδόν διπλάσιες από

τις πραγματικές. Αυτά τα αποτελέσματα οφείλονται στο γεγονός ότι οι τυχαίες επιδράσεις σχετίζονται ισχυρά με τα κατάλοιπα του χρόνου επιβίωσης και το JMRE μοντέλο θεωρεί τον μηχανισμό MNAR για αυτό διορθώνει για πληροφοριακή αποκοπή ενώ στην πραγματικότητα κάτι τέτοιο δεν υπάρχει.

Στην περίπτωση MNAR ελλειπουσών τιμών το JMRE μοντέλο δίνει αμερόληπτες εκτιμήσεις για όλα τα σενάρια επισκέψεων ενώ το μοντέλο μικτών επιδράσεων δίνει συστηματικά μεροληπτικές εκτιμήσεις. Το μοντέλο μικτών επιδράσεων αγνοεί την ύπαρξη πληροφοριακής αποκοπής με αποτέλεσμα να υποεκτιμάει τον ρυθμό μεταβολής των CD4 λεμφοκυττάρων.

Συμπερασματικά, σε περιπτώσεις MCAR ελλειπουσών τιμών οποιαδήποτε μέθοδος και αν χρησιμοποιηθεί θα έχουμε αμερόληπτες εκτιμήσεις. Ενώ υπό MNAR μηχανισμό δημιουργίας ελλειπουσών τιμών πρέπει να χρησιμοποιούνται μοντέλα που λαμβάνουν υπόψη τον μηχανισμό δημιουργίας ελλειπουσών τιμών, όπως είναι το JMRE μοντέλο, ενώ το μοντέλο μικτών επιδράσεων πρέπει να αποφεύγεται γιατί οδηγεί σε μεροληπτικές εκτιμήσεις.

Όσο αφορά τις MAR ελλείπουσες τιμές από την παραπάνω ανάλυση είδαμε ότι τα μοντέλα μικτών επιδράσεων δίνουν αμερόληπτες εκτιμήσεις ενώ τα JMRE μοντέλα δίνουν μεροληπτικές εκτιμήσεις. Ωστόσο, είναι πολύ δύσκολο να γίνει διάκριση μεταξύ MAR και MNAR μηχανισμού ελλειπουσών τιμών. Αν και μόνο αν, είμαστε σίγουροι ότι ο μηχανισμός δημιουργίας ελλειπουσών τιμών είναι MAR τότε πρέπει να χρησιμοποιούνται τα μοντέλα μικτών επιδράσεων. Συμπερασματικά, καταλήγουμε ότι για την ανάλυση επαναλαμβανόμενων μετρήσεων παρουσία ελλειπουσών τιμών πρέπει να εξεταστούν προσεκτικά τα δεδομένα και να εφαρμοστούν τα κατάλληλα μοντέλα.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Andridge, R.R. & Little, R.J.A. (2010). A review for hot deck imputation for survey non-response. *International Statistical Review* 78, 1, 40–64.

Aledort, L. M., M. W. Hilgartner, M. C. Pike, G. F. Gjerset, M. A. Koerper, E. Y. C. Lian, J. M. Lusher, and J. W. Mosley (1992). Variability in serial CD4 counts and relation to progression of HIV-1 infection to AIDS in hemophilic patients. *Br. Med. J.* 304 (6821), 212-216.

Baker, S. G. and N. M. Laird (1988). Regression-analysis for categorical variables with outcome subject to nonignorable nonresponse. *J. Am. Stat. Assoc.* 83 (401), 62-69.

Biggar, R. J. (1990). AIDS incubation in 1891 HIV seroconverters from different exposure groups. *AIDS* 4 (11), 1059-1066.

Breslow, N. E. (1984). Extra-poisson variation in log-linear models. *Appl.Stat.-J. R. Stat. Soc.* 33 (1), 38-44.

Buck, S.F (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of Royal Statistical Society, Series B*, 22, 302-306.

Chi, E. M. and G. C. Reinsel (1989). Models for longitudinal data with random effects and ar(1) errors. *J. Am. Stat. Assoc.* 84 (406), 452-459.

Choi, S. S., S. W. Lagakos, R. T. Schooley, and P. A. Volberding (1993). CD4+ lymphocytes are an incomplete surrogate marker for clinical progression in persons with asymptomatic HIV-infection taking zidovudine. *Ann. Intern.Med.* 118 (9), 674-680.

Crowther M.J, K. R. Abrams (2012), Joint modeling of longitudinal and survival data. *Stata Journal*, 1-20.

Collet D. (1994), *Modeling Survival Data in Medical Research*. Chapman & Hall.

Cox, D. R. and D. Oakes (1984). *Analysis of Survival Data*. Chapman and Hall.

Curran, D., M. Bacchi, S. F. H. Schmitz, G. Molenberghs, and R. J. Sylvester (1998). Identifying the types of missingness in quality of life data from clinical trials. *Stat. Med.* 17 (5-7), 739-756.

Dawson, J. D. (1994). Stratification of summary statistic tests according to missing data patterns. *Stat. Med.* 13 (18), 1853-1863.

Dawson, J. D. and S. W. Lagakos (1991). Analyzing laboratory marker changes in AIDS clinical-trials. *J. Acquir. Immune Defic. Syndr. Hum. Retrovirol.* 4 (7), 667-676.

Degruttola, V. and X. M. Tu (1994). Modeling progression of CD4-lymphocyte count and its relationship to survival-time. *Biometrics* 50, 1003-1014.

Demirtas, H. & Schafer, J.L. (2003). On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out. *Statistics in Medicine*, 22, 2553-2575.

Detels, R., P. A. English, J. V. Giorgi, B. R. Visscher, J. L. Fahey, J. M. G. Taylor, J. P. Dudley, P. Nishanian, A. Munoz, J. P. Phair, B. F. Polk, and C. R. Rinaldo (1988). Patterns of CD4+ cell changes after HIV-1 infection indicate the existence of a codeterminant of AIDS. *J. Acquir. Immune Defic. Syndr. Hum. Retrovirol.* 1 (4), 390-395.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* 39 (1), 1-38. With discussion.

Diem, J. E. and J. R. Liukkonen (1988). A comparative-study of 3 methods for analyzing longitudinal pulmonary-function data. *Stat. Med.* 7 (1-2), 19-28.

Diggle, P. J., K. Y. Liang, and S. L. Zeger (1994). *Analysis of longitudinal data*. Oxford University Press.

Diggle, P. and M. G. Kenward (1994). Informative drop-out in longitudinal data-analysis. *J Roy Stat Soc C-App* 43, 49-93.

Diggle, P. J. (1988). An approach to the analysis of repeated measurements. *Biometrics* 44 (4), 959-971.

Ellenberg, S.S and J. M. Hamilton (1989). Surrogate endpoints in clinical trials-cancer. *Stat. Med.* 8(4), 405-413.

Eyster, M. E., M. H. Gail, J. O. Ballard, H. Almondhry, and J. J. Goedert (1987). Natural-history of human immunodeficiency virus-infections in hemophiliacs—effects of T-cell subsets, platelet counts, and age. *Ann. Intern.Med.* 107 (1), 1-6.

Fahey, J. L., J. M. G. Taylor, R. Detels, B. Hofmann, R. Melmed, P. Nishanian, and J. V. Giorgi (1990). The prognostic value of cellular and serologic markers in infection with human immunodeficiency virus type-1. *N. Engl. J.Med.* 322 (3), 166-172.

Faucett, C. L. and D. C. Thomas (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: A Gibbs sampling approach. *Stat Med* 15, 1663-1685.

Fitzmaurice, G. M., N. M. Laird, and A. G. Rotnitzky (1993). Regression models for discrete longitudinal responses. *Stat. Sci.* 8 (3), 284-299.

Follmann, D. and M.Wu (1995). An approximate generalized linear-model with random effects for informative missing data. *Biometrics* 51, 151-168.

Frison, L. and S. J. Pocock (1992). Repeated measures in clinical-trials - analysis using mean summary statistics and its implications for design. *Stat. Med.* 11 (13), 1685-1704.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1995). *Bayesian data analysis*. Texts in Statistical Science Series. London: Chapman & Hall.

Gilks, W. R., C. C. Wang, B. Yvonnet, and P. Coursaget (1993). Random- effects models for longitudinal data using Gibbs sampling. *Biometrics* 49 (2), 441-453.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika* 73 (1), 43-56.

Goldstein, H. (1989). Restricted unbiased iterative generalized least-squares estimation. *Biometrika* 76 (3), 622-623.

Harville, D. A. and R. W. Mee (1984). A mixed-model procedure for analyzing ordered categorical-data. *Biometrics* 40 (2), 393-408.

Henderson, R., P. Diggle, and A. Dobson (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics* 1, 465-480.

Henrard, D. R., E. Daar, H. Farzadegan, S. J. Clark, J. Phillips, G. M. Shaw, and M. P. Busch (1995). Virological and immunological characterization of symptomatic and asymptomatic primary HIV-1 infection. *J. Acquir. Immune Defic. Syndr. Hum. Retrovirol.* 9 (3), 305-310.

Heyting, A., J. T. B. M. Tolboom, and J. G. A. Essers (1992). Statistical handling of drop-outs in longitudinal clinical- trials. *Stat. Med.* 11 (16), 2043-2061.

Hillis, A. and D. Seigel (1989). Surrogate endpoints in clinical-trials - ophthalmologic disorders. *Stat. Med.* 8 (4), 427-430.

Hogan, J. W. and N. M. Laird (1997a). Mixture models for the joint distribution of repeated measures and event times. *Stat Med* 16, 239-257.

Hogan, J. W. and N. M. Laird (1997b). Model-based approaches to analysing incomplete longitudinal and failure time data. *Stat. Med.* 16 (1-3), 259-272.

Hogan, J. W., X. H. Lin, and B. Herman (2004). Mixtures of varying coefficient models for longitudinal data with discrete or continuous nonignorable dropout. *Biometrics* 60 (4), 854-864.

Hughes, M. D., D. S. Stein, H. M. Gundacker, F. T. Valentine, J. P. Phair, and P. A. Volberding (1994). Within-subject variation in CD4 lymphocyte count in asymptomatic human-immunodeficiency-virus infection - implications for patient monitoring. *J. Infect. Dis.* 169 (1), 28-36.

Jewell, N. and J. Kalbfleisch (1992). Marker models in survival analysis and applications to issues associated with AIDS. In *AIDS epidemiology: Methodological issues*. Birkhauser, Boston.

Johnson, N. L. and S. Kotz (1970). *Continuous Univariate Distributions-1*. Wiley, New York.

Kenward, M. G., E. Lesaffre, and G. Molenberghs (1994). An application of maximum-likelihood and generalized estimating equations to the analysis of ordinal data from a longitudinal- study with cases missing at random. *Biometrics* 50 (4), 945-953.

Kalbfleisch J. D. & Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data* Wiley, 2nd Edition.

Laird, N. M. and J. H. Ware (1982). Random-effects models for longitudinal data. *Biometrics* 38 (4), 963-974.

Laird, N. M. (1988). Missing data in longitudinal-studies. *Stat Med* 7, 305-315.

Lang, W., H. Perkins, R. E. Anderson, R. Royce, N. Jewell, and W. J. Winkelstein (1989). Patterns of T lymphocyte changes with human immunodeficiency virus infection: From seroconversion to the development of AIDS. *JAIDS* 2, 63-69.

Lange, K. L., R. J. A. Little, and J. M. G. Taylor (1989). Robust statistical modeling using the t-distribution. *J. Am. Stat. Assoc.* 84 (408), 881-896.

Lange, N., B. P. Carlin, and A. E. Gelfand (1992). Hierarchical bayes models for the progression of HIV-infection using longitudinal CD4 T-cell numbers. *J. Am. Stat. Assoc.* 87 (419), 615-626.

Liang, K. Y. and S. L. Zeger (1986). Longitudinal data-analysis using generalized linear-models. *Biometrika* 73 (1), 13-22.

Lindstrom, M. J. and D. M. Bates (1988). Newton-raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *J. Am. Stat. Assoc.* 83 (404), 1014-1022.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *J Am Stat Assoc* 90, 1112-1121.

Little, R. J. A. and D. B. Rubin (1987). *Statistical analysis with missing data*. Wiley, New York.

Little, R. J. A. and D. B. Rubin (2002). *Statistical analysis with missing data (2nd edn)*. Wiley, New York.

Longford, N. T. (1993). *Random coefficient models*, Volume 11 of *Oxford Statistical Science Series*. New York: The Clarendon Press Oxford University Press. , Oxford Science Publications.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *J R Statistic Soc B Met* 44, 226-233.

Lyles, R. H., A. Munoz, T. E. Yamashita, H. Bazmi, R. Detels, C. R. Rinaldo, J. B. Margolick, J. P. Phair, and J. W. Mellors (2000). Natural history of human immunodeficiency virus type 1 viremia after seroconversion and proximal to AIDS in a large cohort of homosexual men. *J. Infect. Dis.* 181 (3), 872-880.

Mansour, H., E. V. Nordheim, and J. J. Rutledge (1985). Maximum-likelihood estimation of variance-components in repeated measures designs assuming autoregressive errors. *Biometrics* 41 (1), 287-294.

Margolick, J. B., A. D. Donnenberg, A. Munoz, L. P. Park, K. D. Bauer, J. V. Giorgi, J. Ferbas, and A. J. Saah (1993). Changes in lymphocyte-T and non-lymphocyte-T subsets following seroconversion to HIV-1 - stable CD3⁺ and declining CD3 populations suggest regulatory responses linked to loss of CD4 lymphocytes. *J. Acquir. Immune Defic. Syndr. Hum. Retrovirol.* 6(2), 153-161.

Mellors, J. W., L. A. Kingsley, C. R. Rinaldo, J. A. Todd, B. S. Hoo, R. P. Kokka, and P. Gupta (1995). Quantitation of HIV-1 RNA in plasma predicts outcome after seroconversion. *Ann. Intern. Med.* 122 (8), 573-579.

Mellors, J. W., C. R. Rinaldo, P. Gupta, R. M. White, J. A. Todd, and L. A. Kingsley (1996a). Prognosis in HIV-1 infection predicted by the quantity of virus in plasma. *Science* 272 (5265), 1167-1170.

Mellors, J. W., C. R. J. Rinaldo, P. Gupta, R. M. White, J. A. Todd, and L. A. Kingsley (1996b). Prognosis in HIV-1 infection predicted by the quantity of virus in plasma. *Science* 272, 1124-1125.

Meng, X. L. and D. B. Rubin (1991). Using EM to obtain asymptotic variance-covariance matrices - the sem algorithm. *J. Am. Stat. Assoc.* 86 (416), 899-909.

Mori, M., G. G. Woodworth, and R. F. Woolson (1992). Application of empirical bayes inference to estimation of rate of change in the presence of informative right censoring. *Stat Med* 11, 621-631.

Moss, R. B., F. Ferre, R. Trauger, F. Jensen, A. Daigle, S. P. Richieri, and D. J. Carlo (1994). Inactivated HIV-1 immunogen - impact on markers of disease progression. *J. Acquir. Immune Defic. Syndr. Hum. Retrovirol.* 7, S21-S27.

Munoz, A., V. Carey, A. J. Saah, J. P. Phair, L. A. Kingsley, J. L. Fahey, H. M. Ginzburg, and B. F. Polk (1988). Predictors of decline in CD4 lymphocytes in a cohort of homosexual men infected with human immunodeficiency virus. *J. Acquir. Immune Defic. Syndr. Hum. Retrovirol.* 1(4), 396-404.

Murray, G. D. and J. G. Findlay (1988). Correcting for the bias caused by drop-outs in hypertension trials. *Stat Med* 7, 941-946.

Nunezanton, V. and G. G. Woodworth (1994). Analysis of longitudinal data with unequally spaced observations and time-dependent correlated errors. *Biometrics* 50 (2), 445-456.

O'Brien, T. R., W. A. Blattner, D. Waters, M. E. Eyster, M. W. Hilgartner, A. R. Cohen, N. Luban, A. Hatzakis, L. M. Aledort, P. S. Rosenberg, W. J. Miley, B. L. Kroner, and J. J. Goedert (1996). Serum HIV-1 RNA levels and time to development of AIDS in the multicenter hemophilia cohort study. *JAMA-J. Am. Med. Assoc.* 276 (2), 105-110.

Pantazis N., G.Touloumi (2009). Analyzing longitudinal data in the presence of informative drop-out: The `jmre1` command. *Stata Journal*, 10 (2), 165-190.

Park, T. and C. S. Davis (1993). A test of the missing data mechanism for repeated categorical- data. *Biometrics* 49 (2), 631-638.

Park, T. and S. Y. Lee (1997). A test of missing completely at random for longitudinal data with missing observations. *Stat. Med.* 16 (16), 1859-1871.

Phillips, A. N., C. A. Lee, J. Elford, A. Webster, G. Janossy, A. Timms, M. Bofill, and P. B. A. Kernoff (1991). More rapid progression to AIDS in older HIV-infected people - the role of CD4⁺ T-cell counts. *J. Acquir. Immune Defic. Syndr. Hum. Retrovirol.* 4 (10), 970-975.

Phillips, A. N., C. A. Sabin, J. Elford, M. Bořll, G. Janossy, and C. A. Lee (1994). Use of CD4 lymphocyte count to predict long-term survival free of AIDS after HIV-infection. *Br. Med. J.* 309 (6950), 309-313.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J.Am. Stat. Assoc.* 90 (429), 106-121.

Rochon, J. (1992). Arma covariance-structures with time heteroscedasticity for repeated measures experiments. *J. Am. Stat. Assoc.* 87 (419), 777-784.

Roitt, I., Brostoff, J., and Male, D. Immunology. Fourth Edition. Mosby. 1996

Rotnitzky, A., D. Scharfstein, T. L. Su, and J. Robins (2001). Methods for conducting sensitivity analysis of trials with potentially nonignorable competing causes of censoring. *Biometrics* 57 (1), 103-113.

Rotnitzky, A. and D. Wypij (1994). A note on the bias of estimators with missing data. *Biometrics* 50 (4), 1163-1170.

Rotnitzky, A., J. M. Robins, and D. O. Scharfstein (1998a). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *J Am Stat Assoc* 93, 1321-1339.

Rotnitzky, A., J. M. Robins, and D. O. Scharfstein (1998b). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *J. Am. Stat. Assoc.* 93 (444), 1321-1339.

Roy, J. and X. H. Lin (2002). Analysis of multivariate longitudinal outcomes with nonignorable dropouts and missing covariates: Changes in methadone treatment practices. *J Am Stat Assoc* 97, 40-52.

Rubin, D. (1978). Multiple imputations in sample surveys- a phenomenological bayesian approach to nonresponse. In *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pp. 20-34.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* 63 (3), 581-592. With comments by R. J. A. Little and a reply by the author.

Scharfstein, D. O., A. Rotnitzky, and J. M. Robins (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J Am StatAssoc* 94, 1096-1120.

Schluchter, M. D. (1992). Methods for the analysis of informatively censored longitudinal data. *Stat Med* 11, 1861-1870.

Shih, W. J. and H. Quan (1997). Testing for treatment differences with dropouts present in clinical trials - A composite approach. *Stat. Med.* 16 (11), 1225-1239.

Song, X., M. Davidian, and A. A. Tsiatis (2002). A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics* 58 (4), 742-753.

Stavola, B. and E. Christensen (1996). Multilevel models for longitudinal variables prognostic for survival. *Lifetime Data Analysis* 2 (4), 329-347.

Stein, D. S., J. A. Korvick, and S. H. Vermund (1992). CD4+ lymphocyte cell enumeration for prediction of clinical course of human-immunodeficiency virus disease - a review. *J. Infect. Dis.* 165 (2), 352-363.

Stiratelli, R., N. Laird, and J. H. Ware (1984). Random-effects models for serial observations with binary response. *Biometrics* 40 (4), 961-971.

Stram, D. O., L. J. Wei, and J. H. Ware (1988). Analysis of repeated ordered categorical outcomes with possibly missing observations and time-dependent covariates. *J. Am. Stat. Assoc.* 83 (403), 631-637.

Stubbendick, A. L. and J. G. Ibrahim (2003). Maximum likelihood methods for nonignorable missing responses and covariates in random effects models. *Biometrics* 59 (4), 1140-1150.

Touloumi, G., A. G. Babiker, M. G. Kenward, S. J. Pocock, and J. H. Darbyshire (2003). A comparison of two methods for the estimation of precision with incomplete longitudinal data, jointly modeled with a time-to-event outcome. *Stat Med* 22, 3161-3175.

Touloumi, G., A. G. Babiker, S. J. Pocock, and J. H. Darbyshire (2001). Impact of missing data due to drop-outs on estimators for rates of change in longitudinal studies: a simulation study. *Stat Med* 20, 3715-3728.

Touloumi, G., S. J. Pocock, A. G. Babiker, and J. H. Darbyshire (1999). Estimation and comparison of rates of change in longitudinal studies with informative drop-outs. *Stat Med* 18, 1215-1233.

Touloumi, G., S. J. Pocock, A. G. Babiker, and J. H. Darbyshire (2002). Impact of missing data due to selective dropouts in cohort studies and clinical trials. *Epidemiology* 13 (3), 347-355.

Wei, L. J. and D. O. Stram (1988, Jan). Analysing repeated measurements with possibly missing observations by modelling marginal distributions. *Stat Med* 7 (1-2), 139-148.

Wittes, J., E. Lakatos, and J. Probstfield (1989). Surrogate endpoints in clinical- trials- cardiovascular- diseases. *Stat. Med.* 8 (4), 415-425.

Wu, M. C. and K. Bailey (1988). Analyzing changes in the presence of informative right censoring caused by death and withdrawal. *Stat Med* 7, 337-346.

Wu, M. C. and K. R. Bailey (1989). Estimation and comparison of changes in the presence of informative right censoring - conditional linear-model. *Biometrics* 45, 939-955.

Wu, M. C. and R. J. Carroll (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* 44, 175-188.

Wulfsohn M.S, Tsiatis A.A (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* 53, 330-339.

Tsiatis, A. A., U. Dafni, V. DeGruttola, K. J. Propert, R. L. Strawderman, and M. S. Wulfsohn (1992). The relationship of CD4 counts over time to survival in patients with AIDS: Is CD4 a good surrogate marker? In *AIDS epidemiology: Methodological issues*. Birkhauser, Boston.

Verbeke, G., E. Lesaffre, and B. Spiessens (2001). The practical use of different strategies to handle dropout in longitudinal studies. *Drug Inf. J.* 35 (2), 419-434.

Vonesh, E. F. and R. L. Carter (1992). Mixed-effects nonlinear-regression for unbalanced repeated measures. *Biometrics* 48 (1), 1-17.

Vonesh, E. F., T. Greene, and M. D. Schluchter (2006). Shared parameter models for the joint analysis of longitudinal data and event times. *Stat. Med.* 25 (1), 143-163.

Zeger, S. L. and M. R. Karim (1991). Generalized linear-models with random effects a Gibbs sampling approach. *J. Am. Stat. Assoc.* 86 (413), 79-86.

Zeger, S. L., K. Y. Liang, and P. S. Albert (1988). Models for longitudinal data a generalized estimating equation approach. *Biometrics* 44 (4), 1049-1060.