Dipl. Math.
Heribert Ramroth

Methodische Aspekte bei einer Fall-Kontroll-Studie zum Kehlkopf-Karzinom:
Verallgemeinerte Two-Stage Analysen und Kontrollziehungsverfahren.

Promotionsfach: Epidemiologie / Deutsches Krebsforschungszentrum
Doktorvater: Prof. Dr. rer. nat. H. Becher

Die vorliegende Arbeit geht aus der Rhein-Neckar-Larynx Studie hervor, einer Fall-Kontroll-Studie zum Larynxkarzinom, deren Feldphase am 31. Dezember 2000 abgeschlossen wurde. Schwerpunkt der Studie war die Ermittlung von Ursachen des Larynx-Karzinoms neben Tabak- und Alkoholkonsum im beruflichen Expositionszusammenhang und die Analyse von molekulargenetischen Faktoren anhand von Blutproben. Die epidemiologischen Daten wurden per Interview mittels eines Fragebogens erfaßt.
Die Rhein-Neckar-Larynx Studie ist eine populationsbasierte, häufigkeitsgematchte Fall-Kontroll-Studie, in der die Gruppe der Kontrollpersonen zeitgleich aus der Studienpopulation ausgewählt wurde. Zur Anwendung der statistischen Analyseverfahren spielt im Allgemeinen die Zusammensetzung und somit das Auswahlverfahren dieser Kontrollpersonen eine wichtige Rolle.
Es wurde eine allgemeine Methode vorgestellt, in Populationskontrollen repräsentativ aus der Allgemeinbevölkerung entsprechend der geographischen Populationsverhältnisse auszuwählen. Stichproben von Einwohnermeldedaten, die wie in Deutschland von regionalen Rechenzentren verwaltet werden, werden in einem Kontrollpool zusammengefaßt und bilden den Quelldatensatz, aus dem im Laufe der Studie konkrete Teilnehmer ermittelt werden. Aufgrund von externen Faktoren (Zusammensetzung der Studienregion), aber auch studieninternen Faktoren zu Beginn (finanzielle und organisatorische Gründe) und im Verlauf der Studie (Erhebung von ergänzenden Stichproben) wird ein Kontrollpool die demographischen Verhältnisse der Studienpopulation nicht notwendigerweise repräsentativ widerspiegeln. Als Konsequenz muß bei der Auswahl von Populationskontrollen ein Verfahren angewendet werden, welches Probanden proportional zu den Verhältnissen der Studienregion ermittelt.
Die spezielle Ausgangsituation der Rhein-Neckar-Larynx Studie zur Ermittlung von Kontrollpersonen und das in diesem Fall angewandte Auswahlverfahren wurden dargestellt.

Ausgangspunkt für die methodischen Modellentwicklungen der vorliegenden Arbeit war die erwartete Datensituation der Rhein-Neckar-Larynx Studie:
Die demographischen Daten (Z) Alter, Geschlecht und Wohnort lagen für alle Studienteilnehmer vor. Bei der Erfassung von Interviews (X_1) und Blutproben (X_2) war jedoch damit zu rechnen, daß für Teile der Studienteilnehmer entweder keine Interviewdaten oder keine Blutproben erhoben werden könnten:
- Manche Patienten würden aufgrund der Schwere ihrer Erkrankung nicht mehr in der Lage sein, an einem Interview teilzunehmen.
- Da die Blutprobe auf freiwilliger Basis abzugeben war, würden Teile der Studienteilnehmer sich nur für die Teilnahme am Interview entscheiden.
Außerdem wurde aus Kapazitätsgründen schon in der Planungsphase festgelegt, daß die molekular-genetischen Analysen der Blutproben nur an Teilstichproben aller Teilnehmer durchgeführt werden sollten.

Die bisher bekannten Analyseverfahren konnten in dieser Datensituation nicht angewandt werden:
• Diese Datensituation mit zwei systematisch fehlenden Datenblöcken ist nur bedingt für die üblicherweise verwendeten Imputationsverfahren im missing value Kontext geeignet, da es sich hierbei nicht um vereinzelt fehlende Daten handelt.
• Klassische Analyseverfahren wie die 'complete-case' Analyse verwenden nur die vollständig vorhandenen Datensätze.
• Auf den ersten Blick scheint das Analyseverfahren im Partial Questionnaire Design (PQD) von Wacholder et al. (1994) für die Auswertung dieser Studiensituation geeignet:
Hier werden meist aus finanziellen Gründen schon bekannte Confounder X_1 und X_2 (auch Sekundär-Informationen genannt) nur für Teile der Studienteilnehmer erhoben, die Expositionsvariable Z ist jedoch für alle Studienteilnehmer vorhanden.
Die Motivation im PQD ist zwar unterschiedlich zum h-Design, so daß die 'Rollenverteilung' der Variablen Z, X_1 und X_2 in beiden Designs nicht übereinstimmt, die Datensituation stellt sich jedoch gleich dar. Für das Verfahren im PQD sind zusätzlich Verteilungsannahmen bzgl. der Maximum Likelihood Schätzung erforderlich.
Laut Wacholder et al. (1994) ist das Verfahren jedoch nicht für gematchte Studien geeignet.

Mit den bekannten Verfahren im Two-Stage-Design, die ohne Verteilungsannahmen bzgl. der gemeinsamen Verteilung von X_1 und $X2$ auskommen, ist es aufgrund der Datensituation wiederum nicht möglich, alle Daten der Rhein-Neckar-Larynx Studie mit in die Analyse einzubeziehen.
Das methodische Vorgehen von Wacholder et al. (1994) liefert jedoch einen wichtigen Ansatz für Planung und Design von Studien mit systematisch fehlenden Datenblöcken: Es wird im vorhinein festgelegt, mit welchen verschiedenen Gruppierungen von Daten zu rechnen ist, und zwar eine Gruppe von Studienteilnehmern, von denen alle in der Studie erhobenen Variablen vollständig bekannt sein sollen, zwei Gruppen, in der außer Z nur X1 bzw. X2 vollständig bekannt ist, und möglicherweise eine Gruppe, in der nur Z bekannt ist.
Wird den Studienteilnehmern der Aufbau des Designs vor Studieneintritt mitgeteilt und ihnen die Wahlmöglichkeit eingeräumt, für welche der 4 Gruppen sie sich entscheiden wollen, dann ist Wacholder et al. (1994) die missing at random Voraussetzung (MAR) erfüllt, die für die Durchführung auch von anderen Analyseverfahren im logistischen Regressionsmodell eine wichtige Voraussetzung bildet (Vach 1994).

Zur Lösung dieser Aufgabe wurde der methodische Ansatz der Verfahren im Two-Stage-Design erweitert:
Nach einer Permutation von Variablen und Beobachtungen läßt sich die allgemeine Datensituation im h-Design so aufteilen, daß sich zwei überlappende Two-Stage-Designs ergeben. Das bedeutet, daß alle Beobachtungen in einem von zwei definierten Two-Stage-Designs enthalten sind. Somit war es naheliegend, einen iterativen Ansatz der Two-Stage-Design Verfahren zu entwickeln, in dem die Parameterschätzungen aus dem einen Two-Stage-Design Verfahren in die Auswertung des anderen Two-Stage-Design aufgenommen werden konnten, und umgekehrt. Aufgrund dieser Abhängigkeit in den Parameterschätzungen der beiden Two-Stage-Designs führt das Verfahren zur Konvergenz des im h-Design zu schätzenden Parametervektors. Diese wurde hier empirisch gezeigt.
Es wurde anhand von simulierten Daten gezeigt, welche Effizienzverbesserungen im Vergleich der Verfahren 'complete-case, 'Pseudo-conditional likelihood' (Schill 1993) und dem iterativen

Verfahren im h-Design zu erreichen sind, wenn es sich bei den oben genannten Gruppen um repräsentative Teilstichproben aller Studienteilnehmer handelt. Das Verfahren wurde dabei in einer einfachen und einer allgemeineren Datensituation vorgestellt.
An einem konkreten Datenbeispiel der Rhein-Neckar-Larynx Studie mit den Interviewdaten bzgl. des Rauch- und Alkoholkonsums, sowie der Messung des Enzyms Poly(ADP-ribose) polymerase (PARP) anhand der erhobenen Blutproben, wurde das Verfahren im angewandten Fall demonstriert. Aufgrund für das h-Design nicht ausreichend vorhandener Beobachtungen ohne Interviews wurden auch hier simulierte Daten zusätzlich mit in die Auswertung einbezogen.
Das Analyseverfahren im h-Design wurde im Rahmen dieser Arbeit an einem Beispiel vorgestellt, welches prinzipiell auf Datensituationen ähnlich dem Partial Questionnaire Design erweiterbar ist. Mit diesem Verfahren sind auch gematchte Studien auswertbar, in der sich die Effizienz bzgl. der Varianzschätzungen von Expositions- und Confoundervariablen verbessern läßt, ohne Annahmen bzgl. der gemeinsammen Verteilung der Variablen angeben zu müssen. Es stellt daher eine wichtige Erweiterung von Verfahren im ’missing-value’-Kontext dar.