Kraftfahrt-Bundesamt
Abteilung Statistik

Flensburg, November 2002
durchgesehen im Mai 2005

Flensburger Evaluationsmodell zur Legalbewährung – ein Konzept zur statistischen Evaluation von Maßnahmen des „Driver Improvements“

Franz-Dieter Schade, Kraftfahrt-Bundesamt, Flensburg

1 Zielsetzung

Gegenstand der Ausführungen ist ein optimiertes Verfahren zur statistischen Evaluation von Maßnahmen des „Driver Improvements“, hier als Flensburger Evaluationsmodell zur Legalbewährung (FLEML) bezeichnet. Es geht dabei im Sinne einer summativen Evaluation darum, die Auswirkung der Maßnahmen auf das Verkehrsgeschehen zu ermitteln, um darüber den grundsätzlichen Nachweis der Wirksamkeit der Maßnahmen zu führen.

Mit diesem Verfahren soll ein Maßstab für die Bewertung von Verkehrssicherheitsmaßnahmen gesetzt werden, hier insbesondere für die Qualitätssicherung von Kurs-, Seminar- oder Beratungsmodellen zum Aufbau, zum Erhalt oder zur Wiederherstellung der Kraftfahreignung. Auch diagnose- und regelgeleitete Entscheidungen zur Kraftfahreignung - etwa im Rahmen der medizinisch-psychologischen Begutachtung - gehören zu den hier gemeinten individuellen Maßnahmen, auch wenn sie, wie im Falle eines positiven Gutachtens, nicht mit einer eigentlichen Behandlung, sondern nur mit einer Zuweisung verbunden sind. Die verschiedenen Maßnahmen (Begutachtungs-, Beratungs-, Kurs-, Seminar-, Nachschulungsmodelle) werden im Folgenden pauschal als „Maßnahmen“, die durchführenden Stellen dieser Maßnahmen pauschal als „Anbieter“, die Klienten als „Teilnehmer“ und die in die Evaluation einbezogene Klientengruppe als „Behandlungsgruppe“ angesprochen.

Das Verfahren soll einer Reihe von Kriterien genügen, um gehobenen Ansprüchen sowohl aus Feldern der Praxis als auch der Wissenschaft zu entsprechen. Es soll insbesondere auch den erhöhten Anforderungen an die Evaluation von Kursen zur Wiederherstellung der Kraftfahreignung nach § 70 FeV genügen. Die Anbieter haben dabei in einem Evaluationsverfahren nachzuweisen, dass die Teilnehmer mit ihrer Kursteilnahme eine ausreichende Kraftfahreignung erlangen. Das Verfahren soll

- statistisch-methodisch abgesichert und wissenschaftlich anerkannt sein,
- standardisiert sein, d. h. einem einheitlichen, routinemäßigen Vorgehen folgen,
- die beim Anbieter sowie bei Behörden ohnehin vorliegenden Informationen über den Teilnehmer nutzen,
- diese Informationen statistisch voll ausschöpfen,
- den Datenschutzbestimmungen genügen.

Zu behandeln ist sowohl das Verfahren der Datenerhebung und -auswertung für die Behandlungsgruppe (Teilnehmer der Maßnahmen) wie auch für die Gewinnung von Referenzwerten. Ziel ist es, unter Gesichtspunkten der theoretischen Anforderungen einen „Standard“ für die Evaluation festzulegen. Spätere Überlegungen müssen zeigen, wie unter pragmatischen Gesichtspunkten ein vertretbar reduziertes Verfahren zu beschreiben ist, das für den Anbieter praktikabel und finanzierbar bleibt.

2 Problemfelder der charakterlichen Kraftfahreignung

Gegenstand ist im Folgenden lediglich die charakterliche Kraftfahreignung. Gebräuchlich ist eine Differenzierung von Problemfeldern nach fünf verschiedenen Bereichen, in denen sich Defizite der charakterlichen Kraftfahreignung manifestieren, so wie sie sich auch in den Zuweisungsregeln für die zu ergreifenden Maßnahmen ausdrücken:

1. Problematik von Fahranfängern, die in ihrer Probezeit eine schwerwiegende oder zwei weniger schwerwiegende Zuwiderhandlungen begangen haben,
2. Problematik von „Punktetätern“ ohne (wesentliche) Alkohol- oder Drogenbeteiligung,
3. Problematik von „Punktetätern“ mit Alkoholbeteiligung (jedoch in der Regel unter 1,6 Promille BAK) oder Drogenbeteiligung,
4. Problematik von Alkohol-Ersttätern mit in der Regel mindestens 1,6 Promille BAK,
5. Problematik von Alkohol-Wiederholungstätern.

Da bei spezifischen Maßnahmen im Allgemeinen nicht von einer Breitbandwirksamkeit für alle aufgezählten Problembereiche ausgegangen werden darf, ist für jeden Problembereich eine eigene Evaluation durchzuführen.

3 Evaluationskriterien

Als Mindestforderung an die Wirksamkeit von Maßnahmen gilt, dass die Fahreignung der Teilnehmer nicht durch erneute Verkehrsauffälligkeiten alsbald (wieder) abgesprochen werden muss. *Primäres* (negatives) *Evaluationskriterium* ist also die rechtskräftige Entziehung, die Aberkennung der Fahrerlaubnis oder der Verzicht auf sie.

Gemäß einer weitergehenden Forderung, nicht erst den endgültigen Beweis der Nicht-Eignung abzuwarten, sondern bereits die Anzeichen im Vorfeld zu nutzen, sollte das gesetzliche Punktsystem verwendet werden, zumal es gerade eingerichtet wurde „zum Schutz vor Gefahren, die von wiederholt gegen Verkehrsvorschriften verstoßenden Fahrzeugführern und -haltern ausgehen“ (§ 4 StVG). Das System bewertet Verkehrsverstöße im Verkehrszentralregister (VZR) des Kraftfahrt-Bundesamtes (KBA) nach dem Grad ihrer Schwere, d. h. der mit ihrer Begehung sich manifestierenden Eignungsdefizite. Entsprechend kann als *sekundäres* (negatives) *Evaluationskriterium* das Auffällig-Werden mit Verkehrsverstößen ab einem bestimmten Schweregrad verwendet werden („Verkehrsauffälligkeit“). Als Erheblichkeitsgrenze, die eine Eintragung in das VZR überhaupt erst rechtfertigt, sieht der Gesetzgeber einen Verkehrsverstoß vor, der mit einer Geldbuße von mindestens 40 Euro oder mit einem Fahrverbot belegt wird („*VZR-Auffälligkeit*“).

Für spezielle Problemgruppen kann es nötig sein, die Evaluationskriterien enger zu setzen. So hat sich die Evaluation von Maßnahmen für Alkohol-Verkehrsauffällige vor allem oder sogar ausschließlich an der *einschlägigen* Wiederauffälligkeit, d. h. der Auffälligkeit mit einem Alkoholverstoß, zu orientieren. Die Evaluation von Maßnahmen für Fahranfänger muss dagegen die im Gesetz definierten Grenzen für „schwere oder weniger schwere Zuwiderhandlungen" beachten. Für Maßnahmen, die sich speziell gegen die Begehung grober Verkehrsverstöße richten, etwa solche mit einigem Aggressionsgehalt, kann eine Eingrenzung auf besonders schwerwiegende Verkehrsverstöße sinnvoll sein. Unter inhaltlichen Gesichtspunkten bieten sich dafür Verstöße von drei und mehr Punkten an, da ab dieser Grenze bereits Fahrverbote verhängt werden.

Die beiden vorgeschlagenen Kriterien sind dem Rechtssystem immanent, da aus ihm abgeleitet, und werden vor Behörden, Gerichten und Sachverständigen daher leicht Anerkennung finden.

Exkurs zum Unfallkriterium: Für Maßnahmen, die direkt auf die Verkehrssicherheit abzielen, besteht die Forderung, Unfälle als oberstes Evaluationskriterium zu wählen. Bei leichter Verfügbarkeit von Unfalldaten sollten diese

auch verwendet werden. Es besteht aber keine zwingende Veranlassung, sich auf Unfalldaten zu beschränken, zumal sie nicht über alle Kritik erhaben zu sein scheinen: Eine im KBA erstellte Studie zur Beurteilung von Verkehrssicherheitsmaßnahmen[1] kommt nach Kritik des Unfallkriteriums zu dem Schluss, dass für eine differenzierte Analyse der Verkehrssicherheit (ergänzend zum Unfall) verhaltensnähere Indikatoren heranzuziehen sind. Die Kritik am Unfallkriterium bezieht sich vor allem auf die Zufälligkeit des Unfallgeschehens (prinzipielle Abhängigkeit der Unfallfeststellung von den zufälligen Unfallfolgen, Unfall als letztes Glied einer Kette von Koinzidenzien, geringe Reliabilität wegen der Seltenheit des Ereignisses) und darüber hinaus auf die geringe Validität der verfügbaren Angaben (bloße Augenschein-Plausibilität, nicht erfasste Multikausalität, konzeptuelle Ausblendung wichtiger Aspekte der Verkehrseignung durch Vernachlässigung des Vorunfallgeschehens). Empfohlen wird in der genannten Studie die Verwendung von Daten zur Legalbewährung aus dem VZR, zumal kalifornische sowie australische Großstudienergebnisse belegen, dass die Zahl der in einem festen Zeitraum registrierten Verkehrsverstöße eine genauere Prädiktion zukünftiger Unfälle erlaubt als die Zahl der in diesem Zeitraum registrierten Unfälle.

Vom Erfolg einer Maßnahme soll im Rahmen des vorliegenden Evaluationsmodells gesprochen werden, wenn die Evaluationskriterien in der Behandlungsgruppe signifikant positiver ausfallen als entsprechende Referenzwerte von gerade noch als geeignet zu bezeichnenden Kraftfahrern („hartes Erfolgskriterium“).

Die Forderung, eine rehabilitative Maßnahme solle zu einer *besseren* Verkehrsbewährung führen im Vergleich zu Kraftfahrern, für die zwar erhebliche Eignungsmängel vorliegen, die aber noch rechtmäßig im Besitz einer Fahrerlaubnis sind, erscheint überzogen und unbillig zu sein. Daher kann auch die abgeschwächte Forderung für die Anerkennung eines Maßnahmenerfolgs vertreten werden, dass die Maßnahme nicht zu „*substanziell schlechteren*“ Ergebnissen führen darf als entsprechende Referenzwerte von gerade noch als geeignet zu bezeichnenden Kraftfahrern („weiches Erfolgskriterium“).

Festzulegen ist dabei, was als substanzielle Verschlechterung gelten soll. Eine Empfehlung kann sich orientieren an einer juristischen Auslegung des Begriffs „erhebliches Übersteigen des allgemeinen Risikos“[2]. Danach gilt ein spezielles Risiko, das das allgemein hingenommene, mithin akzeptierte Risiko um ein Drittel übersteigt, bereits als *erheblich*[3]. Die hier vorgenommene Grenzwertsetzung ist allerdings ein Aspekt der „Wagniswürdigung“ und fällt eher in den politischen als in den wissenschaftlichen Entscheidungsbereich.

4 Probleme der Datengewinnung und ihre Lösungen

Bei Verwendung von Daten aus dem VZR sind einige Besonderheiten dieser Datenquelle zu beachten, die, wenn sie nicht angemessen berücksichtigt werden, die Generalisierbarkeit und Vergleichbarkeit der Ergebnisse in schwer abschätzbarer Weise beeinträchtigen. Schwierigkeiten bei der Lösung dieser Probleme haben in der Vergangenheit dazu geführt, dass von geplanten Evaluationsvorhaben Abstand genommen wurde. Inzwischen liegen praktikable Lösungsvorschläge vor.

4.1 Meldeverzug

Die Eintragungen gehen dem VZR je nach Art der entscheidenden Instanz (Fahrerlaubnisbehörde, Bußgeldstelle oder Gericht), Art der Entscheidung (Verwaltungsentscheidung, Entscheidung im Ordnungswidrigkeiten- oder Strafverfahren), Umfang der Entscheidung (mit oder ohne Fahrerlaubnisentziehung) und meldendem Bundesland zum Teil nur stark verzögert zu. Dieser Sachverhalt, die Problematik des so genannten *Meldeverzugs*, mindert die

---

[1] Schade, F.-D. & Heinzmann, H.-J. (2004): Prognosemöglichkeiten zur Wirkung von Verkehrssicherheitsmaßnahmen anhand des Verkehrszentralregisters. Berichte der Bundesanstalt für Straßenwesen, Heft M 155

[2] Nach § 45 IX StVZO (OVG Hamburg, Urt. v. 07.12.1999 - 3 Bf 51/96; s. NVZ 2000, 8, S. 346ff): Eine Gefahrenlage, die das allgemeine Risiko einer Beeinträchtigung der Sicherheit des Verkehrs erheblich übersteigt und deshalb Geschwindigkeitsbegrenzungen auf Autobahnen rechtfertigt, liegt vor, wenn die Unfallhäufigkeit auf dem Streckenabschnitt ohne die Geschwindigkeitsbegrenzung um mindestens *ein Drittel* höher läge als die durchschnittliche Unfallhäufigkeit auf dem gesamten Autobahnnetz (Kursivsetzung durch den Autor).

[3] Die dortige Auslegung bezieht sich auf die Unfallgefahr, kann aber auf das Wiederauffälligkeitsrisiko übertragen werden, da in erster Näherung Wiederauffälligkeitsrisiken und Unfallrisiken einander proportional sind.

Vergleichbarkeit der gewonnenen Daten zwischen verschiedenen Untersuchungsgruppen, sofern sie hinsichtlich der genannten Merkmale sowie hinsichtlich ihrer Beobachtungszeiten nicht streng parallelisiert wurden. In der Vergangenheit hat man das Problem der unvollständigen Aktenlage weitgehend dadurch zu lösen versucht, indem man unter Inkaufnahme eines nicht unerheblichen Informationsverlustes Ereignisse der letzten 12 Monate vor der VZR-Abfrage komplett aus der Evaluation ausschloss[4].

Ein neues Verfahren im Rahmen des „FLEML“ soll Verzerrungen aufgrund unterschiedlicher Meldeverzugszeiten vermeiden und dabei die im VZR vorliegende Information vollständig ausschöpfen. Das Verfahren benötigt dazu für jeden Teilnehmer genaue Datumsangaben zum Beobachtungsbeginn, zum Beobachtungsende sowie zum realisierten VZR-Abfragezeitpunkt. Der aus Beginn und Ende ermittelte individuelle Beobachtungszeitraum wird dabei lediglich als *nominelle Beobachtungszeit* betrachtet.

Wegen des Meldeverzugs liegen aus einer bestimmten Zeitspanne (beispielsweise 12 Monate) um so weniger Verkehrsverstöße im VZR vor, je enger die VZR-Abfrage auf das Beobachtungszeitende folgt. So kommt es vor, dass aus den der VZR-Abfrage unmittelbar vorangehenden 12 Monaten erst 50 % der Verkehrsverstöße dieser Zeit registriert sind, während aus den 12 Monaten davor zum Zeitpunkt derselben VZR-Abfrage bereits 98,5 % eingegangen sind. Dies bedeutet im ersten Fall, dass die nominelle Zeit von 365 Tagen *effektiv* nur wie eine Beobachtungszeit von 182,5 Tage zu zählen ist, während im zweiten Fall eine effektive Beobachtungszeit von 360 Tagen vorliegt. Für eine Umrechnung der nominellen in die effektiven Beobachtungszeiten in Abhängigkeit von der Länge der Beobachtungszeit und ihrem zeitlichen Abstand zur VZR-Abfrage benötigt man die Verteilung der Meldeverzugszeiten. Diese Zeiten, definiert als Laufzeit zwischen Tatdatum (bzw. Entscheidungsdatum) und VZR-Eingangsdatum, werden auf Basis der jeweils letzten amtlichen VZR-Stichprobe ermittelt, und zwar separat nach den genannten Merkmalen Instanzenart, Entscheidungsart, Entscheidungsumfang und Bundesland.

Die weitere Auswertung basiert dann allein auf den effektiven Beobachtungszeiten. Das bedeutet, dass die Evaluationskriterien des 3. Kapitels auf die jeweiligen effektiven Beobachtungszeiten zu beziehen sind. Das hier dargestellte Verfahren erhöht so die Vergleichbarkeit der Ergebnisse. Es nutzt zudem mehr statistische Information, was sich auf die erforderlichen Mindeststichprobengrößen günstig auswirkt.

4.2 Tilgung

Evaluationen auf der Grundlage von VZR-Daten verwenden oft eine retrospektive Datenabfrage („*retrospektiver Evaluationsauftrag*“). Das heißt, statt das VZR regelmäßig auf „frische“ Eintragungen zu einer Person abzufragen, wird nur am Ende des Untersuchungszeitraums eine einzige VZR-Abfrage durchgeführt. Dieses kostensparende Vorgehen ist nur zulässig, wenn sicher gestellt werden kann, dass zum Zeitpunkt der Endabfrage noch alle Eintragungen aus dem interessierenden Beobachtungszeitraum vorliegen, also keine Tilgung der Person oder einzelner Eintragungen vorgenommen sein kann.

Zu bestimmen ist somit für jede Behandlungsgruppe die „*maximale tilgungsfreie Beobachtungsspanne*“, also rückblickend derjenige längste Zeitraum, in dem noch keine für die Evaluation relevanten Eintragungen getilgt sein können. Für den Vergleich verschiedener Behandlungsgruppen (oder Kontrollgruppen) ist dann der kleinste dieser Werte zu verwenden.

4.3 Unvollständige Beobachtungszeit

In vielen Fällen lässt sich eine für die Auswertung wünschenswerte gleichlange Beobachtungsspanne für alle Teilnehmer nicht realisieren. So kommt es bei feststehendem VZR-

[4] Dieses frühere Vorgehen erforderte entweder einen höheren Stichprobenumfang zum Ausgleich des Informationsverlustes oder nach 12 Monaten eine weitere VZR-Abfrage, verbunden mit Kosten und Aktualitätsverlust.

Abfragedatum, aber individuell unterschiedlichen Terminen der Maßnahmenbeendigung zwangsläufig zu unterschiedlich langen Beobachtungszeiten für die Legalbewährung: Personen, die als Erste eines Untersuchungskollektivs in die Rekrutierungsphase gelangen, besitzen dann eine längere Beobachtungszeit als Personen, die zu den Letzten gehören. Während für Erstere die gewünschte Beobachtungsphase vollständig abgeschlossen wird, absolvieren Letztere nur einen Bruchteil davon.

Dieses Problem ist in der Statistik als Problem „rechtszensierter" Daten bekannt. Unter bestimmten Voraussetzungen, insbesondere unter der Annahme, die Rechtszensierung treffe die Personen „zufällig", stehe also nicht in einem systematischen Zusammenhang mit der abhängigen Variable, hier dem Grad der Legalbewährung, kann durch statistische Methoden mit diesem Problem umgegangen werden, ohne dass es zu Ergebnisverzerrungen kommt[5].

5 Methode

5.1 Untersuchungsdesign

5.1.1 Allgemeiner Ansatz

Es wird für das Folgende vorausgesetzt, dass jede betrachtete Behandlungsgruppe hinsichtlich des Treatments (der Maßnahme) homogen ist, mit anderen Worten, die Evaluation wird für jede Maßnahmenart getrennt vorgenommen.

Die Anbieter haben in einem Evaluationsverfahren nachzuweisen, dass die Teilnehmer im statistischen Mittel (nicht aber als Einzelpersonen) mit ihrer Maßnahme (nicht aber im streng kausalen Sinne *durch* ihre Maßnahme) eine ausreichende Kraftfahreignung erlangen. Die Aufgabe entspricht einer „Outcome-Evaluation". Untersucht werden dabei nicht die Durchführungsmodalität der Maßnahme („Prozessevaluation"), die Qualität der Maßnahme („input evaluation"), der Grad der Erreichung und Beeinflussung der Teilnehmer („impact evaluation") – dies alles wird als erfolgreich vorausgesetzt –, sondern im Sinne einer summativen Evaluation für einen Wirksamkeitsnachweis lediglich die Auswirkung der Maßnahme im Verkehrsgeschehen: die Auffälligkeit im Straßenverkehr.

In die Evaluation einzubeziehen ist die *unausgelesene Gesamtmenge* aller Teilnehmer, die in einem festgelegten Zeitraum („*Rekrutierungszeitraum*") ein festgelegtes Ereignis („*Aufnahmeereignis*") zu verzeichnen haben. Bei zu großen Gruppen kann aus Aufwandsgründen auch eine Zufallsstichprobe daraus gewählt werden. Im Folgenden wird davon ausgegangen, dass das Aufnahmeereignis der Abschluss der zu evaluierenden Maßnahme ist. Die so gewonnene Gruppe wird üblicherweise als „sequenzielle Stichprobe aus dem Behandlungsgut" aufgefasst.

Das Evaluationskonzept sieht vor, den Erfolg der Maßnahmen durch Vergleich der Kriteriumswerte der Behandlungsgruppe mit geeigneten *Referenzwerten* festzustellen. Sind solche nicht verfügbar, müssen sie entsprechend der Forderung eines kontrollierten Untersuchungsdesigns anhand einer adäquaten *Kontrollgruppe* ermittelt werden.

5.1.2 Beobachtungszeit

Als Beginn der Beobachtungszeit zur Legalbewährung ist ein „*Startereignis*" zu definieren. Dies kann mit dem Aufnahmeereignis – Abschluss der Maßnahme – identisch sein, es kann auch ein anderes, besonders dokumentiertes Datum sein, wie etwa das Datum der Ausstellung oder Vorlage der Teilnahmebescheinigung oder das Datum der später eingeholten Einwilligungserklärung des Teilnehmers zur VZR-Abfrage. Aufnahmeereignis und Startereignis können im Prinzip logisch wie zeitlich unabhängig gewählt werden. Dies ist der Fall, wenn

---

[5] Blossfeld, H.-P., Hamerle, A. und Mayer, K.U. (1986): Ereignisanalyse. Campus-Verlag, Frankfurt/Main

die Legalbewährungszeit erst mit der Neuerteilung einer Fahrerlaubnis beginnen soll (davon wird im Folgenden ausgegangen).

Neben dieser üblichen Form der Evaluation, bei der die Legalbewährung direkt nach Abschluss der Maßnahme oder zumindest im zeitlich nahen Umfeld geprüft wird („*anschließende Beobachtung*"), kommt eine andere Form vor: die „*abgerückte Beobachtung*". Hierbei nimmt man in Kauf, dass eine größere Zeitspanne zwischen der Maßnahme und der Prüfung auf Legalbewährung verstreicht („Blind-Intervall"). Während beim Verfahren der anschließenden Beobachtung das Zeitintervall zwischen Maßnahmenabschluss und Beobachtungsbeginn, wenn es nicht null ist, eher klein ausfällt und interindividuell kaum variiert, zeichnet sich das Verfahren der abgerückten Beobachtung durch ein im Mittel langes Intervall aus bei oftmals hoher interindividueller Streuung. Beim Verfahren der anschließenden Beobachtung kommt man gewöhnlich zu individuellen Terminen für Beobachtungsbeginn und -ende, während das Verfahren der abgerückten Beobachtung die Festlegung einer kollektiven Beobachtungsspanne mit einheitlichem Start- und Endedatum ermöglicht.

Die abgerückte Beobachtung führt zu einem eigentlich unerwünschten „Blind-Intervall", einer Zeit nach Maßnahmenabschluss, über die mangels Beobachtung keine Aussage möglich ist. Dieser Fall tritt auf, wenn sich der Anbieter erst lange nach Maßnahmenabschluss zu einer Evaluation entschließt (retrospektiver Evaluationsauftrag), so dass zum Zeitpunkt der Datenerhebung, d. h. der VZR-Abfrage, wegen der zwischenzeitlichen Tilgung nicht mehr für den gesamten Zeitraum VZR-Eintragungen vorliegen.

Beispiel für anschließende Beobachtung: Eine Gruppe von 150 Teilnehmern, die eine Maßnahme in den Jahren 2000 bis 2004 absolviert hat, wird in den Jahren 2002 bis 2005 auf VZR-Eintragungen abgefragt, und zwar jeder Teilnehmer zu einem individuell festgelegten Datum 24 Monate nach Abschluss seiner Maßnahme, wobei nur Eintragungen gezählt werden, die ein Tatdatum aus den auf die Maßnahme unmittelbar folgenden 12 Monaten tragen. Damit steht ein Beobachtungsvolumen von 150 Personenjahre zur Verfügung. Gegenstand der Untersuchung ist die Legalbewährung in der Spanne zwischen 0 und 12 Monaten, im Schnitt 6 Monate nach Maßnahmenabschluss.

Beispiel für abgerückte Beobachtung: Eine Gruppe von 150 Teilnehmern, die eine Maßnahme in den Jahren 2000 bis 2004 absolviert hat, wird Ende des Jahres 2007 auf VZR-Eintragungen mit Tatzeitdatum aus dem Jahr 2006 abgefragt. Die Abfrage erfolgt einmalig und für die ganze Gruppe en bloc. Damit steht ein Beobachtungsvolumen von 150 Personenjahre zur Verfügung. Gegenstand der Untersuchung ist die Legalbewährung in der Spanne zwischen 1 Jahr und 7 Jahren, im Durchschnitt etwa 4 Jahre nach Maßnahmenabschluss.

Das Design der abgerückten Beobachtung besitzt zwei Spielarten: Die übliche untersucht die Legalbewährung innerhalb einer gewählten Beobachtungsspanne, die ungewöhnlichere Spielart prüft zu einem gewählten Beobachtungsstichtag für jeden Teilnehmer, ob ein bestimmter Zustand vorliegt, nämlich im Sinne des primären Evaluationskriteriums eine gültige Fahrerlaubnis und im Sinne des sekundären Evaluationskriteriums eine Registrierung im VZR. Gefragt ist zum Beispiel zu einem Stichtag zwei Jahre nach dem Maßnahmenabschluss nach dem Anteil der Teilnehmer, der im Besitz einer gültigen Fahrerlaubnis ist. Diese Beobachtung ist abgerückt, weil bei dieser bewusst einfach gehaltenen Fragestellung nicht geklärt ist, was in der Zwischenzeit geschah, insbesondere ob die Teilnehmer durchgehend oder erneut im Besitz einer Fahrerlaubnis sind oder überhaupt keine neue Fahrerlaubnis nach Maßnahmenabschluss mehr beantragt hatten.

Das abgerückte Design hat zwar inhaltliche Nachteile, jedoch den untersuchungstechnischen Vorteil der einmaligen VZR-Abfrage. Zudem können so auch noch nachträglich Datenbestände eines Anbieters für die Evaluation genutzt werden (retrospektiver Evaluationsauftrag). Im Folgenden wird dennoch vom üblichen und empfehlenswerten Erhebungsdesign einer (nahezu) anschließenden Beobachtung ausgegangen.

Zu definieren ist schließlich das Ende der Beobachtungsphase. Dieses ergibt sich entweder durch Festlegung der individuellen oder kollektiven Beobachtungsdauer oder auch aus dem Zwang bei retrospektiven Aufträgen, möglichst schnell eine VZR-Abfrage durchzuführen (um einer Tilgung zuvorzukommen). In vielen dieser Fälle lässt sich eine gleichlange Beobach-

tungsspanne nicht für alle Teilnehmer realisieren und es kommt zu rechtszensierten Daten (siehe 4.3).

Oft setzen untersuchungstechnische Randbedingungen Einschränkungen für die Wahl der Beobachtungsspannen. Zu unterscheiden ist

- der retrospektive vom prospektiven Evaluationsauftrag,
- der Evaluationsauftrag mit nur einmaliger VZR-Abfrage pro Person vom dem mit mehrmaliger VZR-Abfrage pro Person,
- der Evaluationsauftrag mit gruppenweiser VZR-Abfrage, der eine kollektive Beobachtungsphase erfordert, vom dem mit individueller VZR-Abfrage, der völlig individuelle Start- und Endetermine gestattet.

Jede Kombination aus diesen drei Kriterien ist möglich (wenn auch nicht immer sinnvoll). Im Folgenden wird von dem für eine aussagefähige Evaluation günstigsten Fall ausgegangen: dem prospektiven Evaluationsauftrag mit, wenn nötig, mehrmaliger VZR-Abfrage, gesteuert von individuellen Terminen.

Generell ist für eine Evaluation zu bedenken, dass die Beobachtungszeit im zeitlich engen Zusammenhang zur durchgeführten Maßnahme stehen soll. Daher wird empfohlen, Zeiten, die mehr als fünf Jahre nach Maßnahmenabschluss liegen, für die Evaluation nicht mehr zu betrachten. Denn Forderungen, *einmalig* durchgeführte Maßnahmen müssten auch noch nach mehr als fünf Jahren nachweisbare Wirkungen zeigen, können als unbillig und unrealistisch zurück gewiesen werden. Außerdem sollte die Auswahl „nicht-repräsentativer" Zeitabschnitte vermieden werden. So ist beispielsweise eine Beobachtung der ersten zwei Monate nach Neuerteilung mit einer Beobachtung aus dem dritten Jahr nach Neuerteilung kaum vergleichbar. Daher ist es für die Vergleichbarkeit von Evaluationsergebnissen wichtig, den gewählten Beobachtungszeitraum relativ zum Ende der Maßnahme anzugeben. Dafür sind drei Größen notwendig:

- der mittlere Beobachtungsbeginn in Monaten nach Ende der Maßnahme,
- das mittlere Beobachtungsende in Monaten nach Ende der Maßnahme und
- der „Beobachtungsschwerpunkt" als Mitte zwischen Beginn und Ende.

Wegen der erforderlichen Nähe zwischen Maßnahme und Erfolgskontrolle wird empfohlen, dass der Beobachtungsschwerpunkt einer Behandlungsgruppe nicht weiter als zwei Jahre nach Maßnahmenabschluss liegt.

5.1.3 Ausschlusskriterien

Einheitliche Ausschlusskriterien für die Evaluation dienen dazu, eine aussagefähige Datenbasis für den Vergleich und für die Verallgemeinerung der Ergebnisse zu schaffen. Um Verzerrungen durch besondere Bedingungen zu vermeiden, werden bestimmte Fälle aus der Evaluation ausgeschlossen. Ausschlusskriterien *für Teilnehmer* sind:

a) es fehlt der erfolgreiche Abschluss der Maßnahme im Sinne vorher festgelegter interner Kriterien („Abbrecher");

b) das definierte Startereignis tritt nicht innerhalb einer Maximalzeit nach Aufnahmeereignis ein („verschleppte Fälle"); Empfehlung: neun Monate[6] nach Maßnahmenabschluss sollte

[6] Scheucher, Eggerdinger & Aschersleben (2002) ermittelten bei 66 ihrer Klienten, dass die Fahrerlaubnis durchschnittlich sogar erst nach einem Jahr nach Kursende ausgehändigt wurde (Scheucher, B., Eggerdinger, C. & Aschersleben, G., 2002: 5 Jahre danach – Welche überdauernden Veränderungen werden durch eine Verkehrstherapie für alkoholauffällige Kraftfahrer erreicht? Blutalkohol, 39, 154-173); Werwath fand bei 137 positiv Begutachteten, dass sie eine Fahrerlaubnis innerhalb von durchschnittlich zwei Monaten zurück erhielten, während es bei 179 negativ Begutachteten ca. 9 Monate waren (hierin ist die Zeit der Teilnahme an einer Maßnahme eingeschlossen; Werwath, C. (2000): Zum Stellenwert von Obergutachten im Fahreignungsbegutachtungsprozess. In: Püschel, K. (Hrsg.): Schriftenreihe Forschungsergebnisse aus dem Institut für Rechtsmedizin der Universität Hamburg, Band 2, Verlag Dr. Kovač, Hamburg).

eine Neuerteilung vorliegen, da andernfalls von einem für die Evaluation irrelevanten Problem auszugehen ist (z.B. Krankheit, Auslandsaufenthalt, finanzielle Schwierigkeiten);

c) der Beobachtungsschwerpunkt (siehe oben) liegt zu eng am Startdatum ("Kurzläufer"); Empfehlung: nicht weniger als ein Monat, da die ersten Wochen nach dem Startereignis untypisch sein können[7], z. B. weil die Teilnehmer erst seit kurzer Zeit ihre Fahrerlaubnis wieder besitzen und sie für einen Großteil dieser Zeit noch nicht wieder über ein Fahrzeug verfügen oder sich ihre Mobilitätsgewohnheiten noch nicht wieder stabilisiert haben;

d) der Anteil der effektiven an der nominellen Bebachtungszeit (s. 4.1) ist zu gering („Spätläufer"); Empfehlung: nicht weniger als 20 %, weil das Auswertungsergebnis andernfalls zu empfindlich von statistischen Schwankungen im Meldeverzug abhängt, d. h. das Korrekturverfahren überfordert;

e) es handelt sich um einen seltenen, nach einem vorher festgelegten Katalog definierten Sonderfall („Sonderfälle"); dieser sollte ausgeschlossen werden, um die Evaluation nicht mit untypischen Randproblemen zu belasten[8].

Ausschlusskriterien *für Eintragungen* sind:

f) Verkehrsverstöße mit Tatzeitpunkt außerhalb der individuell festgesetzten Beobachtungsspanne (d. h. vor dem Startdatum oder nach dem Endedatum),

g) Verkehrsverstöße mit (tilgungswirksamem) Entscheidungs- oder Rechtskraftdatum vor Beginn der für alle Untersuchungsgruppen einheitlich festgesetzten tilgungsfreien Beobachtungsspanne (siehe 4.2),

h) Eintragungen, die zwar in der Beobachtungszeit, jedoch nach einem vorher definierten „*terminierenden Ereignis*" auftreten, z.B. nach einer – durchaus auch nur vorläufigen – Fahrerlaubnisentziehung.

Aus praktischen Erwägungen sowie theoretischen Forderungen nach statistischer Unabhängigkeit der Beobachtungsereignisse sollten mehrere Zuwiderhandlungen am selben Tag in geeigneter Weise zu einem einzigen Beobachtungsfall zusammengefasst werden. Damit kann die Rechtsproblematik von Tateinheit und Tatmehrheit (z.B. Rotlichtverletzung wegen zu schnellen Fahrens mit Unfallfolge und darauf folgende Unfallflucht) umgangen werden. Außerdem entstehen ohne diese Regelung Probleme für die Modellvoraussetzung der statistischen Unabhängigkeit der Verkehrsverstöße, auf der das Auswertungskonzept beruht. Eine für die der Evaluation sinnvolle Aggregation von Tatmehrheiten am selben Tag ist z.B. die Auswahl und Weiterverwendung der im Sinne der Evaluationskriterien bzw. des Punktkatalogs schwersten Tat.

Das folgende Schema (Abbildung 1) setzt die angesprochenen Ereignisse und Zeitspannen des Beobachtungsdesigns zueinander ins Verhältnis.

---

[7] Beispielsweise ein Beobachtungsvolumen von 100 Beobachtungsjahren, das sich allein aus der Auswertung von 1000 „Kurzläufern" ergibt, hat sicherlich eine andere Bedeutung als das gleiche Beobachtungsvolumen von Personen, die über eine längere Zeitspanne in Beobachtung standen. Für das Problem rechtszensierter Daten (siehe 4.3) ist aber gerade die Annahme nötig, dass die Legalbewährung praktisch unabhängig von der Länge der Beobachtungszeit ist. Daher sollten Personen, von denen im Vorhinein angenommen werden muss, dass sie diese Annahme besonders eklatant verletzen, aus der Analyse ausgeschlossen werden.

[8] Wenn beispielsweise bei Maßnahmen zur Problematik von Alkohol-Wiederholungstätern in der Regel keine Frauen der höheren Altersstufe teilnehmen, so sollten Fälle, in denen diese Konstellation ausnahmsweise doch vorliegt, nicht in die Evaluation eingehen.

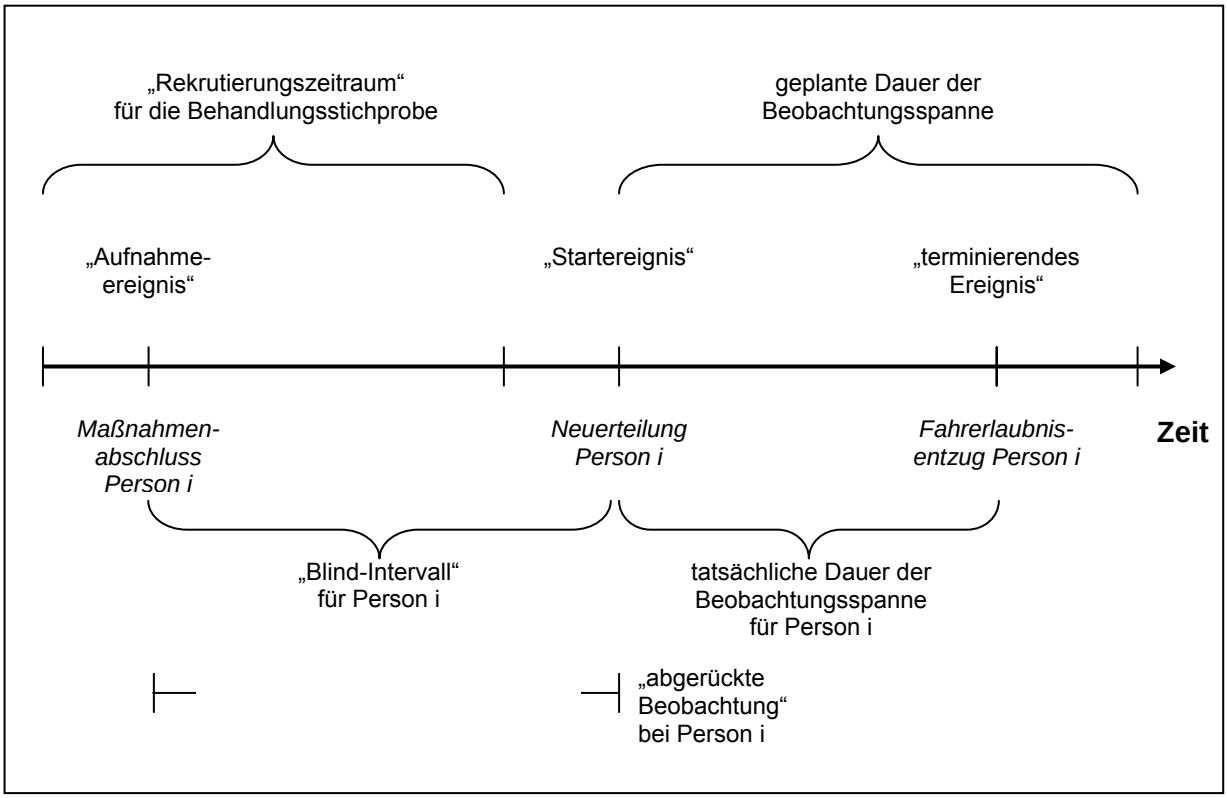

Abbildung 1: Ereignisse und Zeitspannen zur Festlegung des Beobachtungsdesigns

5.2 Kenngrößen zur Evaluation

Getrennt nach dem primären Evaluationskriterium, der Entziehung (oder Aberkennung, Verzicht), und dem sekundären Evaluationskriterium, der VZR-Auffälligkeit oder VZR-Wiederauffälligkeit, sind für die Behandlungsgruppe entsprechende statistische Kenngrößen zu berechnen. Da die Kriterien weitgehend unabhängig voneinander definiert sind, hat eine Evaluation beide Kriterien zu berücksichtigen. Für spezielle Problemgruppen mag es dennoch sein, dass nur eins der beiden Kriterien differenzierbare und aussagekräftige Ergebnisse liefert. Bei der Evaluation von Maßnahmen für die Alkohol-Problemgruppe beispielsweise fallen primäres und sekundäres Evaluationskriterium praktisch zusammen, denn ein Großteil der einschlägigen Eintragungen führt zum Führerscheinverlust, so dass für diese Gruppe eine Evaluation nach dem primären Kriterium ausreichen dürfte. Für die Gruppe der „Punkte-Täter“ dagegen ist das sekundäre Kriterium von größerer Bedeutung, weil hier der Führerscheinverlust erst sehr spät einsetzt.

5.2.1 Kenngröße zum primären Kriterium: „Führerschein-Lebensdauer“

Die Entziehung der Fahrerlaubnis entspricht in der Kette der VZR-relevanten Verkehrsereignisse etwa dem, was in der Terminologie der stochastischen Ereignisanalyse[9] ein „absorbierender Endzustand“ genannt wird. Dies in dem Sinne, dass dieser Zustand nicht mehr (ohne weiteres) verlassen werden kann und damit den bisherigen verhaltensgenerierenden Prozess beendet. Eine solcher Ereignistyp wird effizient nach der Sterbetafel-Methode oder einem der elaborierteren Verfahren (Kaplan-Meier oder Cox-Regression) ausgewertet. Abhängige Variable ist hier in jedem Fall die Zeit bis zum Eintritt einer rechtskräftigen Fahrerlaub-

[9] siehe Fußnote 5

nisentziehung. Auch hier kann und sollte statt der nominellen die effektive Beobachtungszeit verwendet werden, um Verzerrungen durch die Meldeverzugsproblematik zu vermeiden.

Die genannten Verfahren setzen nicht voraus, dass alle Teilnehmer bis zum Eintritt der Fahrerlaubnisentziehung beobachtet werden, sondern ermöglichen die Ermittlung von Kennwerten auch dann, wenn ein Teil schon vorher das Beobachtungsende erreicht (rechtszensierte Daten, siehe 4.3). Benötigt werden dafür zwei, in der Regel als unproblematisch anzusehende Annahmen: 1. Das Verhalten der am Anfang der Rekrutierungsphase in die Stichprobe gelangenden Teilnehmer unterscheidet sich nicht systematisch vom Verhalten der erst gegen Ende hinzu kommenden Teilnehmer. 2. Die Länge der für einen Teilnehmer verfügbaren Beobachtungszeit ist statistisch unabhängig von dessen „Führerschein-Lebensdauer".

Als einfacher statistischer Gruppenkennwert für die Evaluation einer Maßnahme kann die Dauer der Verkehrsbeteiligung dienen, nach der 50 % der Teilnehmer ihre Fahrerlaubnis eingebüßt haben. Dies ist ein anschaulicher Kennwert für die mittlere „Lebensdauer des Führerscheins" im untersuchten Kollektiv.

Die grafische Darstellung über die Zeit (die so genannte Survivor-Funktion, also der Anteil der „Überlebenden" in Abhängigkeit von der Zeit; siehe Abbildung 2) erlaubt in vielen Fällen eine noch informativere Analyse der Daten. Aus einer solchen Grafik kann abgelesen werden, wie viel Prozent der Teilnehmer einer Untersuchungsgruppe nach einem bestimmten Zeitraum noch im Besitz der Fahrerlaubnis sind. Da derartige Kurven im Allgemeinen nicht linear verlaufen, lassen sie sich nicht unmittelbar mit einen einzigen Parameter charakterisieren. Hier kann eine grafische Darstellung des „Survivor-Anteils" (hier dem Anteil derjenigen, die bis zu einem bestimmten Zeitpunkt noch im Besitz ihre Fahrerlaubnis sind) gegen die Zeit weiter helfen. Bei Anwendung einer der in der Literatur empfohlenen Transformationen, z.B. des Logarithmus, kann sich angenähert eine Gerade ergeben (siehe Abbildung 3). Dies wäre ein Beleg für ein einfaches Verhaltensmodell[10], das im Idealfall durch einen einzigen Parameter bestimmt ist. Ein solcher Verhaltensparameter würde dann die betrachtete Gruppe sinnvoller charakterisieren als der oben beschriebene 50-Prozent-Survivor-Wert und einen Vergleichsmaßstab hoher Aussagekraft liefern.

Eine wichtige deskriptive Größe in Survivor-Modellen ist die so genannte *Hazardrate*: die bedingte Wahrscheinlichkeit, dass eine Person, sofern sie bis zum betrachteten Zeitpunkt „überlebte", im nächsten Zeitabschnitt ein kritisches Ereignis („Hazard"; hier die Einbuße des Führerscheins) zu verzeichnen hat. Das Cox-Modell ermittelt systematische Hazardraten-Unterschiede zwischen Untergruppen. Das Modell akzeptiert dabei einen beliebigen Verlauf der Hazardraten über die Zeit, sofern er sich in allen Untergruppen gleichartig (proportional) zeigt. Jede Untergruppe kann dann durch die Proportionalitätskonstante charakterisiert werden, um die sie sich vom allgemeinen Verlauf unterscheidet.

Um Unterschiede zwischen Behandlungs- und Kontrollgruppe zufallskritisch abzusichern, können für die ermittelten Kennwerte Konfidenzintervalle berechnet werden. Überschneiden sich die Konfidenzintervalle der Gruppen nicht, so kann von ihrer Verschiedenheit hinsichtlich der „Führerschein-Lebensdauer" ausgegangen werden. Die vom Statistik-Auswertungsprogramm SPSS durchgeführte Sterbetafel-Analyse beinhaltet einen Signifikanztest auf Unterschiede zwischen den Untersuchungsgruppen. Die Cox-Regression ermittelt für die Behandlungsgruppe (im Vergleich zur gewählten Kontrollgruppe) einen Regressionskoeffizienten (umrechenbar in die besagte Proportionalitätskonstante). Dieser kann signifikanz-

---

[10] Ergibt sich – wie im gewählten Beispiel – bei logarithmischer Transformation der y-Achse eine Gerade, so liegt der Fall vor, dass die bedingte Wahrscheinlichkeit, im nächsten Zeitabschnitt den Führerschein einzubüßen, sofern er am Anfang dieses Zeitabschnitts noch vorhanden war, über eine Konstante ist. Das würde bedeuten, dass das Risiko eines Führerscheinverlusts in der gesamten Beobachtungszeit nach Neuerteilung unverändert bleibt und mit einer einzigen Kennzahl beschrieben werden kann.

statistisch auf Abweichung von null geprüft werden. Bei Signifikanz ist von der Ungleichheit der „Führerschein-Lebensdauern“ in den verglichenen Gruppen auszugehen.

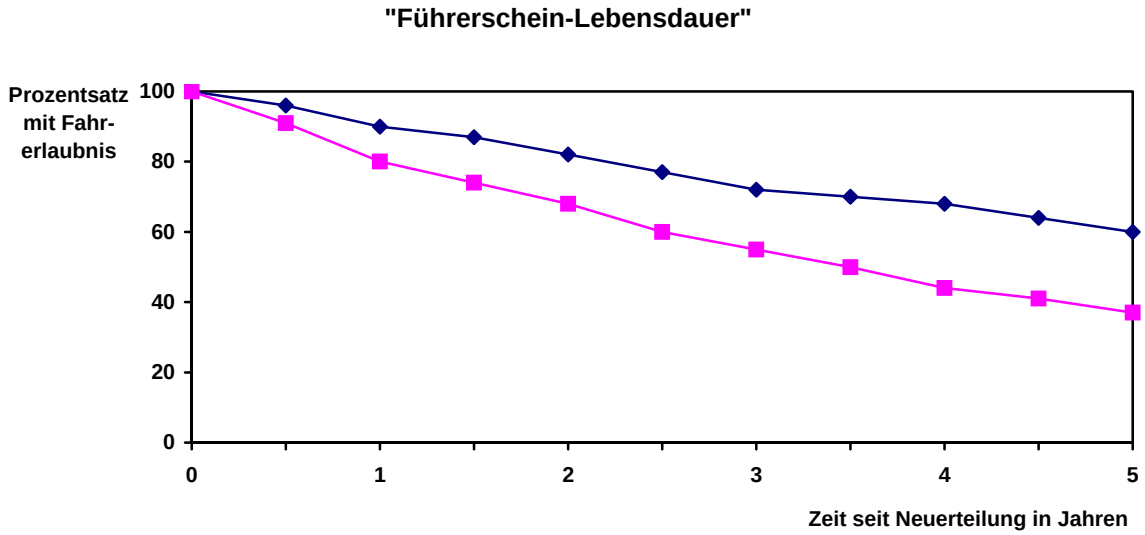

Abbildung 2: „Führerschein-Lebensdauer“, Survivor in Prozent, Vergleich von Behandlungs- und Kontrollgruppe (fiktive Daten)

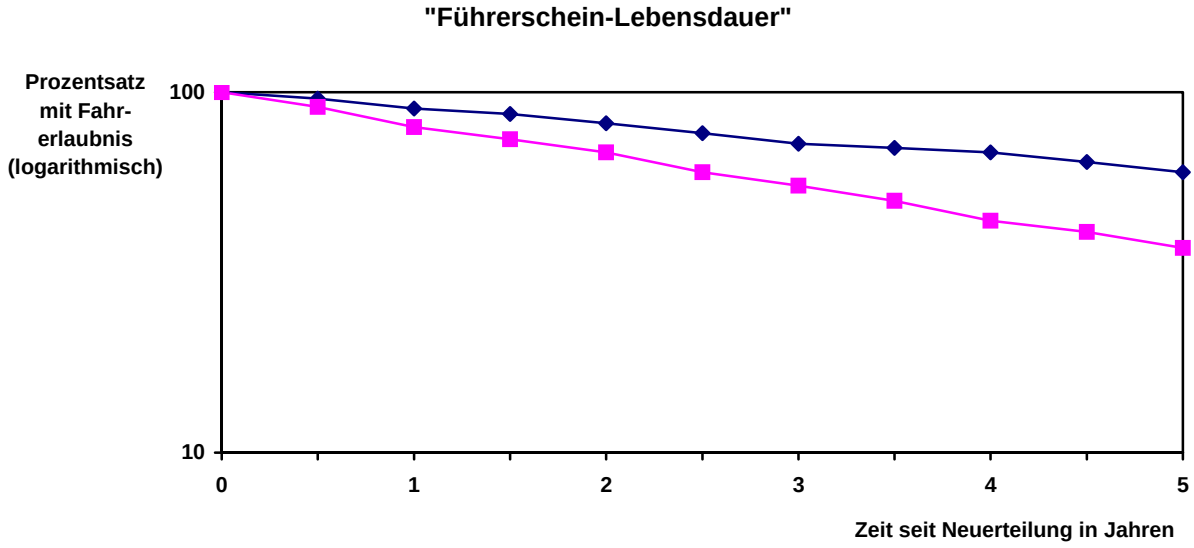

Abbildung 3: „Führerschein-Lebensdauer“, Survivor im logarithmischen Maßstab, Vergleich von Behandlungs- und Kontrollgruppe (fiktive Daten)

5.2.2 Kenngröße zum sekundären Kriterium: „VZR-Auffälligkeitsrate“

Häufigkeitszahlen von seltenen Ereignissen, wie sie Verkehrsauffälligkeiten oder auch Unfälle darstellen, folgen in hinreichend homogenen Personengruppen in guter Approximation einer Poisson-Verteilung[11]. Voraussetzung dafür ist die statistische Unabhängigkeit der Er-

[11] Kenning, J. (1968): Untersuchungen über die Beziehungen zwischen Verkehrsunfällen und Verkehrsverstößen. Inaugural-Dissertation der Mathematisch-Naturwissenschaftlichen Fakultät, Universität Köln.

eignisse: Das Auftreten eines Ereignisses in einem Zeitabschnitt darf die Wahrscheinlichkeit für ein Auftreten im nächsten Zeitabschnitt nicht merklich beeinflussen. Je seltener die Ereignisse der betrachteten Art sind, desto länger dürfen die Zeitabschnitte gewählt werden, für die diese Unabhängigkeitsforderung zu erfüllen ist. Bei VZR-Auffälligkeiten sind dafür durchaus mehrere Monate in Betracht zu ziehen. Die statistische Unabhängigkeit von Verkehrszuwiderhandlungen, die für einen Abstand von einigen Minuten, Stunden oder Tagen noch fraglich sein mag, dürfte im Abstand einiger Monate sehr gut erfüllt sein.

Die Approximation durch die Poisson-Verteilung ist bei VZR-Auffälligkeiten nach Erfahrungen mit KBA-Daten so gut, dass Abweichungen von dieser theoretischen Verteilung, wenn es denn solche geben sollte, erst in extrem großen Stichproben sichtbar werden können. Die Poisson-Verteilung ist durch einen einzigen Parameter, die *Rate* (auch: Hazardrate), vollständig charakterisiert. Die Rate kann interpretiert werden als die durchschnittliche Zahl der Ereignisse pro Zeiteinheit. Ferner kann der Interpretation dienen, dass (für nicht zu lange Zeitabschnitte) das Produkt aus der Rate und der Länge der betrachteten Zeitspanne die Auftretenswahrscheinlichkeit für das Ereignis in dieser Zeit ergibt. So bedeutet beispielsweise eine Hazardrate von 0,01 pro Monat angenähert eine Wahrscheinlichkeit von 0,12 pro Jahr (12 % werden pro Jahr auffällig).

Alle anderen statistischen Angaben, die sich in der Literatur zur VZR-Auffälligkeit oder zu Rückfallbetrachtungen finden, lassen sich bei Gültigkeit des Poisson-Modells direkt und restlos aus diesem einzigen Parameter, der Rate, ableiten und repräsentieren daher *keine eigenständigen* Phänomene, so z.B.

- die mittlere Zahl der VZR-Auffälligkeiten pro Jahr (oder von jeder anderen Zeitspanne),
- die Varianz dieser Zahlen über verschiedene Beobachtungsphasen (z.B. Kalenderjahre),
- der Anteil der Personen mit (mindestens einer) VZR-Auffälligkeit in einem beliebig gewählten Zeitraum, z.B. in zwei, drei oder fünf Jahren („*Wiederauffälligkeitsquote*“),
- der Anteil der Personen mit Mehrfach-Auffälligkeit pro Jahr (oder beliebig gewähltem Zeitraum),
- die Zeit, in der beispielsweise 50 % (oder ein beliebig anderer Anteil) der Personen VZR-auffällig werden (gelegentlich als Maß für die Rückfallgeschwindigkeit verwendet).

Diese Kennwerte sind nicht fundamental und können daher im Allgemeinen nicht empfohlen werden. Als Ausnahmen mögen gelten die Notwendigkeit, Fachfremden die Evaluationsergebnisse anschaulicher zu machen, oder der Wunsch, eine Vergleichbarkeit zu Angaben der älteren Literatur herzustellen.

Der pragmatische Gebrauch der abgeleiteten Maße führt, wenn sie nicht auf dem Poisson-Modell beruhen, zu unauflösbaren Interpretationsproblemen. Wenn beispielsweise die Literatur angibt, dass 30 % in 3 Jahren rückfällig werden, ist dann ein gefundener Wert von 21 % in 2 Jahren besser oder schlechter? Darf linear inter- und extrapoliert werden: 30 % in drei Jahren, also 20 % in zwei Jahren? Wie sollen Wiederauffälligkeitsquoten in den nicht seltenen Fällen der Praxis berechnet werden, in denen Personen unter verschieden langer Beobachtung stehen? – Diese und viele weiteren Probleme sind im Umgang mit der Hazardrate auf Basis des Poisson-Modells einfach und eindeutig gelöst.

Als statistischer Gruppenkennwert für die Evaluation der Maßnahme dient die mittlere (Wieder-) Auffälligkeitsrate der Teilnehmer pro Jahr. Die Auswertung setzt dafür die Anzahl aller Ereignisse (VZR-Auffälligkeiten) in der zu untersuchenden Gruppe zur Gesamtsumme an Beobachtungszeit der Gruppe ins Verhältnis. Genutzt werden hierbei im Sinne einer maximalen statistischen Informationsausschöpfung alle Mehrfachauffälligkeiten in der Beobachtungszeit (im Gegensatz zum einfacheren Verfahren der Berechnung von Auffälligkeits*quoten*, das nur die erste Auffälligkeit im Beobachtungszeitraum verwendet und etwaig weitere ignoriert). Die Auffälligkeits*rate* (in einem stationären Poisson-Prozess) – dies ist ein weiterer Vorteil gegenüber der früher verwendeten Auffälligkeits*quote*, die empfindlich von der (willkürlich) gewählten Länge der Beobachtungsspanne abhängt – stellt ein zeitunabhän-

giges Maß dar[12]. Dieses Maß besitzt zudem die Qualität einer Rationalskala, so dass Werte zu Vergleichszwecken ins Verhältnis zueinander gesetzt werden dürfen.

Um Unterschiede zwischen Behandlungs- und Kontrollgruppe zufallskritisch abzusichern, kann eine so genannte Poisson-Regression durchgeführt werden. Dabei sind so genannte Dummy-Variablen zur Unterscheidung zwischen den Untersuchungsgruppen einzuführen. Die Poisson-Regression ermittelt für die Behandlungsgruppe (im Vergleich zur Kontrollgruppe) einen Regressionskoeffizienten. Dieser kann signifikanzstatistisch auf Abweichung von null geprüft werden. Bei Signifikanz ist von der Ungleichheit der Auffälligkeitsraten in den verglichenen Gruppen auszugehen.

Exkurs zur Poisson-Regression[13]. Das Modell der Poisson-Regression geht, übertragen auf den Anwendungsfall von VZR-Eintragungen, von zwei Modellannahmen aus:

Erstens, dass eine Zählvariable Y, nämlich die Zahl der VZR-Eintragungen pro Zeitabschnitt der Länge b, in einem hinreichend homogenen Unterkollektiv einer Poisson-Verteilung folgt mit dem Parameter λ, dem *Risiko*. Danach gilt bei gegebenem Risiko λ und gegebenem Zeitabschnitt b für die Wahrscheinlichkeit w dafür, dass genau y VZR-Eintragungen vorliegen:

$$w(y, b, \lambda) = \frac{(b\lambda)^y e^{-(b\lambda)}}{y!}$$

Der Erwartungswert für die Zählvariable Y, also der VZR-Eintragungen ist dabei bλ, die Varianz ebenfalls bλ.

Zweitens wird angenommen, dass das Risiko λ selbst eine „einfache" Funktion ist von bestimmten „*unabhängigen Variablen*" x_1, x_2, x_3 ..., deren Kombinationen je homogene Unterkollektive bilden.

Die Funktion soll von der Art sein:

$$\lambda = e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ...)}$$

Die Poisson-Regression hat nun wie jede Regressionsrechnung die Aufgabe, die abhängige Variable aus den unabhängigen Variablen zu schätzen. Die abhängige Variable ist hier der Erwartungswert der VZR-Eintragungen (der Zählvariable Y) in den Unterkollektiven. Berechnet wird er aus den in diesen Unterkollektiven realisierten Merkmalsausprägungen x_{1i}, x_{2i}, x_{3i} ... der unabhängigen Variablen:

$$\hat{Y}_i = b_i e^{(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + ...)}$$

dabei ist b_i die für das Unterkollektiv i realisierte mittlere Beobachtungszeit.

5.3 Kontrollvariablen

In einem Kontrollgruppen-Untersuchungsdesign ist es erforderlich, alle Unterschiede hinsichtlich solcher Faktoren auszuschließen oder zu kontrollieren, die die Vergleichbarkeit von Behandlungs- und Kontrollgruppe stören könnten. Am Besten ist daher eine „Randomisierung", d. h. eine zufällige Zuweisung von Personen zur Behandlungs- oder Kontrollgruppe. Dies aber ist im vorliegenden Gegenstandsbereich grundsätzlich nicht möglich.

Es gibt eine Reihe von Faktoren, von denen die Wahrscheinlichkeit einer Entziehung oder einer VZR-Auffälligkeit abhängt. Dazu gehören soziodemografische Merkmale wie Ge-

---

[12] Das oft beobachtete Abflachen der Auffälligkeitsquote mit zunehmender Beobachtungszeit steht nicht im Widerspruch zur zeitunabhängigen Rate des Poisson-Prozesses, sondern leitet sich rein mathematisch daraus ab.

[13] siehe z.B. Muller, K.E., Nizam, A., Kleinbaum, D.G. & Kupper, L.L. (1998): Applied regression analysis and other multivariate methods. Daxburry Press, London, p. 687-704.

schlecht und Alter sowie regionale Besonderheiten des Verkehrs oder der Verkehrsüberwachung, wie sie sich insbesondere im Unterschied der Bundesländer oder bereits im Stadt-Land-Unterschied zeigen. Bei sehr langfristig angelegten Untersuchungen könnte auch das Jahr der Bewährungskontrolle eine Rolle spielen, sofern angenommen werden muss, dass sich die Verkehrs- oder Überwachungsintensität über die Zeit wesentlich verändert. Sofern diese Faktoren nicht durch strenge Parallelisierung von Behandlungs- und Kontrollgruppe ausbalanciert sind, müssen sie rechnerisch kontrolliert werden. Dazu sind die zu kontrollierenden Variablen als Kovariaten in die Regressionsrechnung aufzunehmen (als zusätzliche unabhängige Variablen x im Sinne des vorangegangenen Abschnitts). In diesem Fall spiegeln die Regressionskoeffizienten die Gruppenunterschiede wider, wie sie bestehen würden, wenn sich die Gruppen hinsichtlich der Kovariaten nicht unterscheiden würden (so genannte Adjustierung).

Als Kontrollvariablen sollten herangezogen werden:

- die zugrunde liegende Problematik als wesentliche Determinante der VZR-Auffälligkeit (zum Beispiel in der nach Abschnitt 2 vorgeschlagenen Differenzierung),
- Geschlecht und Alter der Person als wesentliche Merkmale des soziodemografischen „VZR-Risikos“,
- Bundesland und Regionstyp (nach Stadt-Land) des Wohnsitzes als wesentliche Merkmale der Verkehrsverhältnisse und der Überwachungsintensität in der Region, in der die Verkehrsteilnehmer ihre hauptsächliche Fahrleistung erbringen,
- ferner das Kalenderjahr der Datenerhebung als zeitabhängiges Merkmal der Verkehrsverhältnisse und der Überwachungsintensität.

Die Fahrleistung selbst, obwohl ein sehr wichtiger Faktor für die Verkehrsauffälligkeit, sollte nicht explizit berücksichtigt werden, weil eine Verkehrsauffälligkeit vor dem Gesetz nicht fahrleistungsbezogen, sondern zeitraumbezogen gewertet wird: Zu vergleichen und zu bewerten sind die Gefahren, die von zwei Gruppen in *gleichen Zeiträumen*, nicht auf gleichen Streckenlängen ausgehen[14].

5.4 Kontrollgruppenbildung

Die Gewinnung einer speziellen Kontrollgruppe für jeden Anbieter wäre zweifellos sehr aufwändig. Das vorliegende Konzept geht von der Grundidee aus, nicht eigens für jede Behandlungsgruppe eine Kontrollgruppe zu ziehen, sondern einen gemeinsamen Datenpool aus Daten des KBA zu nutzen, der durch statistische *Adjustierung* (s.o.) an das Klientel des jeweiligen Anbieters angepasst wird. Um die Vergleichbarkeit bei diesem Vorgehen zu gewährleisten, müssen aber die wesentlichen Stör- und Einflussfaktoren kontrolliert werden (siehe 5.3).

Referenzwerte sollen - dies ist ja ihr Zweck - vergleichbar sein. Die oben beschriebene Auswertungsmethode der Regression (Cox-Regression oder Poisson-Regression) erlaubt durch Berücksichtigung der Kontrollvariablen die Berechnung *adjustierter* Kennwerte. So ist es rein rechnerisch möglich, für jeden Anbieter einen speziell auf seine spezifische Klientenstruktur adjustierten Referenzwert zu bilden. So kann auch für Anbieter mit ungewöhnlich hohem Anteil an Klienten einer bestimmten Geschlechts- und Altersgruppe mit Wohnsitz in einem bestimmten Regionstyp ein Referenzwert ermittelt werden.

---

[14] Es gibt vor dem Gesetz keinen Rabatt für Vielfahrer.

Die folgende Betrachtung dient dazu, das hier vertretene Rationale[15] einer Kontrollgruppenbildung verständlich zu machen: Nach ihrer Kraftfahreignung lassen sich, zumindest ideell, sechs Kollektive unterscheiden, nämlich die Gruppe von Personen,

1. die nach den Ergebnissen eines idealen Eignungstests keinerlei Auffälligkeiten im Verkehr erwarten lassen würden,
2. die man nach den Ergebnissen eines idealen Eignungstests als ausreichend geeignet bezeichnen müsste, obwohl eine geringe, aber deutlich von null verschiedene Wahrscheinlichkeit dafür besteht, dass sie verkehrsauffällig werden oder sogar die Fahrerlaubnis einbüßen,
3. bei denen man im Zuge der Ersterteilung einer Fahrerlaubnis nach Vorliegen aller Voraussetzungen pragmatisch von einer Eignungsvermutung ausgeht, wohl wissend, dass eine gewisse Wahrscheinlichkeit besteht, dass sie verkehrsauffällig werden oder sogar die Fahrerlaubnis einbüßen,
4. für die Eignungszweifel vorliegen, insbesondere auch Eintragungen mit insgesamt 14 bis 17 Punkten im VZR, die aber noch rechtmäßig im Besitz einer gültigen Fahrerlaubnis sind,
5. denen nach einer vorangegangenen Entziehung (Aberkennung, Verzicht) eine Fahrerlaubnis neu erteilt wurde (weil sie ihnen nicht hinreichend begründet verweigert werden konnte),
6. denen die Fahrerlaubnis rechtskräftig entzogen wurde.

Die Gruppen 1 und 2 sind nur genannt worden, um das gesamte Spektrum der Möglichkeiten bis hin zum Ideal darzustellen. Die Gruppe 2 soll dabei deutlich machen, dass ein realistisches Menschenbild, das der Fehlbarkeit des Menschen Rechnung trägt, ein gewisses Restrisiko des Versagens *als gesellschaftliche Last* hinnehmen muss. Die Gruppe 3 hebt sich hiervon noch weiter ab, da kein idealer Eignungstest verfügbar ist und man aus pragmatischen Gründen auf eine Eignungsvermutung[16] angewiesen ist. Die Gruppe 5 besitzt eine deutlich erhöhtes Risiko, erneut verkehrsauffällig zu werden oder gar die Fahrerlaubnis einzubüßen[17]. Dabei ist das Risiko in Gruppe 5 noch wesentlich größer ist als das in Gruppe 4.

Wir können von der Gruppe 1 mit perfekter Eignung bis zur Gruppe 6 mit schweren Eignungsdefiziten von einer graduellen Abstufung der Kraftfahreignung sprechen. Für die weitere Argumentation festzuhalten ist, dass die Gruppe 5 die Grenze der gesellschaftlich gerade noch hinnehmbaren Eignungsmängel markiert. Eine mittlere Führerschein-Lebensdauer oder eine Verkehrsauffälligkeitsrate, die in einem „substanziellen Maße“ ungünstiger ausfällt als in dieser Gruppe, ist demnach nicht hinnehmbar.

Die in den ungünstigen Legalbewährungsdaten der Gruppe 5 zum Ausdruck kommende verminderte Eignung dieser Personen wird vom Gesetzgeber faktisch hingenommen, erkennbar am fehlenden Willen, das Verfahren einer Neuerteilung zu verschärft. An dieser Praxis, also der „normativen Kraft des Faktischen“, und nicht an strikten theoretischen Forderungen hat sich die Gewinnung von Vergleichswerten zu orientieren: Das vom Gesetzgeber bei Neuerteilung im Durchschnitt hingenommene abgesenkte Maß der Eignung, gemessen an den beiden Kriterien „Führerschein-Lebensdauer“ und „VZR-Auffälligkeitsrate“, ist der gesellschaftlich faktisch anerkannte Vergleichsmaßstab. Die Referenzwerte sind daher zu

---

[15] Es gibt sehr begrüßens- und verfolgenswerte weitergehende Vorschläge, z.B. von Jacobshagen in einem Arbeitspapier niedergelegt, die jedoch heute noch nicht realisierbare Anforderungen an die Datenlage (besonders auf Seiten des VZR) stellen.

[16] siehe auch Kroj, G. (Hrsg., 1995): Psychologisches Gutachten Kraftfahreignung. Deutscher Psychologen Verlag, Bonn, S.42

[17] Eine Auswertung einer Stichprobe der im Jahr 1996 im VZR eingegangenen Mitteilungen zu Verkehrsverstößen zeigt, dass der einzutragende Verstoß in 4,3 % der Fälle mit einer Entziehung der Fahrerlaubnis verbunden ist, sofern die Person bislang keine VZR-Eintragung besaß. Bei einer Vorbelastung mit 4 bis 8 Punkten beträgt die Häufigkeit einer Entziehung 4,9 % und steigt bei einer größeren Vorbelastung auf 5,5 %. Ging aber bereits eine Neuerteilung nach Entziehung voraus, so führen 10,3 % der einzutragenden Verstöße zu einer (erneuten) Entziehung.

bestimmen auf der Grundlage eines ausreichend großen Kollektivs von *unausgelesenen* Personen, denen nach einer Entziehung die Fahrerlaubnis neu erteilt wurde. Eine Differenzierung dieses Kollektivs nach den zur Wiedererlangung der Fahrberechtigung durchgeführten Maßnahmen ist nicht zulässig: Wer die Fahrberechtigung *rechtmäßig* erworben hat, aus welchen Voraussetzungen heraus und auf welchem Weg auch immer, besitzt zunächst – bis zum Beweis des Gegenteils – eine anerkannt ausreichende Fahreignung. Maßnahmen, die, gemessen an den beiden Evaluationskriterien, eine vergleichbare Fahreignung sicherstellen, müssen daher als ausreichend wirksam anerkannt werden.

Mit einer analogen Argumentation kann eine mittlere Führerschein-Lebensdauer oder eine Verkehrsauffälligkeitsrate, die in einem „substanziellen Maße“ ungünstiger als in der Gruppe 4 liegt, als Grenze für solche Maßnahmen gelten, die sich auf „Punkte-Täter“ spezialisieren.

Die Orientierung an empirisch gewonnenen Referenzwerten erfordert es, zumal unter den Bedingungen einer wandelbaren Praxis, diese Referenzwerte in regelmäßigen Abständen verlässlich zu ermitteln. Eine verbesserte Praxis wird so nach einiger Zeit verschärfte Referenzwerte nach sich ziehen, die wiederum dazu beitragen werden, die Praxis zu verbessern - ein Regelkreis, der im Sinne der Verkehrssicherheit durchaus erwünscht ist.

6 Auswertung

6.1 Datenerhebung, Datenaufbereitung für Behandlungsgruppe

Die Datenerhebung und -aufbereitung wie auch die statistische Auswertung werden aus Datenschutzgründen im KBA durchgeführt (sofern keine anderen Wege ermöglicht wurden). Zuständig ist eine Sondergruppe im Fachbereich Statistik unter qualifizierter wissenschaftlicher Leitung. Diese Gruppe arbeitet personell, räumlich und organisatorisch unabhängig von den Zentralen Registern. Sie hat außer einer standardisierten Datenschnittstelle keinen Zugang zu den Registern.

Das Verfahren der Datenerhebung und -verarbeitung gestaltet sich wegen der Tilgungsbestimmungen unterschiedlich, je nachdem, ob vor der zu evaluierenden Maßnahme die Fahrerlaubnis entzogen wurde oder nicht („Maßnahmen *im* Zusammenhang oder *ohne* Zusammenhang einer Entziehung“).

Verfahren bei Maßnahmen im Zusammenhang mit einer Entziehung: Am Ende der für die Evaluation festgelegten Rekrutierungsphase (wegen der Tilgung spätestens aber 10 Jahre nach dem Ereignis, das zur Entziehung, Aberkennung oder dem Verzicht führte) meldet der Anbieter alle Teilnehmer, für die in dieser Zeit ein erfolgreicher Maßnahmenabschluss zu verzeichnen ist (s. Ausschlusskriterium a). Der erfolgreiche Abschluss liefert das Datum des Aufnahmeereignisses. Die VZR-Abfrage wird für die nach dem Datum des Aufnahmeereignisses zusammengefassten Subgruppen zu den mit dem Anbieter vereinbarten Zeitpunkten durchgeführt. Bei der Wahl der Abfragezeitpunkte sollte sicher gestellt werden, dass für keinen Teilnehmer einer Subgruppe mehr als 5 Jahre seit Aufnahmedatum verstrichen sind (wegen der Bestimmungen zur Tilgungshemmung bei einfachen Ordnungswidrigkeiten). Die aus dem Register zurück gemeldeten Eintragungen sind daraufhin zu prüfen, ob eine Neuerteilung vorliegt und ob das Ausschlusskriterium b erfüllt ist (Neuerteilung innerhalb einer Maximalfrist nach Maßnahmenabschluss).

Verfahren bei Maßnahmen ohne Zusammenhang mit einer Entziehung: Besteht kein Zusammenhang mit einer Entziehung, so kann hier nicht von einer weitreichenden Tilgungshemmung ausgegangen werden. Am Ende der für die Evaluation festgelegten Rekrutierungsphase (wegen der Tilgungsfristen spätestens aber zwei Jahre nach dem Beginn) meldet der Anbieter die Teilnehmer, für die in dieser Zeit ein erfolgreicher Maßnahmenabschluss zu verzeichnen ist (s. Ausschlusskriterium a). Der erfolgreiche Abschluss liefert das

Datum des Aufnahmeereignisses und ist in der Regel zugleich das Startdatum für die Beobachtungsphase.

Die weiteren Schritte der Datenverarbeitung: Die mit dem Anbieter vereinbarten Informationen (bzw. der Standard-Datenkranz der amtlichen VZR-Statistik) werden aus dem Ergebnis der VZR-Abfrage übernommen, d. h., im Falle digitaler Speicherung aus dem Datensatz herausselektiert oder im Falle der herkömmlichen Speicherung in Aktenform nach Fotokopieren der Unterlagen durch Kodierer mit Hilfe einer Eingabemaske auf elektronische Datenträger gebracht. Anschließend werden die Angaben einer weitgehend computergestützten Plausibilitätsüberprüfung unterzogen und nötigenfalls manuell am Bildschirm nachbearbeitet.

Die weitere Verarbeitung umfasst die folgenden Punkte (stichwortartig):

1. Bestimmung von Beobachtungsbeginn und Beobachtungsende für jeden individuellen Fall, dazu das etwaige Auftreten eines terminierenden Ereignisses, hier die – auch nur vorläufige – Einbuße der Fahrerlaubnis, wodurch das Beobachtungsende vorverlegt wird; Berechnung der individuellen Beobachtungsspanne.
2. Anwendung der Ausschlusskriterien aus Abschnitt 5.1.3 für alle Personen und bei den verbliebenen Personen für alle VZR-Eintragungen.
3. Zählung der VZR-Auffälligkeiten und Summation der Beobachtungszeiten pro Untersuchungsgruppe.
4. Setzen eines „Statusmerkmals“ von null auf eins im Falle eines terminierenden Ereignisses vor regulärem Beobachtungsende laut Punkt 1.

6.2 Gewinnung von Referenzdaten

Verfahren für Maßnahmen im Zusammenhang mit einer Entziehung: Für den Aufbau eines Referenzdatensatzes wird folgendes Verfahren vorgeschlagen: Aus dem Zugang zum VZR wird durch das KBA laufend eine geeignete Stichprobe von Personen gewonnen, denen nach einer Entziehung (Aberkennung oder Verzicht) - aus welchen Gründen auch immer (Angaben dazu liegen nicht vor) - eine Fahrerlaubnis neu erteilt wurde. Diese Personen verbleiben für die Dauer der anschließenden Beobachtungsphase in einer für Forschungszwecke eigens eingerichteten Datenbank. Die Auswertung benötigt keine Einwilligung der Betroffenen, da keine Einzeldaten, sondern nur statistische Kennwerte ermittelt und veröffentlicht werden.

Das Gesamt der Neuerteilungen umfasst als kleine und in der Regel völlig vernachlässigbare Teilmenge notwendigerweise auch die zu evaluierenden Fälle. Der verzerrende Effekt, der damit in Kauf zu nehmen ist, arbeitet tendenziell für die Anbieter, denn er senkt die Wahrscheinlichkeit, eine etwaige Verschlechterung gegenüber dem Referenzwert festzustellen.

Die Personen der Kontrollgruppe müssen genauso wie die Personen der Behandlungsgruppen regelmäßig auf etwaige Eintragungen im VZR abgefragt werden, um eine Wiederauffälligkeit oder einen Fahrerlaubnisentzug feststellen zu können. Daneben wird die Dauer der jeweiligen Beobachtungszeit als Kovariate verarbeitet (gemäß Empfehlungen in Abschnitt 5.1.3 auf 60 Monate eingeschränkt). Im Fall einer Entziehung endet die Beobachtungszeit mit dem Datum der vorläufigen Entziehung.

Exkurs: Für das Modell der Stichprobenziehung kommt zunächst eine repräsentative Zufallsauswahl in Frage. Solche Stichproben können jedoch unter bestimmten Umständen einen Nachteil haben: Faktoren mit seltenen Merkmalsausprägungen sind entsprechend schwach repräsentiert. Dies wird zum gravierenden Problem, wenn eine Behandlungsgruppe in hohem Ausmaß von solchen Faktoren bestimmt ist. Beispielsweise wird eine für die Bundesrepublik repräsentative Stichprobe (selbst bei nicht unerheblichem Umfang) nur wenige Fälle weiblicher Alkohol-Wiederholungstäter umfassen – zuwenig, um für Anbieter, die sich dieser Klientel besonders verschrieben haben, Referenzwerte von annähernd ähnlicher Verlässlichkeit zu erstellen wie für Anbieter, die nur männliche Klienten haben. Ähnliche Probleme sind in der Bundesstatistik lange bekannt und führten zur Empfehlung

einer nach Gruppen geschichteten Auswahl mit disproportionaler Aufteilung[18]. Weibliche Personen würden nach diesem Verfahren überproportional in die Stichprobe aufgenommen werden. Bei der Stichprobengewinnung zur Erstellung von Referenzwerten könnte von solchen Methoden zur Erhöhung der Präzision Gebrauch gemacht werden.

Verfahren für Maßnahmen ohne Zusammenhang mit einer Entziehung: Wegen derzeit noch nicht erkennbarem Bedarf hier noch nicht ausgeführt.

6.3 Kennwertberechnung

Für die Auswertung werden die Daten der Behandlungsgruppe und die der Kontrollgruppe zusammengeführt.

6.3.1 Auswertung zum primären Kriterium „Führerscheins-Lebensdauer"

Aus dem Statusmerkmal nach Punkt 4 der Datenverarbeitungsschritte und der Dauer bis zur Entziehung bzw. bis zum Beobachtungsende kann das primäre Evaluationskriterium, die „Führerscheins-Lebensdauer" abgeleitet werden.

Sterbetafel-Analyse: Dazu werden im Auswertungsmodell der Sterbetafel-Analyse diese Daten, nämlich die Zeitspanne und die Statusvariable, als Input-Variablen für die Prozedur „Survival" des Statistik-Auswertungsprogramms SPSS[19] verwendet. Eine weitere kategoriale Variable kennzeichnet die Untersuchungsgruppen (Behandlungs- und Kontrollgruppe), so dass das Programm eine Sterbetafel gesondert für jede Untersuchungsgruppe erzeugt.

Gesucht wird als deskriptive Statistik die Zeit, nach der 50 % der Personen keine Fahrerlaubnis mehr besitzen. Das Programm weist diese Zahl, sofern sie durch reale Beobachtungsfälle repräsentiert ist, in einer besonderen Zeile aus. Verwendbar ist zu Vergleichszwecken neben diesem Median auch das erste Quartil (der 25-%-Wert). Dieser Wert muss allerdings durch Interpolation aus der Ergebnistabelle selbst ermittelt werden.

SPSS führt einen signifikanzstatistischen Vergleich der „Überlebenskurven" der beiden Untersuchungsgruppen mit Hilfe eines Chi-Quadrat-Tests nach Gehan durch. Die Null-Hypothese ist dabei, dass alle beobachteten Unterschiede zwischen dieses Gruppen lediglich auf Stichprobenfehlern beruhen.

Cox-Regression: Eine anspruchsvollere und leistungsfähigere, wenn auch weniger anschauliche Auswertungsmethode bietet die so genannte Cox-Regression, die ebenfalls mit SPSS durchgeführt werden kann. Diese Methode erlaubt es, zusätzlich eine beliebige Zahl von Kontrollvariablen zu berücksichtigen und bietet damit den Vorteil der Adjustierung und einer schärferen Testung der Hypothese auf Unterschiede zwischen den Untersuchungsgruppen.

Dazu sind die zu kontrollierenden Variablen als unabhängige Variablen in die Regressionsrechnung aufzunehmen. In diesem Fall spiegeln die Regressionskoeffizienten die Gruppenunterschiede wider, wie sie bestehen würden, wenn sich die Gruppen hinsichtlich der Kontrollvariablen nicht unterscheiden würden (Adjustierung). So werden Untersuchungsgruppen, die sich in der Zusammensetzung nach den Kontrollvariablen unterscheiden, direkt vergleichbar. Ähnlich ist es zu Vergleichszwecken möglich, auf eine bestimmte Altersverteilung (z.B. eine Durchschnittsbevölkerung) abgestellte Kennwerte zu berechnen (altersstandardisierte Ergebnisse).

---

[18] siehe z.B. BÖLTKEN, F. (1976): Auswahlverfahren, Teubner, Stuttgart, 1. Aufl. S. 262ff. Beschrieben als „Prinzip der vergleichbaren Präzision" der amtlichen Statistik in KRUG, W., NOURNEY, M. & SCHMIDT, J. (1994): Wirtschafts- und Sozialstatistik, Oldenbourg, München, 4. Aufl. S. 118-122

[19] NORUSIS, M.J. (1994): SPSS Advanced Statistics 6.1. Chicago.

6.3.2 Auswertung zum sekundären Kriterium „VZR-Auffälligkeitsrate“

Aus den Daten von Punkt 3 der Datenverarbeitungsschritte wird die Ereignisrate als Quotient aus der Zahl der VZR-Auffälligkeiten und der Summe der Beobachtungszeiten für jede Untersuchungsgruppe berechnet. Damit liegt das sekundäre Evaluationskriterium, die VZR-Auffälligkeitsrate pro Jahr, für deskriptive Zwecke vor.

Um Unterschiede zwischen Behandlungs- und Kontrollgruppe zufallskritisch abzusichern, kann eine so genannte Poisson-Regression[20] durchgeführt werden. Neben der abhängigen Variable, die Zahl der VZR-Eintragungen pro Person, und der unabhängigen Variable des Treatments, die Zugehörigkeit zu Behandlungs- oder Kontrollgruppe, kann als so genannte Kovariate auch die individuelle Dauer der Beobachtungsspanne berücksichtigt werden. Das Regressionsverfahren ermittelt für die Behandlungsgruppe (im Vergleich zur Kontrollgruppe) einen Regressionskoeffizienten, der signifikanzstatistisch auf Abweichung von null geprüft werden kann. Bei Signifikanz ist von der Ungleichheit der Auffälligkeitsraten in den verglichenen Gruppen auszugehen.

Die Größe des Einflusses einer Behandlung lässt sich anhand des relativen Risikos RR, hier dem Verhältnis des Auffallensrate in der Behandlungsgruppe zu der in der Kontrollgruppe, mit Hilfe der Exponentialfunktion aus dem Regressionskoeffizienten β einfach bestimmen: $RR = \exp(\beta)$. Das relative Risiko einer Behandlungsgruppe von beispielsweise RR = 0,90 bedeutet eine gegenüber der Vergleichsgruppe um 10 % verminderte Ereignisrate. Dieser Fall würde bei einem β von -0,105 auftreten (denn $e^{-0,105} = 0,90$).

Die Methode erlaubt es, zusätzlich eine Zahl von Kontrollvariablen zu berücksichtigen und bietet damit neben der Adjustierung des Behandlungseffekts eine schärfere Testung der Hypothese auf Unterschiede zwischen den Untersuchungsgruppen (s. Abschnitt 5.2). Dazu sind die zu kontrollierenden Variablen als zusätzliche unabhängige Variablen in die Regressionsrechnung aufzunehmen. In diesem Fall spiegelt der Regressionskoeffizient für den Behandlungseffekt die Gruppenunterschiede wider, wie sie bestehen würden, wenn sich die Gruppen hinsichtlich der Kontrollvariablen nicht unterscheiden würden (Adjustierung). So werden Untersuchungsgruppen, die sich in der Zusammensetzung nach den Kontrollvariablen stark unterscheiden, direkt vergleichbar.

Um einen Vergleich der Behandlungsdaten des jeweiligen Anbieters mit anderen Werten (z.B. der Literatur) zu ermöglichen, können mit Hilfe von Regressionsschätzformeln Werte für die VZR-Auffälligkeitsrate ermittelt werden, die zu erwarten gewesen wären, hätte der Anbieter ein „durchschnittliches“ Behandlungskollektiv. Dieses Vorgehen ist als Verfahren der Normierung bekannt[21].

7 Signifikanzstatistische Schlüsse

Der Vergleich der nach den Evaluationskriterien gewonnenen statistischen Kennwerte des Behandlungserfolgs mit den Referenzwerten soll zu einem aussagefähigen Schluss führen. Dieser Schluss muss daher gegen Zufallsergebnisse aufgrund von Stichprobenfehlern abgesichert sein.

7.1 Testlogik

Nach Abschnitt 3 ist vom Evaluationserfolg dann zu sprechen

1. gemäß hartem Erfolgskriterium, wenn die Evaluationskriterien in der Behandlungsgruppe signifikant positiver ausfallen, als es die Referenzwerte fordern,

---

[20] siehe Fußnote 13

[21] Es dient z.B. in der Epidemiologie in der Form der „Altersnormierung“ dem Vergleich zwischen Ländern mit verschiedenem Bevölkerungsaufbau.

2. gemäß weichem Erfolgskriterium, wenn die Evaluationskriterien in der Behandlungsgruppe die Referenzwerte nicht signifikant um mehr als einen „substanziellen“ Betrag ins Negative unterschreiten.

Der erste Fall ist einfach zu behandeln: Man testet den der Behandlung zugeordneten Regressionskoeffzienten auf einem üblichen Signifikanzniveau (Irrtumswahrscheinlichkeit) von 0,05 oder 0,01 auf Abweichung von null. Da die Hypothese gerichtet ist – es wird ja eine positive Abweichung von der Kontrollgruppe erwartet –, wird ein einseitiger Signifikanztest durchgeführt. Ist Signifikanz erreicht, so kann die Behandlung als erfolgreich gelten. Im Falle, dass die Abweichung in negativer Richtung ausfällt, erübrigt sich ein Signifikanztest, da die Nullhypothese, die Behandlung sei unwirksam, dann ohnehin nicht zurückgewiesen werden kann. Ein Erfolg gilt in diesem Fall als nicht nachgewiesen.

Der zweite Fall ist komplizierter, da hierbei eine teststatistische Zusatzforderung[22] erfüllt sein muss: eine ausreichende Teststärke („test power“). Will man, wie hier, aus einer Nicht-Signifikanz Schlüsse ziehen, so muss der statistische Test von seinem Design (und seiner Stichprobe) her empfindlich genug sein, einen gewissen, vorher als „substanziell“ definierten Verschlechterungseffekt mit hinreichender Wahrscheinlichkeit aufdecken zu können. Ein Test, der wegen mangelnder Teststärke nur mit beispielsweise 60-prozentiger Wahrscheinlichkeit einen gewissen Verschlechterungseffekt als signifikant nachweisen könnte, würde im Falle einer Nicht-Signifikanz keinen überzeugenden Beweis dafür liefern, dass keine substanzielle Verschlechterung existiert (da er in durchschnittlich 40 % der Anwendungsfälle versagen, d. h. einen wirklich vorhandenen Effekt fälschlich als nicht signifikant abtäte).

Diese Fehlermöglichkeit, nämlich einen wirklich vorhandenen substanziellen Effekt nicht zu erkennen, nennt man Fehler 2. Art oder Beta-Fehler (im Gegensatz zum Alpha-Fehler, einen nicht vorhandenen Effekt fälschlich als signifikant zu bezeichnen). Entsprechend hat man in Studien, in denen man aus der Nicht-Signifikanz den Schluss auf die Ungültigkeit der Alternativhypothese ziehen möchte, den Beta-Fehler zu kontrollieren und auf üblicherweise 0,05, 0,01 oder gar 0,001 zu begrenzen. Eine Schwierigkeit besteht gewöhnlich darin, einen Kompromiss zwischen der Forderung, bereits kleinste Verschlechterungen gegenüber dem Referenzwert aufdecken zu können, und der Forderung nach einer organisatorisch und finanziell noch vertretbaren Stichprobengröße zu finden.

Vorgeschlagen wurde hier (im Abschnitt 3), von den Anbietern zu fordern, dass – auch für schwierigstes Klientel – der Referenzwert mit großer Sicherheit um nicht mehr als im Mittel 33 % ins Negative unterschritten werden darf. Wir haben den „substanziellen Effekt“ damit auf eine Verschlechterung um 33 % angesetzt und die Forderung erhoben, mit großer Sicherheit eine solche oder gar größere Verschlechterung aufzudecken. Unter „großer Sicherheit“ soll eine Irrtumswahrscheinlichkeit im Sinne des Beta-Fehlers von 0,01 bei einseitigem Test verstanden werden. Die einseitig formulierte Alternativhypothese lautet im vorliegenden Fall: „Die Behandlungsgruppe ist um einen ‚substanziellen Effekt‘ von 33 Prozent (oder mehr) schlechter als die Kontrollgruppe“. Die Nullhypothese lautet: „Die Behandlungsgruppe ist im Vergleich zur Kontrollgruppe genau so gut, wenn nicht besser, jedoch nicht um mehr als 33 % schlechter“.

Wird ein statistischer Test auf Abweichung zwischen Behandlungs- und Kontrollgruppe von mindestens 33 % *nicht* signifikant, so bedeutet dies bei einem Beta-Fehler von 0,01, dass mit mindestens 99-prozentiger Sicherheit keine „substanzielle“ Verschlechterung im obigen Sinne besteht. Dies sollte ausreichen, dem Anbieter – nach weichem Erfolgskriterium – die Unbedenklichkeit für die Maßnahme zu bescheinigen. Tritt allerdings eine Signifikanz auf, so muss die Maßnahme als ungeeignet gelten.

---

[22] Würde man auf diese Forderung verzichten, so könnte man eine Signifikanz in negativer Richtung leicht dadurch verhindern, indem man eine derart kleine Stichprobe wählt, dass im Falle selbst eines dramatisch schlechten Behandlungseffekts die Signifikanzgrenze unerreichbar ist.

7.2 Stichprobenumfang

Hartes Evaluationskriterium: Bei festgelegtem Alpha-Fehler gilt, dass eine größere Stichprobe die Teststärke steigert, d. h. den Beta-Fehler (einen vorhandenen positiven Behandlungseffekt nicht zu entdecken) senkt. Der Beta-Fehler gibt hier unmittelbar das Risiko des Anbieters an, ein für ihn nicht akzeptables, nämlich nicht signifikantes Ergebnis zu erzielen. Es ist also in seinem eigenen Interesse, eine möglichst große Behandlungsstichprobe, gemessen an Personenjahren, in die Evaluation einzubringen. Für ein Evaluationshandbuch sollte eine Tabelle erstellt werden mit den bei einseitigem Test nötigen Stichprobenumfängen bei einem Alpha von 0,05, Beta-Werten von 0,10, 0,05, 0,01 und 0,001 sowie als realistisch angenommenen Effektstärken, d. h. den erwartbaren Verbesserungen der Behandlungsgruppe gegenüber der Kontrollgruppe, zwischen 10 und 100 Prozent.

Schon jetzt ist abschätzbar, dass einige hundert Personenjahre in der Behandlungsgruppe für diese statistische Fragestellung wünschenswert sind.

Weiches Evaluationskriterium: Um den Beta-Fehler auf den geforderten Wert einzuschränken, gibt es zwei Möglichkeiten: Man kann den Alpha-Fehler von üblicherweise 0,05 oder 0,01 auf 0,10 oder höher anheben oder die Stichprobe ausreichend vergrößern. Einer starken Anhebung des Alpha-Fehlers wird der Anbieter nicht zustimmen können, weil dies sein Risiko vergrößert, fälschlich eine substanzielle Verschlechterung als signifikant attestiert zu bekommen. Damit bleibt nur die Möglichkeit den Stichprobenumfang ausreichend festzulegen. Hierfür sollten in einem Evaluationshandbuch Tabellen für verschiedene Alpha-Niveaus erstellt werden, die auf Seiten der Kontrollgruppe von einem Beobachtungsumfang zwischen 4.000 und 12.000 Personenjahren ausgehen.

Auf Seiten der Behandlungsgruppe werden nach erster Abschätzung Stichproben in der Größenordnung von 1000 Personenjahre erforderlich sein.

8 Zusammenfassung

Vorgestellt wird auf der Grundlage einer Differenzierung der Problemfelder „Fahranfänger“, „Punktetäter“ und „Alkoholtäter“ ein Konzept zur statistischen Outcome-Evaluation von Maßnahmen des „Driver Improvements“ anhand von Legalbewährungsdaten aus dem VZR (FLEML), dazu eine Reihe von Empfehlungen zur Durchführung. Als primäres bzw. als sekundäres Evaluationskriterium dienen die Merkmale „Entziehung der Fahrerlaubnis“ und „VZR-Auffälligkeit“. Für bisherige Probleme, nämlich Meldeverzug, Tilgungszeiten und unvollständige Beobachtungszeiten, werden Lösungen besprochen.

Geschildert werden Möglichkeiten des Beobachtungsdesigns wie anschließende und abgerückte Beobachtung, retrospektiver und prospektiver Untersuchungsauftrag, einmalige und mehrmalige VZR-Abfrage, ferner wichtige Ein- und Ausschlusskriterien für Personen und Eintragungen. Näher ausgeführt werden Art und Bedeutung der empirisch zu ermittelnden Kenngrößen „Führerschein-Lebensdauer“ und „VZR-Auffälligkeitsrate“ wie auch die dahinter stehenden statistischen Verhaltensmodelle. Ein breiter Raum wird eingeräumt für die Darstellung von Kontrollgruppenbildung und Kontrollvariablen.

Danach werden die Möglichkeiten und Wege der Datenerhebung und Datenaufbereitung für die Behandlungsgruppe sowie der Gewinnung der Referenzdaten dargelegt. Schließlich wird die Auswertung beschrieben: die Kennwertberechnung und ihre zufallskritische Absicherung durch Signifikanztests einschließlich der dazu gehörigen Testlogik und Überlegungen zum nötigen Stichprobenumfang.