U. Grömping, U. Weimann
Ford Werke AG, Köln, Germany

S. Menzler
Tierärztliche Hochschule Hannover, Institut für Biometrie, Epidemiologie und Informationsverarbeitung, Hannover, Germany

Split-Register Study: A New Method for Estimating the Impact of Rare Exposure on Population Accident Risk Based on Accident Register Data

Abstract

The data situation for quantifying the proportion of accidents avoided by the introduction of active safety systems is incomplete, since there is generally no data available on the accidents avoided by the technology in question. In this paper, a split-register approach is suggested and compared with the classical case-control approach known from epidemiologic applications. Provided a set of assumptions hold, which can reasonably be made in such data situations, the split register approach allows inferences on the population accident risk. For both approaches the benefits of basing the analysis on the results of a logistic regression to adjust for confounding factors are outlined. The biasing effects of violating key assumptions are discussed and the split-register approach is demonstrated using the example of the active safety system ESP with data from the German in-depth accident study GIDAS.

Notation

A accident register vehicles

D accident type that is suspected to be preventable by technology, e.g. loss of stability accidents

$\overline{D}$ other accidents (e.g. accidents where loss of stability played no role)

E vehicles equipped with active safety technology of interest, e.g. ESP

$\overline{E}$ vehicles without active safety technology of interest, e.g. without ESP

x covariates

P(D\E) probability of event D given event E

OR odds ratio

RR relative risk

Introduction

For investigations with the aim of quantifying the proportion of accidents avoided by active safety technologies like ESP, the researcher is confronted with an incomplete data situation. Since there is generally no data available on avoided accidents, conclusions usually have to be drawn from accident register, and possibly additionally population census data.

One way to deal with this data situation is to use the case-control approach, which is well known from epidemiology, where it was initially introduced for investigating rare diseases with long time of onset. In accident research concerned with quantifying the proportion of accidents preventable by introducing or promoting a certain active safety technology, there is a further difficulty. In addition to the accidents of interest possibly but not necessarily being rare, the exposure, which in this case would be a hopefully beneficial active safety technology, is often rare as well at the time of its investigation as a new technology. In this paper, an alternative approach called split-register is suggested for accident research based on accident registers (=accident databases), which is better suited for situations with rare exposures. Furthermore, it allows inferences on the population accident risk, which is not directly possible with the classical case-control approach.

In both approaches, logistic regression can be applied to adjust for confounding from third factors. The benefits and reasoning of this approach are briefly discussed in this paper.

The effect of violation of assumptions on the estimates is discussed and finally the split-register approach is demonstrated on the example of the active safety technology ESP with data from the German in-depth accident study GIDAS.

Analysis Approaches

Data Situation

Consider the population “all vehicles”, which are all vehicles registered in a certain region in a certain

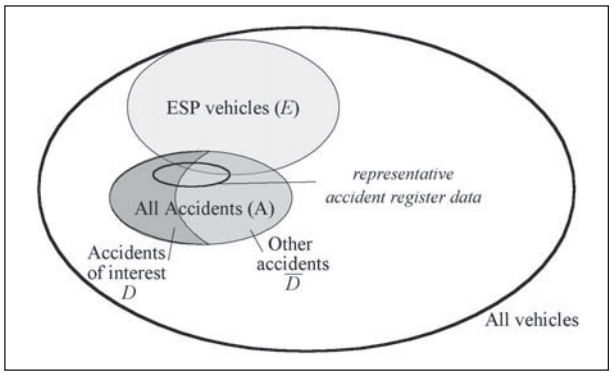

Fig. 1: Graphical illustration of example data situation

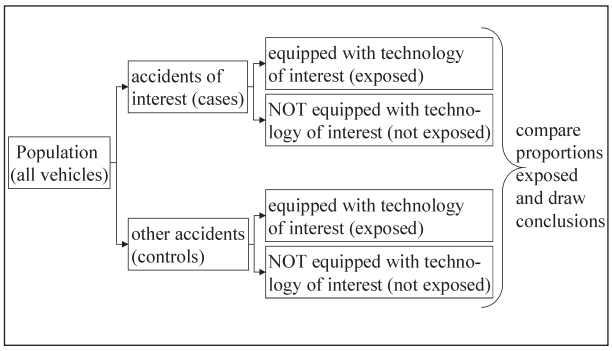

Fig. 2: Schematic of case control approach

time frame. In this population of all vehicles, there is a subset of those vehicles E, which are equipped with the active safety technology of interest, e.g. ESP. The negation of an event is denoted by a bar, so all vehicles not equipped with the technology of interest are denoted by $\bar{E}$. Another subset in the population of all vehicles are those vehicles that have been involved in an accident and are registered in the accident register – this is not equivalent to the subset of all vehicles involved in an accident, because not all accidents are registered. The vehicles registered in the accident register can be partitioned into those vehicles D, which were involved in the accident type that is assumed to be preventable by the technology of interest, for example loss-of-stability accidents in case of ESP, and those which were involved in other accidents $\bar{D}$. In the following, the accidents which are assumed to be preventable by the technology of interest will be referred to as the accidents of interest (figure 1).

The Case Control Approach

In accident analysis, case control studies are studies in which vehicles which were involved in accidents of interest (cases) are compared with other vehicles (controls). The idea behind including controls in the analysis is that they can be used to estimate the exposure rate to be expected, if no association between equipment with the technology and the rate of accidents of interest were present.

The selection of cases and controls has to be independent of the exposure status, which in this case is whether the vehicle is equipped with the active safety technology of interest, see figure 2. For selecting the cases, the definition what constitutes an accident of interest has to be clearly understood. If not all the cases available are selected for the study, then the study cases should be selected so that they are representative of all cases.

The controls should come from the same population at risk for the accidents of interest as the cases. They should be representative of the target population, or appropriate weighting factors for later analysis should be known. In order to avoid so-called confounding bias, sometimes the controls are chosen so that they match the cases with respect to other variables called covariates which might influence the rate of accidents of interest. This approach leads to less vehicles being included in the study and therefore less confidence in the results. If not all confounding variables are considered in matching, or if no matching has been carried out, a logistic regression can be applied prior to further analysis of results, to adjust for covariates.

The criteria for selecting cases in the context of this paper depend on the definition of what constitutes an accident of interest. In the ESP example the accidents of interest would be all loss of stability accidents and a clear definition of how accidents are classified into loss of stability accidents and other accidents is necessary. For the controls, there are several possibilities of setting selection rules. However, selection into the control group must not depend on presence or absence of the exposure.

There are case-control studies in accident research, in which the controls are recruited from non-accident vehicles, e.g. by randomly selecting license holders who live in the same geographical area as the cases [1]. Many case-control studies are register-based, however. They use certain accident types as cases, other accident types as controls [2, 3], in the example this could for example be all accidents D or a subset of D.

Because there is no absolute information on prevented accidents, only a relative estimate of the proportion of prevented accidents can be made. In the case control approach, the odds ratio can be used to do this.

With the odds of having an accident of interest given the exposure and covariates

$$odds_D \mid E = \frac{P(D \mid E, \mathbf{x})}{1 - P(D \mid E, \mathbf{x})} \tag{1}$$

and the odds of having an accident of interest given that the vehicle was not equipped with the technology of interest and covariates

$$odds_D \mid \overline{E} = \frac{P(D \mid \overline{E}, \mathbf{x})}{1 - P(D \mid \overline{E}, \mathbf{x})} \tag{2}$$

the odds ratio is given by

$$OR_D = \frac{odds_D|E}{odds_D|\overline{E}} = \frac{\mathrm{P}(D|E, \mathbf{x})/(1 - \mathrm{P}(D|E, \mathbf{x}))}{\mathrm{P}(D|\overline{E}, \mathbf{x})/(1 - \mathrm{P}(D|\overline{E}, \mathbf{x}))}. \tag{3}$$

The odds ratio is the factor, by which the odds of having an accident of interest with an “unexposed vehicle“ is multiplied to find the odds for an “exposed“ vehicle. In this context, exposure means being equipped with the active safety technology of interest.

If the exposure has no influence on the odds of having an accident of interest, the odds ratio is equal to one. If the proportion of accidents of interest is smaller for vehicles equipped with the technology of interest than for other vehicles, which indicates the desired preventative effect of the technology, then the odds ratio is smaller than one. If this proportion is larger for the “exposed” than the “unexposed” vehicles, then a harmful effect of the technology is indicated and the odds ratio is larger than one.

The main advantage of the case control approach in the case of accident register data is that the analysis can be performed without the need to collect further data. Disadvantages of the case-control approach include the high bias risks, in particular selection bias for controls and confounding bias, and that only the odds ratio, but not the absolute reduction in risk can be estimated.

The Split-Register Approach

As mentioned above, one possibility to select cases and controls for the case control approach is to use all vehicles from D as cases, and all vehicles from $\overline{D}$ as controls. In the split-register approach, the same groups are determined.

In this case, the “cases” and the “controls” put together make up the whole population of accident register vehicles. This is very useful, if this accident register can be assumed to be representative for all accidents in a population of vehicles.

With the further assumption, that given the covariate information x, the probability of causing accidents of the type $\overline{D}$ (the control group or “other accidents“) does not depend on the fact whether the vehicle is equipped with the active safety technology of interest,

$$\mathrm{P}(\overline{D} \mid E, \mathbf{x}) = \mathrm{P}(\overline{D} \mid \overline{E}, \mathbf{x}) = \mathrm{P}(\overline{D} \mid \mathbf{x}), \tag{4}$$

the odds ratio from the accident database coincides with the relative risk in the population.

The relative risk of accidents of interest for vehicles with vs. vehicles without the active safety technology of interest in the population is given by

$$RR_D = \frac{\mathrm{P}(D \mid E, \mathbf{x})}{\mathrm{P}(D \mid \overline{E}, \mathbf{x})}. \tag{5}$$

It measures the risk of having an accident of interest in a vehicle equipped with the active safety technology of interest, compared to one not equipped with it, given covariates x.

Often (4) is a reasonable assumption to make. In the ESP example, the accidents of interest are loss of stability accidents. The group of other accidents includes for example parking accidents or accidents that do not involve any yaw motion of the vehicle at all. For those “other accidents“ it is reasonable to assume that equipping vehicles with ESP has no influence on the probability of their occurrence.

In the split-register approach, the analysis explicitly acknowledges that all information is conditional on the fact that an accident has happened. Thus, the probabilities P(D|E, x) and P(D|$\overline{E}$, x) cannot be directly estimated. Instead, the accident register yields estimates for the probabilities P(D|E, x, A) and P(D|$\overline{E}$, x, A), and in analogy to (3) an odds ratio $OR_D|A$ can be defined as

$$OR_D \mid A = \frac{\mathrm{P}(D \mid E, \mathbf{x}, A)/(1 - \mathrm{P}(D \mid E, \mathbf{x}, A))}{\mathrm{P}(D \mid \overline{E}, \mathbf{x}, A)/(1 - \mathrm{P}(D \mid \overline{E}, \mathbf{x}, A))}. \tag{6}$$

Even though this model is conditional on the fact that an accident has happened, it can be shown that the unconditional relative risk (5) of having an accident of interest is equal to the conditional odds ratio (6), see appendix A.

This is a very useful result, because the quantity one minus the relative risk can be interpreted as the proportion of accidents of interest among those vehicles not equipped with the active safety technology that can be prevented by equipping them with it.

The quantity of interest in most investigations is the avoidable proportion of all accidents rather than the avoidable proportion of the accidents of interest D only. It is given by

$$1 - RR_{A,\mathbf{x}} = \mathrm{P}(D \mid \overline{E}, \mathbf{x}, A)(1 - RR_D), \tag{7}$$

where the equality is shown in appendix A. In a split-register study, this quantity can be estimated for given combinations of covariates, provided assumption (4) holds. Logistic regression is a necessary first step to obtain this estimate.

An advantage of the split-register approach, as well as of the case control approach, is that application is simple and fast, if data are already available. Additionally, the relative risk and population risk can be estimated.

A disadvantage of the split-register approach, as well as of the case control approach is the high bias risk, in particular confounding bias which causes violation of the key assumption (4). Consequences of such a violation of assumptions is discussed further below.

Logistic Regression

Many researchers restrict attention to specific subsets of vehicles, in order to make exposed and unexposed vehicles (e.g. ESP and non-ESP) as comparable as possible. Consequently, only small numbers of vehicles are usable in the analysis.

Alternatively, all vehicles can be used. The important confounding information then has to be incorporated via a logistic regression. For example, in addition to ESP as the exposure of main interest, the power of the vehicle, vehicle equipment like ABS, power steering, the tire profile, weather and light conditions during the accident, the driver's age, gender, blood alcohol level can be adjusted for by carrying out a logistic regression.

Logistic regression is a standard technique implemented in statistics software packages like SAS, Minitab, SPSS or S-plus. It is a method for modeling probabilities of an event depending on other variables. For example, logistic regression can be applied for modeling the probability that an accident from A belongs to the accidents of interest D; as mentioned above, it would be desirable to incorporate information on exposure and further covariates, which can be done in logistic regression, similar to a linear regression approach. Since probabilities have to lie between 0 and 1, it is helpful to apply a transformation that allows using a linear function for modeling without running the risk of getting implausible values for the probabilities. The most commonly chosen transformation is the so-called logit-transformation that is the natural log of $p/(1–p)$ with p denoting the probability. The probabilities $P(D|\overline{E}, x, A)$ and $P(D|E, x, A)$ would for example be modeled by modeling their logits with a coefficient for the presence or absence of exposure and a linear function in x with coefficients to be estimated in order to achieve a good fit between data and model. It is a nice feature of logistic regression that the coefficients have an interpretation as logarithms of odds ratios. For further reference, see for example [6, 7].

Bias Investigation

In the section on the split register approach, it was shown, that the population accident risk reduction following from introduction of an active safety technology to all vehicles can be estimated. This only holds under the key assumption (4), that the probability of having an accident of type $\overline{D}$, a “control type“ accident, is not influenced by presence of the technology under investigation. In this section, consequences of violations of this key assumption are discussed.

The most natural violation of this assumption is that some accidents which belong to D were wrongly allocated to $\overline{D}$. In this case, the presence or absence of the active safety technology of interest has an influence on the probability of $\overline{D}$, which is a violation of assumption (4). For a preventive exposure (e.g. ESP), i.e. if E has a beneficial influence on the accidents wrongly allocated to $\overline{D}$, then this leads to a dilution of the estimated impact of the exposure, i.e. the beneficial effect of the technology under

investigation is underestimated. Thus, although it is very desirable to avoid diluting study results, a wrong allocation of some accident types does not invalidate a proven benefit of a new technology.

If assumption (4) is violated the other way round, i.e. if exposure increases the probability of $\bar{D}$, then the benefit of the technology under investigation is overestimated. This violation may for example arise, if more risk prone customers buy vehicles with the technology in question so that their risk proneness increases their risk of having an accident that cannot be prevented by the technology. This bias can be mitigated by including variables in the logistic regression model that may account for risk proneness of the drivers, for example power of the vehicle or gender of driver.

Example: Quantifing the Effectiveness of ESP to Prevent Loss of Stability Accidents Based on GIDAS Data

The example is based on GIDAS data from 1994 to the middle of 2003, which is assumed to be representative for the accident situation in the area of Germany, where the data is collected for that period of time. For further information on data collection methods in the GIDAS study, see http://www.gidas.org/. In order to ensure independence between the different vehicles in the analysis, only one vehicle per accident was included in the analysis. The population of accidents in the register is partitioned into loss of stability accidents D and other accidents $\bar{D}$. The exposure in this case is whether the vehicle was equipped with ESP (E) or not ($\bar{E}$). The definition, whether an accident was a loss of stability accident, was based on information independent of whether the vehicle was equipped with ESP. The classification of the accidents depended on the accident type, driving speed, road condition, and whether the driver stated he made an evasive manoeuvre. By this procedure, 48.47% of all accidents were classified as loss of stability accidents.

The total proportion of vehicles equipped with ESP were 2.72%, the first accidents involving vehicles equipped with ESP occurred in the year 2000.

A large number of variables that could possibly affect the probability of causing a loss of stability accident was considered in the logistic regression.

Logistic regression can only be carried out with vehicles for which there is a value for every variable that is to be included in the model. Only 1631 of the 6211 vehicles had complete information for all variables considered in logistic regression, and to only use those 1631 vehicles would have meant a considerable loss of information. Therefore a missing value imputation algorithm has been used to fill missing variables, see [4,5].

Twelve variables were found to have significant influence on the probability of causing a loss of stability accident and were included in the logistic regression model, so that they could be adjusted for when calculating the odds ratio.

The adjusted odds ratio is estimated to be

$$\text{est}(OR_D|A)=\text{est}(RR_D) = 0.5587,$$

i.e. $\text{est}(1–RR_D)=44.13\%$ is the proportion of loss of stability accidents among non-ESP vehicles that can be avoided by equipping all these vehicles with ESP. This result holds regardless of the covariate scenario.

Since the proportion of loss of stability accidents in all accidents crucially depends on the covariates, the covariates need to be taken into account for estimating the proportion of all accidents (A) that can be prevented. An overview of some scenarios is given in table 1.

	scenario 1	scenario 2	scenario 3	scenario 4	scenario 5
age of driver	18	38	38	38	60
passengers	yes	no	no	yes	no
tread depth	2	5	5	5	8
rain	yes	no	yes	no	no
day/night time	night	day	day	day	day
estimated accident avoidance potential of ESP installation	35.6%	21.3%	26.3%	19.0%	12.5%

Tab. 1: Estimated accident avoidance potential of ESP installation for several combinations of covariates

For a young driver with passengers in the car and worn tires with tread depth of only 2mm, on a rainy night, a vehicle equipped with ESP has an estimated accident avoidance potential of 35.6%. To the other extreme, for a vehicle equipped with tires with tread depth of 8mm, driven by a driver aged 60 without any passengers on a dry day, the estimated accident avoidance potential is only 12.5%. For an average age driver of 38 years, with passenger and tires with average tread depth of 5mm, on a dry day, the estimated accident avoidance potential of ESP is 19%.

Conclusions

Split-register-studies offer an interesting possibility of being able to estimate population relative risks in accident research. They are similar to case-control studies in many respects, especially the risk of confounding bias. They differ from case-control studies in that they are better for rare exposures like very new technologies and have a more sound line of mathematical arguments that allows direct conclusions on the population. They are a recommended strategy because of their simplicity and should be used in combination with logistic regression, in order to avoid confounding bias as much as possible.

References

1 C. TURNER, R. McCLURE: Quantifying the Role of Risk-Taking Behaviour in Causation of Serious Road Crash-Related Injury. Accident Analysis and Prevention 36 (3), 383–389, 2004

2 T. RUEDA-DOMINGO, P. LARDELLI-CLARET, J. D. LUNA-DEL-CASTILLO, J. J. JIMÉNEZ-MOLEÓN, M. GARCÍA-MARTÍN, A. BUENO-CAVANILLAS: The Influence of Passengers on the Risk of the Driver Causing a Car Collision in Spain– Analysis of Collisions from 1990 to 1999. Accident Analysis and Prevention 36 (3), 481–489, 2004

3 K. K. W. YAU: Risk Factors Affecting the Severity of Single Vehicle Traffic Accidents in Hong Kong. Accident Analysis and Prevention 36 (3), 333–340, 2004

4 S. OTTO: Quantifizierung des Einflusses aktiver Sicherheitssysteme auf die Unfallwahrscheinlichkeit und Identifikation von sicherheitsrelevanten Attributen basierend auf Realunfalldaten. Diplomarbeit, Fachbereich Statistik, Universität Dortmund, Germany, 2004

5 SAS Institute Inc. SAS/STAT® Software: Changes and Enhancements, Release 8.2. SAS Institute Inc., Cary, NC, 2001

6 A. J. DOBSON: Introduction to Generalized Linear Models, Second Edition, Chapman & Hall, 2001

7 D. W. HOSMER, S. LEMESHOW: Applied Logistic Regression, John Wiley & Sons Inc., 1989

Appendix A: Equality of Odds Ratio and Relative Risk in Split-Register Approach

Equation (4) means that the probability of causing accidents of the type $\overline{D}$, the control group or "other accidents", does not depend on the fact whether the vehicle is equipped with the active safety technology of interest, given the covariate information x.

If this key assumption (4) holds, then the odds ratio of accidents of interest for vehicles with vs. vehicles without ESP, given the vehicle has had an accident is equal to relative risk of accidents of interest for vehicles with vs. vehicles without ESP in the population.

Since D=D∩A, which can be appreciated by regarding figure 1,

$$\begin{aligned} P(D \mid E, \mathbf{x}) &= \frac{P(D, E, \mathbf{x})}{P(E, \mathbf{x})} \\ &= \frac{P(D, E, \mathbf{x}, A) P(A, E, \mathbf{x})}{P(E, \mathbf{x}) P(A, E, \mathbf{x})} \\ &= P(D \mid E, \mathbf{x}, A) P(A \mid E, \mathbf{x}) \end{aligned} \tag{8}$$

and analogously $\mathrm{P}(\overline{D} \mid E,\mathbf{x}) = \mathrm{P}(\overline{D} \mid E,\mathbf{x},A)\,\mathrm{P}(A \mid E,\mathbf{x})$, $\mathrm{P}(D \mid \overline{E},\mathbf{x}) = \mathrm{P}(D \mid \overline{E},\mathbf{x},A)\,\mathrm{P}(A \mid \overline{E},\mathbf{x})$, and $\mathrm{P}(\overline{D} \mid \overline{E},\mathbf{x}) = \mathrm{P}(\overline{D} \mid \overline{E},\mathbf{x},A)\,\mathrm{P}(A \mid \overline{E},\mathbf{x})$. With these results, it is straightforward to rewrite $OR_{D|A}$ in the desired way:

$$
\begin{aligned}
OR_D \mid A &= \frac{\mathrm{P}(D \mid E, \mathbf{x}, A)(1 - \mathrm{P}(D \mid \overline{E}, \mathbf{x}, A))}{\mathrm{P}(D \mid \overline{E}, \mathbf{x}, A)(1 - \mathrm{P}(D \mid E, \mathbf{x}, A))} \\
&= \frac{\mathrm{P}(D \mid E, \mathbf{x}, A)\,\mathrm{P}(\overline{D} \mid \overline{E}, \mathbf{x}, A)}{\mathrm{P}(D \mid \overline{E}, \mathbf{x}, A)\,\mathrm{P}(\overline{D} \mid E, \mathbf{x}, A)} \\
&= \frac{\mathrm{P}(D \mid E, \mathbf{x}, A)\,\mathrm{P}(\overline{D} \mid \overline{E}, \mathbf{x}, A) P(A \mid E, \mathbf{x}) P(A \mid \overline{E}, \mathbf{x})}{\mathrm{P}(D \mid \overline{E}, \mathbf{x}, A)\,\mathrm{P}(\overline{D} \mid E, \mathbf{x}, A) P(A \mid E, \mathbf{x}) P(A \mid \overline{E}, \mathbf{x})} \\
&= \frac{\mathrm{P}(D \mid E, \mathbf{x}) P(\overline{D} \mid \overline{E}, \mathbf{x})}{\mathrm{P}(D \mid \overline{E}, \mathbf{x}) P(\overline{D} \mid E, \mathbf{x})} \\
&= \frac{\mathrm{P}(D \mid E, \mathbf{x}) P(\overline{D} \mid \mathbf{x})}{\mathrm{P}(D \mid \overline{E}, \mathbf{x}) P(\overline{D} \mid \mathbf{x})} \\
&= \frac{\mathrm{P}(D \mid E, \mathbf{x})}{\mathrm{P}(D \mid \overline{E}, \mathbf{x})} \\
&= RR_D . \qquad (9)
\end{aligned}
$$

As a further step, the proportion of all accidents that can be prevented by introducing the active safety technology of interest can be obtained:

$$
\begin{aligned}
1 - RR_{A,\mathbf{x}} &= \frac{\mathrm{P}(A \mid \overline{E}, \mathbf{x}) - \mathrm{P}(A \mid E, \mathbf{x})}{\mathrm{P}(A \mid \overline{E}, \mathbf{x})} \\
&= \frac{\mathrm{P}(D \mid \overline{E}, \mathbf{x}) + \mathrm{P}(\overline{D} \mid \overline{E}, \mathbf{x}) - \mathrm{P}(D \mid E, \mathbf{x}) - \mathrm{P}(\overline{D} \mid E, \mathbf{x})}{\mathrm{P}(D \mid \overline{E}, \mathbf{x}) / \mathrm{P}(D \mid \overline{E}, \mathbf{x}, A)} \\
&= \frac{\mathrm{P}(D \mid \overline{E}, \mathbf{x}) + P(\overline{D} \mid x) - \mathrm{P}(D \mid E, \mathbf{x}) - \mathrm{P}(\overline{D} \mid \mathbf{x})}{\mathrm{P}(D \mid \overline{E}, \mathbf{x}) / \mathrm{P}(D \mid \overline{E}, \mathbf{x}, A)} \\
&= \frac{\mathrm{P}(D \mid \overline{E}, \mathbf{x}) - \mathrm{P}(D \mid E, \mathbf{x})}{\mathrm{P}(D \mid \overline{E}, \mathbf{x}) / \mathrm{P}(D \mid \overline{E}, \mathbf{x}, A)} \\
&= \mathrm{P}(D \mid \overline{E}, \mathbf{x}, A)(1 - RR_D) \qquad (10)
\end{aligned}
$$

The subscript x in (10) shows that the relative risk of all accidents depends on the settings of the covariates. Logistic regression offers an estimate for $\mathrm{P}(D \mid \overline{E}, x, A)$, and therefore an estimate for $RR_{A,x}$ can also be obtained.