Methods for the creation of injury risk functions based on real world accident data

Dipl.-Ing. Florian Spitzhüttl
Dipl.-Ing. Henrik Liers

Berlin, November 2016
Methods for the creation of injury risk functions based on real world accident data

Dipl.-Ing. Florian Spitzhüttl
Dipl.-Ing. Henrik Liers

Dresden, November 2016
I Contents

I Contents .................................................................................................................. I

II Symbols ................................................................................................................. III

III Abstract ................................................................................................................. V

1 Introduction ............................................................................................................. 1

2 Principles of injury risk functions ............................................................................. 2
  2.1 Definition ............................................................................................................. 2
  2.2 Dependent variable ............................................................................................. 2
  2.3 Independent variables ......................................................................................... 3
  2.4 Mathematical function ......................................................................................... 4
  2.5 Application .......................................................................................................... 6
  2.6 Challenges .......................................................................................................... 8

3 Statistical models for the creation of injury risk functions ....................................... 10
  3.1 Curve fitting based on relative frequencies ......................................................... 11
  3.2 Logistic regression ............................................................................................... 14
  3.3 Contour Lines of Equal injury Severity ................................................................. 18
  3.4 Survival analysis ................................................................................................. 23
  3.5 Limitations of statistical models ......................................................................... 25
  3.6 Measures for model quality ................................................................................. 26

4 Creation of injury risk functions for pedestrians ...................................................... 32
  4.1 Selection and filtering of the dataset .................................................................... 32
  4.2 Selection of the dependent variable .................................................................... 34
  4.3 Selection of the independent variable ................................................................ 36
  4.4 Selection of the statistical model ........................................................................ 50
  4.5 Computation of injury risk functions .................................................................. 51
  4.6 Discussion ............................................................................................................ 60
  4.7 Limitations and user references .......................................................................... 68
<table>
<thead>
<tr>
<th>CONTENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 Creation of injury risk functions for cyclists............................ 69</td>
</tr>
<tr>
<td>6 Summary and outlook ...................................................................... 79</td>
</tr>
<tr>
<td>7 References ................................................................................... 81</td>
</tr>
<tr>
<td>8 List of figures ............................................................................ 84</td>
</tr>
<tr>
<td>A Censored data ................................................................................ VII</td>
</tr>
<tr>
<td>B Extension of the “Contour Lines of Equal injury Severity” ................ XI</td>
</tr>
<tr>
<td>C Examples for remarkable car – cyclist accidents ........................... XIII</td>
</tr>
<tr>
<td>D Influence of the secondary impact on the injury severity ................ XV</td>
</tr>
</tbody>
</table>
II Symbols

Capital letters

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIS</td>
<td>Abbreviated Injury Scale</td>
</tr>
<tr>
<td>BRPX</td>
<td>Contact point (first contact) at the vehicle in x-direction</td>
</tr>
<tr>
<td>CDC</td>
<td>Collision Deformation Characteristics (SAE J224)</td>
</tr>
<tr>
<td>D</td>
<td>Domain of a function $f: D \rightarrow \mathbb{R}$</td>
</tr>
<tr>
<td>F</td>
<td>Death- / injury probability</td>
</tr>
<tr>
<td>H</td>
<td>Absolute cumulative frequency</td>
</tr>
<tr>
<td>ISS</td>
<td>Injury Severity Scale</td>
</tr>
<tr>
<td>ISSx</td>
<td>Metric ISS</td>
</tr>
<tr>
<td>MAIS</td>
<td>Maximum AIS value</td>
</tr>
<tr>
<td>MAIS_{car}</td>
<td>MAIS of injuries caused by vehicle</td>
</tr>
<tr>
<td>MAIS_{sec}</td>
<td>MAIS of injuries caused by secondary collision</td>
</tr>
<tr>
<td>R</td>
<td>Relative cumulative frequency</td>
</tr>
<tr>
<td>$R^2$</td>
<td>Coefficient of determination</td>
</tr>
<tr>
<td>S</td>
<td>Survival function</td>
</tr>
<tr>
<td>X</td>
<td>Period until occurrence of an event</td>
</tr>
</tbody>
</table>

Small letters

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>Slope of a linear function</td>
</tr>
<tr>
<td>b</td>
<td>Ordinate intercept</td>
</tr>
<tr>
<td>e</td>
<td>Euler’s number ($e \approx 2,71828183$)</td>
</tr>
<tr>
<td>h</td>
<td>Absolute frequency</td>
</tr>
<tr>
<td>i, j, k</td>
<td>Control variables</td>
</tr>
<tr>
<td>n</td>
<td>Number of cases</td>
</tr>
<tr>
<td>p</td>
<td>Probability</td>
</tr>
<tr>
<td>r</td>
<td>Relative frequency</td>
</tr>
<tr>
<td>t</td>
<td>Period</td>
</tr>
<tr>
<td>v</td>
<td>Velocity, speed</td>
</tr>
<tr>
<td>x</td>
<td>Independent variable</td>
</tr>
<tr>
<td>y</td>
<td>Dependent variable</td>
</tr>
<tr>
<td>$\epsilon$</td>
<td>Coefficient of restitution</td>
</tr>
</tbody>
</table>
## Indices

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Initial value</td>
</tr>
<tr>
<td>car</td>
<td>Passenger car</td>
</tr>
<tr>
<td>class</td>
<td>Classification result</td>
</tr>
<tr>
<td>coll</td>
<td>Collision</td>
</tr>
<tr>
<td>crit</td>
<td>Critical</td>
</tr>
<tr>
<td>cyc</td>
<td>Cyclist</td>
</tr>
<tr>
<td>fatal</td>
<td>Fatally injured, deceased</td>
</tr>
<tr>
<td>ISSx</td>
<td>Metric ISS</td>
</tr>
<tr>
<td>i,</td>
<td>Control index</td>
</tr>
<tr>
<td>j</td>
<td>Control index</td>
</tr>
<tr>
<td>l</td>
<td>Lower threshold</td>
</tr>
<tr>
<td>MAIS1</td>
<td>MAIS = 1</td>
</tr>
<tr>
<td>MAIS2+</td>
<td>MAIS ≥ 2</td>
</tr>
<tr>
<td>MAIS3+</td>
<td>MAIS ≥ 3</td>
</tr>
<tr>
<td>n</td>
<td>Final value (variable)</td>
</tr>
<tr>
<td>ped</td>
<td>Pedestrian</td>
</tr>
<tr>
<td>rel</td>
<td>Relative</td>
</tr>
<tr>
<td>thres</td>
<td>Threshold</td>
</tr>
<tr>
<td>tot</td>
<td>Total</td>
</tr>
<tr>
<td>u</td>
<td>Upper threshold</td>
</tr>
</tbody>
</table>

## Abbreviations

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AEB</td>
<td>Autonomous Emergency Braking</td>
</tr>
<tr>
<td>AIC</td>
<td>Akaike Information Criterion</td>
</tr>
<tr>
<td>AUC</td>
<td>Area Under Curve</td>
</tr>
<tr>
<td>BASSt</td>
<td>Federal Highway Research Institute (Germany)</td>
</tr>
<tr>
<td>BIC</td>
<td>Bayesian Information Criterion</td>
</tr>
<tr>
<td>engl</td>
<td>English</td>
</tr>
<tr>
<td>Euro NCAP</td>
<td>European New Car Assessment Programme</td>
</tr>
<tr>
<td>FAT</td>
<td>Research Association of Automotive Technology (Germany)</td>
</tr>
<tr>
<td>FARS</td>
<td>Fatality Analysis Reporting System</td>
</tr>
<tr>
<td>GIDAS</td>
<td>German In-Depth Accident Study</td>
</tr>
<tr>
<td>IRF</td>
<td>Injury Risk Function</td>
</tr>
<tr>
<td>JDME</td>
<td>Year of vehicle model's market introduction</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver Operating Characteristics</td>
</tr>
<tr>
<td>TBI</td>
<td>Traumatic Brain Injury</td>
</tr>
<tr>
<td>VDA</td>
<td>Association of Automobile Industry</td>
</tr>
<tr>
<td>VUFO</td>
<td>Institute for Traffic Accident Research at Dresden University of Technology</td>
</tr>
</tbody>
</table>
III Abstract

Injury risk functions gain importance in the field of vehicle safety. Whether for the understanding of injury mechanisms, the definition of protection criteria or the transfer of simulation results on real accident scenario. It often requires statements about the correlation between technical accident parameters, such as loads or accident severity, and the predicted injury severity. So far, there are no established, standardized and generally accepted definitions for the creation of statistical models. The scope of this report includes the development of a consistent and coherent approach for creating injury risk functions based on real world accident data. Thereby the focus lies in the analysis and evaluation of applied statistical models. One result are injury risk functions for pedestrians and cyclists in full frontal car collisions.

The fundamental principles of injury risk functions are gathered and explained by an extensive analyse. As result of both a literature review and consultations with specialists in the departments of various automobile manufacturers, the report analyses and evaluates the established statistical models:

- Curve fitting based on relative frequencies
- Binary Logistic regression
- Multinomial Logistic regression
- Contour Lines of Equal injury Severity
- Survival analysis

regarding applicability for the creation of injury risk functions.

The developed procedure is applied to car-pedestrian- and car-cyclist-accidents based on the real accident database GIDAS (German In-Depth Accident Study). Therefore, the report defines boundary conditions based on the specific problem, formulates a statistical model and creates injury risk functions for the applied sample afterwards. Essential objectives are the identification of the injury values, the analysis and evaluation of potential influencing parameters and the selection of an appropriate statistical model. Based on the findings the center of interest lies in the calculation of specific injury risk function for pedestrians and cyclists in full frontal vehicle collisions. Thereby, model quality measures represent important values for an estimation of the model’s ability to explain occurrences within the sample.
Main influencing parameters include collision speed (pedestrians) and relative speed (cyclists), person’s age (creation of homogeneous age groups: children, adults, elderly), type of impact at the vehicle (full frontal) and vehicle model’s age (passenger cars with market introduction as from 2000). By using binary logistic regression, these are included in the creation of the models for MAIS2+ and MAIS3+ injured adult and elderly pedestrians. Related figures visualize the injury risk functions. Achieving decent model quality measures enables an application to the assessment of vehicle safety systems. The agreed boundary conditions cause very small case numbers for the groups of children and all fatally injured pedestrians. Unfortunately, this leads to unreliable models for these groups so far.

The created pedestrian models can analogous be applied to car-cyclist-accidents by using the more appropriate relative speed parameter instead of collision speed. However, acceptable model quality measures cannot be achieved and a valid evaluation of injury risks is not ensured. Much more than for pedestrian accidents the complex impact constellations at passenger car - cyclists collisions suggest a wide heterogeneity of the groups. Some deeper analysis regarding collision constellations and predictive parameters promises more resistant probability models here.
1 Introduction

Injury risk functions (IRF) represent substantial elements for the assessment of vehicle safety systems. They are able to predict expected injury severities depending on one or more accident parameters. They are particularly qualified for the assessment process when system warnings or system intervention could change the consequences of a real accident.

Currently there are several statistical models for the creation of injury risk functions in use but no established, standardized and generally accepted method existent at all (Banglmaier et al. [6], Hasija [14], Kröyer et al. [22], Niebuhr et al. [32], [33]). Functions generated by different approaches may lead to similar results or can also show considerably distinctions (e.g. Hasija [14]). Therefore, results of the assessment of vehicle safety systems based on the real world accident scenario is strongly dependent on the chosen statistical model.

Harmonization and recommendation of an established method for the creation of injury risk functions based on real world accident data is of great interest for a consistent assessment of vehicle safety systems. An ISO document for the creation of injury risk functions based on biomechanical samples is currently under revision [18]. A harmonized procedure would also be desirable in the field of vehicle safety research respectively the usage of real world accident data.

Just few injury risk functions based on accident data containing large number of cases have been published so far ([15], [32], [37], [41]). The objective of this report giving an overview of currently used statistical models. Followed by the selection of an appropriate one the focus is on the creation of specific injury risk functions for pedestrians and cyclists (based on the GIDAS (German In-Depth Accident Study) real world accident database) as a basis for the assessment of vehicle safety systems.
2 Principles of injury risk functions

2.1 Definition

In contrast to other academic terms, there exists no injury risk curve a clear definition nor description for “injury risk function” (IRF). IRFs gain currency nearly exclusively in the field of biometry / biomechanics as well as in the field of vehicle safety research. So to say, they provide an interdisciplinary interface of such two areas. The majority of biomechanical studies contribute to the understanding of injury mechanisms in traffic accidents. Their usage is increasing as method for vehicle safety studies. Nevertheless, differences in objective, used parameters and applied methods can be determined which will be discussed within this report.

An injury risk function describes the probability of the occurrence of a specific injury criterion (dependent variable) in dependence of one or more variables (independent variable) for a specific population.

2.2 Dependent variable

The injury criterion of the injury risk function is known as dependent variable. It is mostly binary resp. dichotomous but can also be multinomial.

The fields of vehicle safety and traffic accident research generally mean the occurrence of the (minimum) injury severity of a certain level. Therefore a measurable or investigated (usually ordinal scaled) value \( z \) is transferred into the binary dependent variable \( y \) using a specific and related threshold \( z_{thres} \).

\[
y = \begin{cases} 
1 & \text{for } z \geq z_{thres} \\
0 & \text{for } z < z_{thres}
\end{cases}
\]  
(2.1)
As an example, the following definition describes the dependent variable \( y \) for receiving at least one injury with a severity according to the Abbreviated Injury Scale (AIS, further information in [10]) of 2 or above (complies a maximum AIS of the person (MAIS) of 2 or more; described as “MAIS2+”):

\[
y = \begin{cases} 
1 & f \text{ or MAIS} \geq 2 \\
0 & f \text{ or MAIS} < 2 
\end{cases}
\]  

(2.2)

It is possible to create injury risk functions for the injury severity of:

- The whole person (e.g. official definition of injury severity, MAIS, ISS, etc.)
- Body region (e.g. AIS\(_{\text{Head}}\))
- Single injuries (e.g. AIS)

Injury risk functions are most common created for injury criterions like:

- \( p_{\text{MAIS2+}} \) probability of being injured with at least MAIS2+
  
  \[
  \text{MAIS2+} = \begin{cases} 
  0 & \text{MAIS0, MAIS1} \\
  1 & \text{MAIS2, MAIS3, MAIS4, MAIS5, MAIS6} 
  \end{cases}
  \]

- \( p_{\text{MAIS3+}} \) probability of being injured with at least MAIS3+
  
  \[
  \text{MAIS3+} = \begin{cases} 
  0 & \text{MAIS0, MAIS1, MAIS2} \\
  1 & \text{MAIS3, MAIS4, MAIS5, MAIS6} 
  \end{cases}
  \]

- \( p_{\text{severe}} \) probability of at least being seriously injured
  
  \[
  \text{Severe} = \begin{cases} 
  0 & \text{uninjured, slightly injured} \\
  1 & \text{seriously injured, fatally injured} 
  \end{cases}
  \]

- \( p_{\text{fatal}} \) probability of being fatally injured
  
  \[
  \text{Fatal} = \begin{cases} 
  0 & \text{uninjured, slightly injured, seriously injured} \\
  1 & \text{fatally injured} 
  \end{cases}
  \]

2.3 Independent variables

Independent variables are parameters, which contributes to the prediction of the injury severity. They influence the dependent variable and consists of physical parameters, which are supposed to predict the injury criterion.

Be that

\[
x = [x_1, x_2, x_3, \cdots, x_n]^T
\]

(2.3)

is a vector to describe the independent variable \( x_1, x_2, x_3, \cdots, x_n \).

The independent variables may have one of the following characteristics:

- Nominal scale (e.g. male/female, Audi/BMW/Mercedes/OPEL/VW/…)
- Ordinal scale (e.g. AIS 1 - 6)
- Cardinal scale (e.g. model year, age)
In addition, a distinction is made between categorical (discrete) and metric (continuous and quasi-continuous) values. In the field of vehicle safety and traffic accident research, these are primarily values describing the collision- and impact severity:

- Physical parameter
  E.g. load, torque, speed, acceleration, etc.

Moreover, they can also include:

- Physiological/ individual parameter
  E.g. gender, age, weight, height, pre-existing illness, etc.

- Situational parameter
  E.g. side of impact, seatbelt usage, point of impact, etc.

- Vehicle specific characteristics
  E.g. vehicle age, existence of passive safety systems, front shape, etc.

In biomechanics the injury criterion often include physical parameter, correlating well with an injury severity measure (of a specific body region) (cf. [18]). Examples are HIC₁₅, Nᵢⱼ as well as chest deflection or chest acceleration.

2.4 Mathematical function

The mathematical function constitutes the connection between the probability of the occurrence $p$ of the dependent variable $y$ and the independent variable(s) $x$:

$$p(y = 1) = f(x)$$  \hspace{0.5cm} (2.4)

Because of the dichotomous (0 or 1) characteristics of the dependent variable $y$, the probabilities of the complementary events sum up to 1:

$$p(y = 0) + p(y = 1) = 1$$ \hspace{0.5cm} (2.5)
Generally, there exists no mandatory mathematical function $f(x)$ describing the injury risk function. Since it is a probability function, following boundary conditions must be fulfilled (exceptions see below):

A1 Both extreme values must not fall below 0 or exceed 1:

$$0 \leq p(x) \leq 1 \quad \text{for} \quad x \in D \quad (2.6)$$

A2 The probability of the occurrence of a certain injury severity level has to be 0, if all independent variables $x$ equal 0:

$$p(x = 0) = 0 \quad (2.7)$$

A3 The probability of the occurrence of a certain injury severity level must be 1, if all metric dependent variables $x$ converge towards $\infty$ (or all categorical/nominal variables equal such value, that lead to the highest probability of the injury severity level):

$$p(x \to \infty) = 1 \quad (2.8)$$

A4 The probability for the occurrence of a certain injury severity at an equivalent accident scenario must be greater or equal (monotonous increasing) when any independent variable is greater:

$$p(x_i \geq l) \geq p(x_i \geq k) \quad \text{for all} \quad k, l \in D \text{ with } l \geq k \quad (2.9)$$

A5 The probability for the occurrence of a certain injury severity $z_k$ must be greater or equal than the probability for the occurrence of another higher injury severity $z_l$:

$$p(z \geq z_k) \geq p(z \geq z_l) \quad \text{for all} \quad z_k, z_l \in D \text{ with } z_l > z_k \quad (2.10)$$

There may be exceptions for such conditions depending on the statistical model. The common use of collision speed of the opponent vehicle as predictor for the injury risk of pedestrians is an example for not fulfilling condition A2. Injuries may happen even at collision speed $v_{coll} = 0 \, \frac{km}{h}$ because of pedestrian’s speed as well as through falling as a reaction of dodging or frightening and related secondary collisions. If these predictors extend the model, the injury risk function needs to meet condition A2 again when no other influences are relevant.
An example for an exception when the model does not need to fulfill condition A4 in any case is the selection of the predictor: "age of the pedestrian". Due to the anatomy of children (skeleton growth, muscle growth) and the different impact points of children at the vehicle due to lower height compared to adults, the injury risk $p(age)$ increases not necessarily monotonous at every age range.

Therefore, the obligatory status of each condition depends on the exact model.

The task when creating injury risk functions is to find an appropriate function that satisfies all boundary conditions and adequately represent the empirically observed data. The shape of the function is defined by specifications of the mathematical function and by empirical coefficients. The application of logistic functions is common, because of their sigmoid shape (S-shaped). They approximate the points $p(x = 0) = 0$ and $p(x \to \infty) = 1$ well. Furthermore, they are able to illustrate the dose-effect-relation, which is often present in the field of biomechanics and vehicle safety ("the more [load] … the more [level of injury severity]"). Beside the logistic function, there are a few more functions, each with their own advantages and disadvantages. Chapter 3 deals with such statistical models for the creation of injury risk functions.

2.5 Application

The main objective of an injury risk function in the field of traffic accident research is its application when assessing active and passive vehicle safety systems. The function is either used as a tool for the assessment of the expected variation of the accident severity caused by active and accident sequence intervening vehicle safety systems or for a comparative evaluation of passive activities. The probability of a discrete injury severity is evaluated by subtraction of several injury risk functions referring to different dependent variable thresholds.
Figure 2.1 – Examples of injury risk functions including reading example

Figure 2.1 exemplarily shows two injury risk functions for the probability of different injury severities of pedestrians (red – at least seriously injured, blue – at least slightly injured). These are generated by binary logistic regression based on real accident data. Probability values for different injury severities are stated in dependence of the collision speed (= speed of the vehicle in the moment of the frontal collision with the pedestrian). It demonstrates this procedure as an example for the application of an emergency brake function towards an impending pedestrian collision at the vehicle front (Pedestrian AEB). If it is possible to reduce the collision speed (of the real accident) from 50 km/h to 40 km/h (exemplarily as the figure shows) by system intervention and/or driver’s reaction, probability of each injury severity changes.

It is important that this method is a statistical approach and does not apply to any single case. Usually hundreds or even thousands of real accidents (without safety system) are simulated with a safety system. Afterwards the result can be compared in assistance of the injury risk function.

Empirical data usually determines injury risk functions retrospectively and the assessment of future safety systems uses them prospectively. They provide an important possibility to evaluate the consequences of modified (but not prevented) accidents. They may also be used for a single vehicle safety function (e.g. qualification of an e-call by prediction of the expected injury severity in dependence of the accident severity determined by the vehicle as well as individual driver attributes).
2.6 Challenges

In general, creation of injury risk functions involves several challenges influencing reliability significantly. Praxl [38] has evaluated the following four factors for biomechanics:

- **Sampling**
  
The uncertainty due to statistical conclusions from a sample to the population is named “effect of the sample”.

  Low sample sizes (in biomechanical studies usually less than 100) often lead to the necessity of interpolation/imputation or analyses of statistically insufficient dataset.

  In addition, the distribution of people is not the same as the distribution of the total population. So there are physiological differences in biomechanical investigations with corpses due to examination of mostly elderly bodies, resulting in different effects compared to living people ([13], [45]).

- **Censoring of data**
  
The uncertainty due to inaccuracy in determining the tolerance limits is called “effect of censoring”

- **Test severities**
  
The uncertainty of injury severity prediction due to the chosen test load value is referred by Praxl as “effect of the test”

- **Statistical model**
  
  Bias due to chosen statistical method for the creation of the injury risk function are summarized as “effect of statistics”.
Some of the mentioned challenges for biomechanical tests can be avoided in the analysis of real accident data, whereas others have to be taken into account to the same extent.

There is one big advantage of real accident data, that mostly there is a large number of cases within the database. This sets the sample size challenge to a much lower extent. On the other hand, it is important to observe the representativeness/generality of the dataset. The specific investigation criteria and objectives of the data acquisition lead to biases in the population. For example, the American database “Fatality Analysis Reporting System” (FARS) only contains fatally accidents. Although it contains a large amount of data, it does not allow generally valid conclusions regarding the risk of injury.

In contrast, the GIDAS database contains accidents in which at least one person was injured. Therefore, GIDAS is more generally valid than the FARS database, but does not contain accidents with exclusively material damage. In general, such problems can be found in a similar manner in almost every accident database. Chapter 4.1 gives a detailed review of the used dataset.

Censoring of real world accident data is not a big question. It is almost always left- and right censored, because of representing single multivariate events and being un reproducible.

The selection of the area of the tested load magnitude is an important decision and uncertainty factor for biomechanical tests. For real accident data, there is no importance. The existence of reconstruction data provides various parameters, which can be used for the specification of the accident severity. Consequently, there is no need for simulation or dummy tests for the definition of the load. The distribution of the injury severity (dependent variable) and the influencing factors (independent variables) in the dataset is, on the other hand, an important consideration for the estimation of confidence intervals of an injury risk function.

The underlying statistical model also has an important influence to injury risk functions of real accident data. Therefore, the following chapter considers three statistical models for the creation of injury risk functions in detail.
3 Statistical models for the creation of injury risk functions

As mentioned in the previous chapter, there is no established standard for the creation of injury risk functions based on real accident data. However there is a necessity resulting from the variety of possible statistical models and specific properties of the underlying data. Various studies (often in the field of biomechanics) already dealt with different approaches and their similarities, differences and inaccuracies (et al. [6], [14], [22], [24], [34], [38]).

This section shows the results of a literature review and a survey of specialist departments of various automobile manufacturers. This gives an overview about used statistical models and their advantages and disadvantages regarding the creation of injury risk functions based on real accident data. The critical reflection considers aspects of dissemination among experts, applicability in the field of vehicle safety and required data basis. Subsequent sections examine following statistical models, which represent established approaches:

- Curve fitting based on relative frequencies
- Binary Logistic regression
- Multinomial Logistic regression
- Contour Lines of Equal injury Severity
- Survival analysis

There are more models, which are not / only rarely used or slight modifications / special cases of the observed models. The report lists some with a literature reference but does not described them in detail:

- Probit Regression Model Agresti [1], Schild [42]
- Consistent Threshold Method Nusholtz/Mosier [35]
- Certainty Method Mertz et. al. [29]
- Mertz/Weber Method Mertz/Weber [28]
- Modified Logistic Regression Model Nakahira [31], Banglmaier [6]
- Risk Curve Evaluation Criteria Nakahira [31]
3.1 Curve fitting based on relative frequencies

The curve fitting based on relative frequencies is a simple model for the description of an injury risk in dependence of an independent variable. Therefore, each independent variable (e.g. collision speed) value calculates the relative frequency of a certain injury severity (e.g. MAIS 2+) in the sample. Subsequently, the model performs curve fitting based on the observed data using the method of least squares. This means the method represents a regression method. A logistic function is appropriate for modeling the link between independent and dependent variables as described in section 2.4.

Figure 3.1 exemplary shows the observed accident data (left side) based determination of the single relative frequencies for each collision speed (right side) for a minimum injury severity $MAIS \geq 2$ (MAIS 2+). Figure 3.2 shows the related graph. Based on the observed data it clearly points out an increasing trend of being MAIS 2+ injured at increasing collision speed $v_{coll}$. It is important to state, that data points in lower definition area (low $v_{coll}$) are based on a larger amount of observed accidents. This leads to tendencies better accuracy than data points at higher collision speed occurring more rarely.

<table>
<thead>
<tr>
<th>Case-No.</th>
<th>$v_{coll}$</th>
<th>MAIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>35</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>45</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>43</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>18</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1072</td>
<td>54</td>
<td>1</td>
</tr>
</tbody>
</table>

![frequ. MAIS2+ per $v_{coll}$ (km/h)](image)

<table>
<thead>
<tr>
<th>$v_{coll}$</th>
<th>$n_{MAIS1}$</th>
<th>$n_{MAIS2+}$</th>
<th>$n_{tot}$</th>
<th>frequ. MAIS2+</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2</td>
<td>-</td>
<td>2</td>
<td>0%</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>0%</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>-</td>
<td>6</td>
<td>0%</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>1</td>
<td>6</td>
<td>17%</td>
</tr>
<tr>
<td>5</td>
<td>22</td>
<td>9</td>
<td>31</td>
<td>29%</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>2</td>
<td>9</td>
<td>22%</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>2</td>
<td>10</td>
<td>20%</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>1</td>
<td>10</td>
<td>10%</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>25%</td>
</tr>
<tr>
<td>10</td>
<td>40</td>
<td>4</td>
<td>44</td>
<td>9%</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Figure 3.1 – Example for relative frequencies of MAIS2+ injuries over collision speed from GIDAS

This is caused by characteristics of the population. It is useful to observe the distribution of the independent variable ($v_{coll}$ in the example) within the whole population. Figure 3.2 also shows, that not all discrete $v_{coll}$ steps (according to the discretization $1\ km/h$) include sufficient accidents. Especially areas of higher $v_{coll}$ show missing data and statistical outliers due to a small number of cases.
Subsequently, the selection of an appropriate regression function is required. Possible approaches are polynomial (simple linear, linear square, linear cubic, etc.), exponential, logistical et al. The functional shape is determined by minimizing the statistical error with the least squares method. Figure 3.3 shows the examples of a linear regression with different polynomial degrees and boundary conditions.

**Figure 3.2 – Observed shares (relative frequencies) from GIDAS**

Proportion of MAIS2+ injured pedestrians at full frontal car impacts in dependence of collision speed (GIDAS, n = 1,072)
A big advantage of the model is the adjustment of the curve based on the observed data. The selection of the curve shape (regression function) is arbitrary, but affects the curve progression significantly. An additionally forced zero crossing may fulfill condition A2.

In contrast, the function is not necessarily monotonous increasing (no fulfilment of condition A4) and can meet values \( p(x) < 0 \) (without constraints) or \( p(x) > 1 \) (no fulfilment of condition A1 or A3). This would mean there is be no probability statement possible. Multivariate considerations are only possible in a very limited way. There is no weighting of the data points to each other.
In summary, the statistical model of curve fitting based on relative frequencies is only applicable in a limited way on the creation of injury risk functions. Conditions mentioned in section 2.4 may not be fulfilled and no multivariate consideration of several influences is possible.

3.2 Logistic regression

Logistic regression (or logit model) uses a regression approach to determine the probability of an event (dependent variable) in dependence of various influencing parameters (independent variables) [4]. Therefore, it is a statistical classification function. Logistic regression is often used, because the link between dependent and independent variable is non-linear.

The chapter distinct between the binary logistic regression with a dichotomous dependent variable and the multinomial logistic regression with a nominal dependent variable and more than two discrete characteristics. Usually logistic regression means the binary form, if there are no further information.

Binary logistic regression estimates the probability function for the occurrence of a dichotomous dependent variable (e.g. fatally injured: yes/no). Therefore, such (observed) influencing factors (independent variables) are determined, which best describe the affiliation to one of the two dependent variable groups.

The assumption of a latent variable $z_j$ as a linear combination of the independent variable $x_i$ with:

$$z_j = \beta_0 + \sum_{i=1}^{n} \beta_i \cdot x_{i,j}$$  \hspace{1cm} (3.1)

enables the link between dependent and observed independent variable:

$$y_j = \begin{cases} 1 & \text{for } z_j > 0 \\ 0 & \text{for } z_j \leq 0 \end{cases}$$  \hspace{1cm} (3.2)

An iterative algorithm estimates the regression constant $\beta_0$ and the regression coefficients $\beta_i$ (often-called logit-coefficients) through the maximum likelihood method.
Unlike linear regression methods, the dependent variable is not metric, but nominal. There is no prediction of actual values, but their likelihood occurrence. Therefore, the model uses the logistic function:

\[
p(y = 1) = \frac{1}{1 + e^{-z}} \tag{3.3}
\]

with \( e \approx 2.71828183 \) (Euler's number)

Regression coefficients \( \beta_i \) give an indication on the impact strength of the respective independent variable \( x_i \). The logistic function moves (\( f: D \rightarrow [0; 1] \)) in the whole domain within the interval \([0; 1]\), which fulfills condition A1 of a probability function and shows a sigmoid shape (s-shaped), strictly monotonous increasing (cf. condition A4). In addition, it is symmetrical to the inflection point \( (y = 1) = 0,5 \).

Figure 3.4 shows the example of a binary, logistic regression curve out of empirical data as a probability function for MAIS2+ injured pedestrians in full frontal collisions with a passenger car in dependence of the collision speed \( v_{\text{coll}} \) (no restriction of vehicle age, including grazing).

![Figure 3.4 – Example of a binary logistic regression curve](image)
As with all parametric models, the parameters (here: regression coefficients) of the binary logistic regression are estimated based on empirically observed data. In addition, a multivariate estimation is possible in order to consider several independent variables, which describe the dichotomous expression of the dependent variable. The sigmoid shape shows the dose-response relationship very well, cannot adopt probabilities less than zero or greater than one (domain $W = \{ p(x) | 0 < p(x) < 1 \}$, cf. condition A1) and is strictly monotonically increasing in the entire domain (compare condition A4).

Hasija [14] tested several statistical models regarding god and poor correlating data. The binary logistic regression shows a good function curve for good correlating data and a poor function curve for poor correlating data. Other methods sometimes also produce good function curves based on poor correlating data.

A disadvantage of the logistic function is the mathematical inability of showing a zero crossing. Therefore, condition A2 ($p(x = 0) = 0$) cannot be satisfied. The probability of a relevant offset of zero depends on the sample size. The smaller the sample size, the more likely is a relevant offset [38].

Also negative are the predefined curve shape caused by the logistic function and the symmetry without a reference to the real distribution. Basically the specification of function shape applies for all parametric models and is therefore not a special issue of logistic regression.

The curve fit works particularly well in areas with high data density, whereas it may be inaccurate in areas with a low amount of underlying data points. When creating binary logistic regression curves for several injury severities higher probability values for lower injury severities is mathematically possible. This would lead to intersections of curves (e.g. MAIS2+ and MAIS3+) in areas with a low data density (cf. condition A5). This issue can be avoided by using conditional probabilities [15] or by the use of multinomial logistic regressions (see section 3.2).

In summary, the statistical model of binary logistic regression is an established and the most common method for the creation of injury risk functions based on real accident data. Because of the analytical shape of the function, it is applicable for system assessment. A consideration of several independent variables is possible (multivariate evaluation), whereby several influences can be presented as a linear combination in the so-called logit. The method ensures the reflection of the real data in total (for a real dataset with $n$ fatally injured people the model calculates exactly $n$ fatally injured people). The disadvantage like the missing zero crossing may be plausible with a corresponding independent variable (e.g. collision speed of accident oppo-
nent) or is less problematic at comparative / relative considerations (e.g. with vs. without a system). At other independent variables (e.g. relative speed), this is not the case. The model works is especially appropriate for the creation of injury risk functions, because of its alignment to real, empirical data (few assumptions necessary).

The multinomial logistic regression is an extension of the binary logistic regression, where the dependent variable accepts more than two states. The procedure and the advantages and disadvantages are analog in most points. Therefore, the following paragraph concentrates on differences to the binary model of the logistic regression.

Since the dependent variable does not have to be dichotomous anymore (if it is dichotomous, the model is identical to the binary logistic regression), it may have an ordinal scale level with more than two values. Therefore, the multinomial logistic regression model is able to model single MAIS values or a classification in slightly injured, seriously injured and fatally injured in common. Intersections of several curves of different injury severities are excluded (see condition A5), which represents an additional advantage to the binary logistic regression.

A major disadvantage of multinomial regressions is the use of the equal regression coefficients $\beta_i$ of the function curves for the different injury severities. That means that the independent variable(s) have an identical influence on various characteristics of the dependent variable, which appears quite questionable. Particularly at non-linear MAIS, this method seems not appropriate.

Also the inclusion of additional independent variables for a better description of the relationships is not possible for different injury severities. The only difference of the several curves are the intersection with the y-axis ($\beta_0$).

In summary, the multinomial regression is an extension of the established binary logistic regression to avoid intersections of probability curves of various states of the dependent variable. This satisfies condition A5. The risk functions for different characteristics of the dependent variable can be read directly. Because of the major disadvantages like the identical selection of the independent variable and the identical regression coefficients, the binary logistic regression seems more suitable with the alternative approach of conditional probability (Helmer [15]).
The probit regression model (see Agresti [1], Moore [30], Schild [42]) is very similar to the logistic regression (often called logit model in this context). The main difference is the assumption of different dependent variable distributions. So the probit model assumes a normal distribution instead of a logistic distribution (equation of the distribution function, see formula (3.4)). Figure 3.5 compares the density functions of both distribution assumptions.

\[ F(z) = \phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt \]  

(3.4)

In comparison, the models result in very slight differences. Only in areas when approximating the ordinate axis or the probability value of 100 %, there occurs a small variation.

![Comparison of the density functions of the logit model and the normal distribution of the probit model](image)

**Figure 3.5 – Comparison of the density functions of the logit model and the normal distribution of the probit model [36]**

### 3.3 Contour Lines of Equal Injury Severity

The method of “Contour lines of equal injury severity” is a model-based approach, which has been published in 2013 by Niebuhr et al. [32] and extended in the beginning of 2015 [33] (applied to different body regions).
The model describes a methodology for creating a consistent family of injury risk functions for pedestrians. The basic model is a phenomenological approach that assumes an exponential relation between injury severity and collision speed. This means the model approach bases on the assumption of influences and not on real accident data.

The model uses the injury severity with the assistance of the metric \( ISS_x \)-scale (see [19]) as dependent variable. Below, the definitions of the \( AIS_x \) and the \( ISS_x \) are mentioned.

- **Metric \( AIS_x \)-scale:**
  
  \[
  AIS_{x[i]} := \frac{25}{(e^8 - 1)} \cdot \left( e^{AIS_{[i]}} - 1 \right) \approx 0.1696 \cdot \left( e^{AIS_{[i]}} - 1 \right) \tag{3.5}
  \]

  with \( AIS := [AIS_1 \cdots AIS_9]^T \) and \( D = \{AIS_i \in \mathbb{N} | 0 \leq AIS_i \leq 6\} \)

- **Metric \( ISS_x \)-scale:**
  
  \[
  ISS_x := \sum_{i=1}^{3} AIS_{x[i]} \tag{3.6}
  \]

Table 3.1 shows the comparison between the non-metric \( MAIS \) and the metric \( ISS_x \)-scale.
Table 3.1 – Comparison MAIS and ISS to the metric ISSx

<table>
<thead>
<tr>
<th>MAIS</th>
<th>ISS</th>
<th>ISSx</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1 ... 3</td>
<td>0,29 ... 0,87</td>
</tr>
<tr>
<td>2</td>
<td>4 ... 12</td>
<td>1,08 ... 3,25</td>
</tr>
<tr>
<td>3</td>
<td>9 ... 27</td>
<td>3,24 ... 9,71</td>
</tr>
<tr>
<td>4</td>
<td>16 ... 48</td>
<td>9,09 ... 27,27</td>
</tr>
<tr>
<td>5</td>
<td>25 ... 75</td>
<td>25 ... 75</td>
</tr>
<tr>
<td>6</td>
<td>75 *</td>
<td>75 *</td>
</tr>
</tbody>
</table>

* by definition

The approach uses the collision speed \( v_{\text{coll}} \) as independent variable. It states the existence of a critical speed value \( v_{\text{crit}} \), where the injury probability equals one: \( p(v_{\text{crit}}) = 1 \). It defines an exponential function depending on the collision speed \( v_{\text{coll}} \) and the injury severity of the ISSx-scale for a full-frontal impact as:

\[
p_{\text{ISSx}}(v_{\text{coll}}) = \min_{v \geq 0} \left\{ \left( \frac{v_{\text{coll}}}{v_{\text{crit}}} \right)^{\text{ISSx}}, 1 \right\}
\]

(3.7)

In addition, a distinction of the impact type between full-frontal impact, grazing and near miss is made (see appendix B) and whether a head impact has occurred or not (“head impact” - HI and “avoided head impact” - AHI).

Figure 3.6 shows the curves for the following injury severities at a full-frontal impact with head impact:

- \( \{\text{ISSx} > 1\} = \{\text{MAIS2 +}\} \)
- \( \{\text{ISSx} > 2.5\} \approx \{\text{ISS} > 9\} \)
- \( \{\text{ISSx} > 5\} = \{\text{ISS} > 15\} = \text{polytrauma} \)
- \( \{\text{ISSx} > 10\} \approx \{\text{fatal}\} \)
Figure 3.6 – injury risk functions for pedestrians with head impact (Contour lines of equal injury severity) [32]

The extension at the latest publication adds [34] three more model updates.

A distinction is made between age groups \( j \):

- **Children** 0 – 14 years
- **Adults** 15 – 60 years
- **Seniors** older than 60 years

Advantages of the “contour lines of equal injury severity” are mainly the strictly monotonically increasing function (cf. condition A4) and the included constraint point \( p(v_{coll} = 0) = 0 \). In addition, the distinction of the impact constellation between full-frontal impact, grazing and near miss seems reasonable because of major differences in the corresponding accident sequence and energy exchange process between passenger cars and pedestrians.
Furthermore, a constant speed $v_{\text{crit}}$ is assumed, where each pedestrian suffers the related injury severity with a probability value of $p(v_{\text{crit}}) = 100\%$. Corresponding values are partially taken from old literature and vehicles of the current passenger car fleet have improved pedestrian protection systems as well as different front contour shapes. In the latest adaption of the model, this parameter is not constant for all injury severities and age groups, but is parameterized using empirical data. The use of the metric $ISSx$-scale creates a good scalability and improves the quality of the curve. However, the $ISSx$ is barely used in the field of vehicle safety and model results are hardly comparable with other curves.

A disadvantage is the fixed function shape (exponential or linear). Here, the extension of the model and the resulting adjustment to empirical data are an important development. The proposed linear modeling of grazing impacts has not been further examined yet. The correct modeling of grazing collisions could be used to evaluate pedestrian protection systems in a much more realistic manner. Another disadvantage is the univariate ($v_{\text{rel}}$ or $v_{\text{coll}}$) function creation. Although additional influences such as age, type of impact and head impact are included in the overall model, a multivariate approach is not implemented (so far) in the function itself\(^2\). The model only applies to car-pedestrian collisions. An application on cyclists or other accident scenarios has not been published yet.

In summary, the missing link to real data and the strong dependence of $v_{\text{crit}}$ (unusual, (yet) not validated with real data such as e.g. GIDAS) needs to be observed with caution. The extension of the model considers both in a better way. Currently the $ISSx$ is used seldom. The newest extension function shows only small differences compared to logistic regression function.

\(^1\) For other methods, there usually exists no $v_{\text{crit}}$. Real accident data gives no information on the existence of any critical speed.

\(^2\) Following studies additionally considered pedestrians age (and other measures).
3.4 Survival analysis

The survival analysis is a model, describing the duration until a particular event happens. The term “survival” derives from the analysis of survival time up to the time of death or component failure. The survival probability defines the probability that the time to death is longer than \( t \). “Death” refers to any event, where the field of traffic accident research or biometrics describe an injury severity or other biomechanical injury criterion. The “time” until the occurrence of the event then relates to the independent variable, for example the collision speed.

The survival function \( S(t) \) is the function for the survival probability:

\[
S(t) = p (X > t)
\]  

(3.8)

With

\( t \) - Variable (any time)

\( X \) - Event (time until “death “)

\( p \) - Probability for variable until the event \( X \)

Boundary conditions:

\[
S(t = 0) = 1
\]  

(3.9)

\[
S(t) \leq S(u) \quad \text{for } u < t
\]  

(3.10)

\[
S(t) \to 0 \quad \text{for } t \to \infty
\]  

(3.11)
Its complementary function describes the injury probability:

\[ F(t) = p(X \leq t) = 1 - S(t) \]  \hspace{1cm} (3.12)

The failure density function is the first derivative of \( F(t) \) and describes the rate of the event per time unit:

\[ f(t) = \frac{dF(t)}{dt} \]  \hspace{1cm} (3.13)

The model of the survival analysis is similar to the model of the binary logistic regression and is often used in the field of biometrics. In contrast, the model is rarely used so far in the field of accident research.

Equivalent to binary logistic regression the dependent variable is a dichotomous expression and nonlinear dependent to the categorical or metric independent variable(s). The assumption of the independent variable distribution is a major priority. Assuming a logistic distribution results exactly in the same curve as via logistic regression. In addition, the model takes censoring of data into account (for more information about censoring see appendix A). Studies on the effect of different distributions and differently censored data on injury risk functions is found in Hasija [14] and Praxl [38].

Figure 3.7 and Figure 3.8 illustrate scatter plots of different injury risk functions with various sample size. They show a similar shape compared to the model of logistic regression. The most important difference is the intersection with the y-axis at \( p(x = 0) = 0 \). For further information about the model of survival analysis, see Hosmer et. al. [17] or Kleinbaum et. al. [20].
Advantages and disadvantages of the survival analysis model are very similar to those of the binary logistic regression. This way, the model generates a “good” function curve in case of a good correlating dataset [14]. The biggest advantage is possibility (depending on the assumed distribution) to represent the injury probability for \( p(\nu_{\text{col}} = 0) = 0 \). Furthermore, it considers the distribution of the independent variable. However, this distribution is usually unknown for real accident data. Apart from that Hasija [14] has shown, that “good” function curves can also be generated for poorly correlating data.

In summary, the model of survival analysis is an appropriate model for the generation of injury risk functions. It is very similar to the logistic regression and even identical when choosing a logistic distribution. The survival analysis also offers the possibility to process interval censored or exact data, but real accident data is usually right- and left censored (see 4.1). In the field of biomechanics and the use of dummy / lab tests, the survival analysis is an established applied method. Its usage is not usual in the field of accident data so far.

3.5 Limitations of statistical models

The interpretation of results of statistical models always requires a certain individual responsibility. Each model includes some freedom of influencing the function. These vary from dataset (filter criteria, subgroups, outliers), to independent variable(s) (parameters, scale levels, scale) or the mathematical function itself. Therefore, all mod-
eled injury risk functions and their premises request a critical observation and examination.

At the same time, all models contain limitations in the modeling process itself and their interpretation. The selection of appropriate dependent and independent variables must always consider availability, accuracy and terms of usability as predictive parameter for the application of the injury risk function. This results in restrictions not always leading to the selection of the “best” parameter but the “most appropriate”. The modeling process must determine such a one.

Important to notice is the exclusive application of injury risk functions in the context of statistical methods. Single case statements must not be done.

3.6 Measures for model quality

Mathematically, even the usage of a small population or poor correlated data can produce “good” curves (cf. [14]). Therefore, it is essential to evaluate the quality of the calculated models and to assess the usability of the created functions. So there are several quality measures for the usage of statistical analysis methods.

Quality measures assess the overall model quality. This enables the possibility to make a statement on the independent variables’ ability to explain the dependent variable values. In other words, it gives an idea how well the function is able to distinct between the expressions 1/0, injured/ uninjured (or similar). At the same time model overfitting should be prevented (overfitting is the adaption of the model using too many independent variables, so it describes the selected data in a very good/exact way). Some information criteria such as the “Akaike Information Criterion” (AIC) or “Bayesian Information Criterion” (BIC) take the risk of overfitting into account.
The assessment of the calculated models uses different quality measures:

- Classification results
- Pseudo-R-square-statistics
  - Cox und Snell-R²
  - Nagelkerke-R²
- Receiver Operating Characteristic (ROC)
- Information criteria
  - Akaike Information Criterion (AIC)
  - Bayesian Information Criterion (BIC)

There are a few more quality measures such as the McFadden R², the Likelihood-Ratio-Test, the Hosmer-Lemeshow-Test, the Press’s Q-Test and others, which are not discussed further in this report (further literature see [2], [4], [7], [11], [39]).

Outlier diagnostics (which data points seem to be outliers, decision on exclusion, acceptance or model adaption) contribute to model improvements additionally to the consideration of model quality measures.

Classification results

Classification results represent one possibility to describe the quality of the adaption to observed data. They compare between empirically observed group assignment and (predicted) regression model group assignment (threshold usually \( p_{thres} = 50\% \))

with:

\[
y_{\text{class}, i} = \begin{cases} 
1 & \text{for } p_i(\mathbf{x}_i) \geq p_{thres} \\
0 & \text{for } p_i(\mathbf{x}_i) < p_{thres}
\end{cases}
\] (3.14)

Figure 3.9 shows an example of the comparison of the “detection rate”. True positives (1 observed, 1 predicted) be named as sensitivity and true negatives as specificity.

<table>
<thead>
<tr>
<th>observed</th>
<th>predicted</th>
<th>% true</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>82</td>
<td>7</td>
</tr>
<tr>
<td>1</td>
<td>21</td>
<td>23</td>
</tr>
<tr>
<td>% in total</td>
<td>77,4</td>
<td>22,6</td>
</tr>
</tbody>
</table>

Figure 3.9 – Model quality consideration: example of classification results
However, the use of classification results for assessing probability models is controversial. It is not recommended to use them as sole quality measure.

**Pseudo-R²-Statistics**

“The so-called pseudo-R² statistics try to quantify the portion of explained “variation” by the logistic regression model” [4]. They are similar to the coefficient of determination R² at linear regression and describe the overall model’s quality. For this purpose, they compare the zero model (all regression coefficients set to zero) with the full model.

**Cox-Snell-R² domain:**

\[
W_{Cox-Snell} = \{ R^2_{Cox-Snell} \in \mathbb{R} | 0 \leq R^2_{Cox-Snell} < 1 \} \tag{3.15}
\]

Literature considers \( R^2_{Cox-Snell} > 0,2 \) values as “acceptable” model quality and \( R^2_{Cox-Snell} > 0,4 \) values as “good” model quality [4].

**Nagelkerkes-R² domain:**

\[
W_{Nagelkerkes} = \{ R^2_{Nagelkerkes} \in \mathbb{R} | 0 \leq R^2_{Nagelkerkes} \leq 1 \} \tag{3.16}
\]

Literature refers \( R^2_{Nagelkerkes} > 0,5 \) values as “very good” model quality. Quality measure of Nagelkerke is preferred towards Cox-Snell, due to its clear content interpretation [4].

**Receiver Operating Characteristics (ROC)**

The ROC also serves for model’s sensitivity and specificity evaluation. Therefore, it determines relative frequency distribution of true positives (sensitivity) and true negatives (specificity) and displays each possible parameter value on the ordinate and abscissa of a diagram (see Figure 3.10). The area under curve (AUC) of the ROC is the measure for model quality. A diagonal result in an AUC-value of 0.5 and corresponds to a random binary event (e.g. Laplacian coin). A perfect model reaches an AUC value of 1.
Figure 3.10 – Model quality consideration: example for ROC AUC

It is not recommended to use the ROC/AUC for model selection because it only reflects how often a model was right or wrong while the extent of the error remains unconsidered.

Information criteria

Information criteria like AIC, AICc and BIC represent model quality measures for comparing two models with different degrees of modeling / independent variables. Using these criteria makes a statement about model improvement by including or excluding any parameter. However, a single criterion value contains no substance for interpretation. Information criteria are used to evaluate several models with identical approach. Since they represent no absolute value, it is not possible to compare two different approaches, but two (or more) identical curves of different degree of modeling (different dependent variable(s)).

Criteria evaluate model quality under consideration of not becoming unnecessarily complex (”overfitted”, referring to the number of parameters). Thereby, they represent a criterion for selection of the appropriate degree of modeling. Among others, the most important ones are the Akaike information criterion (AIC) and the Bayesian information criterion (BIC).
The AIC assumes an asymptotical convergence of the negative and logarithmic maximum of the likelihood-function \( \ln(\ell(\hat{\theta})) \) to the number of estimated parameters \( k \) for infinite large samples [2]. Therefore, the term \( 2k \) considers the number of variables as a “penalty term”.

\[
AIC = -2 \ln(\ell(\hat{\theta})) + 2k \tag{3.17}
\]

Burnham and Andersen ([8] and [7]) recommend the usage of the extension of the AIC, the AICc, for small samples.

They also see disadvantages of \( BIC \) towards \( AIC/AICc \) since it is not information derived. In contrast Schwarz [43] sees a disadvantage in the sample independent penalty term at the AIC, which prefers large samples with many parameters. Therefore, he recommends the use of the BIC, where the factor of the penalty term grows with the sample size:

\[
BIC = -2 \cdot \ln(\ell(\hat{\theta})) + k \cdot \ln(n) \tag{3.18}
\]

**Outlier diagnostics**

In addition to model quality consideration, single cases within the sample can influence the model significantly. This especially is the case when either the described model is inappropriate for the data or the model cannot describe a particular variable expression. For this reason, outlier diagnostic are useful. It can achieve a model improvement or determine and discuss the limits of the model. For each outlier must be decided, if it should be excluded, accepted or an entire model adaption is necessary.
Exemplarily the following two real accidents out of the GIDAS database mention a noticeable injury severity compared to their collision speed:

- Passenger car – pedestrian accident with $v_{\text{coll}} = 14 \text{ km/h}$ . The pedestrian has been fatally injured (MAIS = 6).
  (GIDAS case 1060868)  
The very low collision speed does not match the maximum injury severity. However, a single case analyses clarifies a link of the fatal injuries (among others contused cervical cord, dens fracture) with the secondary impact on the road/curb and the very high age of the pedestrian (male, 96 years old). This shows that especially for pedestrians collision speed (speed of the car at the time of collision) is not sufficient as sole predictive parameter. The age of the pedestrian should be part of the prediction model.

- Passenger car – pedestrian accident with $v_{\text{coll}} = 67 \text{ km/h}$ . The pedestrian has been slightly injured (MAIS = 1).
  (GIDAS case 30090252)  
Despite high collision speed and a full frontal impact, the pedestrian just received bruises, abrasions and a contused laceration at the head (each injury severity AIS = 1). The lucky collision constellation with a head- and shoulder impact at the windscreen area and the low person’s age (male, 23 years old) have been crucial. Another possible influence was the condition of the pedestrian (slightly intoxicated and under influence of narcotics).
4 Creation of injury risk functions for pedestrians

Chapter 2 and 3 explain the principles and bring possible statistical models for injury risk functions together. On such basis, this chapter performs and describes the model process on a concrete dataset. This results in functions describing the injury probability for pedestrians in a collision with a passenger car.

Model creation procedure runs the following steps:

- Selection and filtering of the dataset
- Selection of the dependent variable
- Selection of the independent variable
- Selection of the statistical model
- Computation of injury risk functions

The procedure must be understood as an iterative process, since the single steps have various cross-influences affecting each other. Thus, the selection of the dataset is depending on chosen injury criterion (dependent variable) and filtering is depending on the independent variables. Evaluation of the generated curves using model quality parameters (see section 3.6), in turn, influences the selection of the independent variable and potentially even the selection of the dataset. Furthermore, outliers, being identified according to the created models, may require a readjustment of the dataset or comprehensive changes within the model definition.

Subsequent to the model definitions, it is important to discuss the included assumptions and resulting limits.

4.1 Selection and filtering of the dataset

The creation of the injury risk functions for collisions of pedestrians and passenger cars uses the real accident database German in-Depth Accident Study (GIDAS).

GIDAS is a collaboration project of the Federal Highway Research Institute (BASt) and the Research Association for Automotive Technology (FAT) of Germany (Figure 4.1). It exists since 1st July 1999 including data of the research areas Dresden and Hannover. In these areas about 2,000 accidents per year are investigated and recorded to the GIDAS database. Each case is encoded with about 3,400 variables. Following the documentation, experienced engineers reconstruct each accident.
Figure 4.1 – Structure of the GIDAS project

Since investigation areas are topographically representative for national average, following a well-defined sampling plan and the large number of cases GIDAS database can be used for representative statements about German accident statistics. For further information see [46].

Like every database, also GIDAS contains few incomplete datasets or unknown data. As the database is very comprehensive and has a large number of cases, the study omits data imputation (mathematical method to complete missing data points) and excludes cases with unknown data. The applied master dataset is created by various filter queries and considering relevant restrictions. It contains all dependent and independent variables.
The master dataset uses the current GIDAS database status (as of 30th June 2015). The described investigation methodology leads to database biases. The evaluation of the created curves needs to take these into account (see chapter 4.6). Real accident data is always left and right censored. Unlike biomechanical samples, they are almost unrepeatable and always represent multivariate single events. The knowledge of the exact parameter distribution is almost never present for real accident data.

The study creates the available dataset for calculating the injury risk functions for accidents of passenger cars and pedestrians as following (n = sample size = number of accidents):

<table>
<thead>
<tr>
<th>GIDAS sample (07/2015): n = 27,051</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection of all accidents with available information of side of the impact, injury severity and age of the pedestrian, collision speed, age of the vehicle model:</td>
</tr>
<tr>
<td>First collision car – pedestrian: n = 2,134</td>
</tr>
<tr>
<td>only full frontal impact: n = 1,397</td>
</tr>
<tr>
<td>Year of vehicle model’s market introduction off 2000: n = 305</td>
</tr>
<tr>
<td>Exclusion of grazing and special cases*: n = 288</td>
</tr>
<tr>
<td>Master dataset for IRF creation: n = 288</td>
</tr>
</tbody>
</table>

* e.g. further collision of the pedestrian with another moving vehicle, pedestrians pinched between two vehicles, suicide attempts, kneeling/lying pedestrian etc.

**Figure 4.2 – Sample of accidents involving pedestrians and passenger cars out of GIDAS**

Grazing and special cases are excluded by automatization and single case analyses. Thus, the elimination of a certain subjective influence at the inclusion or exclusion of special cases is not entirely possible.

**4.2 Selection of the dependent variable**

The selection of the injury criterion (dependent variable) depends on the objective. It is therefore necessary to find a suitable injury criterion for the creation of the injury risk function to predict the probability of an injury severity. This study creates injury risk functions for various injury values, since the required criterion is often based on the application.
Many analysis use the MAIS (maximum value of all single injury severities based on the Abbreviated Injury Scale (AIS), further information in [10]), since this value describes the physical injury severity very well. However, the MAIS does not give a direct statement about a person’s decease. There may occur a fatally injured person with an $MAIS = 3$ caused by traffic accident consequences, whereas another person with an $MAIS = 5$ survives. The official injury severity ($PVERL$, definition according to [44]) realizes such statement as it distinguishes between:

- Uninjured
- Slightly injured (ambulant therapy and injuries without hospitalization)
- Seriously injured (stationary therapy within a hospital)
- Fatally injured (within 30 days as a result of the accident)

However, such definition involves the disadvantage to describe the physical injury severity worse than the MAIS, because it primarily orientates on the time of hospitalization. Occasionally persons are classified as seriously injured, even though they suffer from no injury ($MAIS = 0$), but stay for more than 24 hours in the hospital for monitoring.

Just indicating the most severe single injury is a disadvantage of the MAIS. It may happen that very different injured persons reach the same $MAIS$ and therefore formally being assigned to the same lethality probability. Exemplarily a person with a polytrauma consisting of a craniocerebral trauma $2^{nd}$ degree (AIS = 3), hemopneumothorax (AIS = 3) and a liver rupture (AIS = 3) relate to $MAIS = 3$ as well as a person “just” suffering a (complex) tibial fracture (AIS = 3). An approach that attempts to consider such effects is called “Injury Severity Score” (ISS, see [5]), which sums the squares of the three highest AIS-values out of six different body regions. Since the classification of the AIS is not metric but ordinal, MAIS and ISS are also not metric. Therefore, a comparability of different injury severities is only possible in an ordinal classification. Junge et al [19] attempt to formulate a metric ISSx.
The following represent appropriate and frequently used dependent variables for injury risk functions in the field of traffic accident research:

- Official injury severity
- AIS (single injuries, body regions)
- MAIS (overall injury severity)
- ISS / ISSx

Since every scale involves its characteristically advantages and disadvantages, it is important to perform the selection according to the objective. Additionally availability of the data is relevant. Most national traffic accident statistics only contain the official injury severity and have no information about the MAIS. This report favours the MAIS as a measure for injury severity, based on given requirements regarding applicability, database availability and adequate significance. Additionally, it considers the official injury definition “killed”, in order to make additional statements about fatally injured pedestrians.

The creation of injury risk functions uses the following dependent variables:

- MAIS2+

\[
y_{MAIS2^+} = \begin{cases} 
1 & \text{for } MAIS \geq 2 \\
0 & \text{for } MAIS < 2 
\end{cases}
\] (4.1)

- MAIS3+

\[
y_{MAIS3^+} = \begin{cases} 
1 & \text{for } MAIS \geq 3 \\
0 & \text{for } MAIS < 3 
\end{cases}
\] (4.2)

- Fatally injured (by official definition)

\[
y_{fatal} = \begin{cases} 
1 & \text{for } PVERL = "fatal" \\
0 & \text{for } PVERL \neq "fatal" 
\end{cases}
\] (4.3)

The evaluation of the injury severity by AIS-scale uses the AIS definition of 2005, update 2008.

4.3 Selection of the independent variable

The selection of the independent variable(s) for the description of the previously mentioned dependent variable bases on the analysis of potential prediction variables. Many studies about evaluating injury risks already dealt with potential parame-
ters for the prediction of the probability of occurrence. Influences are not always directly visible and even good a correlation does not necessarily imply a causality. Accuracy of parameters, especially reconstruction parameters (e.g. “How accurate is collision speed in accident databases?” particularly in the low speed region because of minor deformations on the vehicle etc.), should be questioned critically and their applicability needs to be discussed. Important issues are:

- Applicability and significance
- Availability in the dataset
- Distribution
- Accumulation at certain (encoding) values
- Encoding errors
- Systematic errors due to data investigation method

The statistical model can consider parameter either directly through their impact on the model itself or indirectly through dataset filtering. One has to keep in mind that filtering limits the generality of the resulting function.

For reasons of clarity, the report divides potential parameters into:

- Collision parameters
- Vehicle parameters
- Individual characteristics of the pedestrian

**Collision parameters** are parameters, which describe or divide types of collisions or which quantify the severity of a collision. This includes characteristically measures like speed (collision speed, relative speed, initial speed, difference of speed, etc.), impact-spot(s) on the vehicle, the impact type (full impact, grazing impact), the direction of motion of the parties and some more.

**Collision speed:**

In reference to passenger car - pedestrian accidents the collision speed $v_{\text{coll}}$ defines the vehicle speed (opponent of the pedestrian) in longitudinal direction at the time of collision.

**Advantages:**

- Correlates very well (especially at full frontal collisions) with impact intensity/severity and thus with injury severity (because of usually low pedestrian speed)
- Commonly available in the GIDAS-database (from reconstruction and simulation)
- Comprehensible parameter in general

**Disadvantages:**

- Does not consider pedestrian speed
- Does not describe energy input (especially relevant for grazing)

**Applicability as prediction variable:**

- Commonly used parameter for injury risk functions for pedestrians
- Most relevant parameter with sufficient availability for most accidents between passenger cars and pedestrians
- Limitations especially at very high and very low value area
- Appropriate for many applications (system assessment)

**Relative speed:**

Relative speed $v_{rel}$ describes the approach speed between both participants center of gravity at the time of collision (see Figure 4.3).

**Advantages:**

- Correlates very well (especially at full frontal collisions) with impact intensity/severity and thus with injury severity
- Commonly available in the GIDAS-database (from reconstruction and simulation)
- Comprehensible parameter in general
- Considers speed of both participants

**Disadvantage:**

- Does not describe energy input (especially relevant for grazing)
Applicability as prediction variable:

- Occasionally used for injury risk functions
- Available for most passenger car - pedestrian accidents (GIDAS-database)
- Appropriate for many applications (system assessment)
- More appropriate than collision speed for some special cases (e.g. stationary car)
- Very similar to passenger car’s collision speed due to low pedestrian speed

Difference of speed (delta-v) of the passenger car:

The difference of speed $\Delta v_{\text{car}}$ describes the vectorial change of speed (pre-crash to post-crash) of the passenger car during collision (see Figure 4.4).

Advantages:

- Considers speed and mass of each participant
- Correlates very well with impact intensity/severity in general and thus with injury severity (especially for vehicle occupants)
- Commonly available in the GIDAS-database

Disadvantage:

- The change of speed of the passenger car is very low at collisions with pedestrians due to low pedestrian mass (compared to the passenger car),
Applicability as prediction variable:

- Not appropriate due to large collision opponents’ mass difference
- Limited applicability for system assessment (renewed reconstruction necessary)

Difference of speed of the pedestrian:

The difference of speed $\Delta v_{ped}$ describes the vectorial change of speed (pre-crash to post-crash) of the pedestrian during collision (see Figure 4.4).

Advantages:

- Considers speed and mass of each participant and thus the energy input to the pedestrian
- Correlates very well with impact intensity/severity and thus with injury severity

Disadvantages:

- Uncertain/doubtful accuracy due to difficult determination of the parameter
- Rarely/not available in accident databases (trial record in GIDAS since 2010)
- Not commonly used parameter for people

Applicability as prediction variable:

- Not (yet) usable due to lack of availability in the database
- Parameter is not used for IRFs so far
- Very limited applicability for system assessment (hard to calculate in virtual cases (after system interference), renewed reconstruction necessary)

Impact type:

The impact type can distinguish the collision in full impact and grazing impact (more nominal characteristics possible). However, the distinction of the impact type is not sufficient as a sole parameter. In combination with another measure, like collision speed, it is promising, because it assists describing grazing impacts much better in the model. Grazing impacts mostly lead to significant rotatory component of the resulting motion of the pedestrian. This results in a significant low injury risk even at higher collision speed. However, there is no generally accepted definition of grazing and full impact.
Advantages:

- Distinction to different impact situations possible
- Consideration of rotatory components of the resulting pedestrian motion

Disadvantages:

- Not sufficient as sole measure
- No existing general definition
- Limits vary per vehicle model and can only be determined by expensive single case analysis

Applicability as prediction variable:

- Significant model improvements possible by consideration in combination with another collision parameter
- Definition must be set and considered at evaluation and discussion of the results

Further collision parameters:

There are further parameters that can describe the energy input and thus the collision severity. Those are measures such as change in pedestrian’s momentum $\Delta \vec{p}_{\text{ped}} = f(m_{\text{ped}}, \Delta \vec{v}_{\text{ped}})$, strain energy $U = f(m_{\text{ped}}, m_{\text{car}}, v_{\text{rel}}, \epsilon)$, kinetic energy $\Sigma E_{\text{kin}} = f(m_{\text{ped}}, m_{\text{car}}, v_{\text{ped}}, v_{\text{car}})$ et al, which additionally consider rotatory components of motions.

However, these measures are not yet or rarely included in existing databases. Calculations currently still lead to concerns regarding accuracy. Furthermore, applicability for system assessment is significantly harder in comparison to evaluations using collision speed etc. It must be mentioned that the greatest influence on energetic measures mostly results from quadratic part of the passenger car’s collision speed $v_{\text{coll}}$. Therefore, combined measures have limited higher significance. Considering an increasing availability in accident databases, such approaches might be promising for the future.

Vehicle parameters are parameters to describe the vehicle more accurate or rather limit the dataset regarding observed vehicles. This includes characteristic parameters such as vehicle age (in respect to constructive aspects, NCAP requirements etc.), vehicle category, mass at collision, car chassis (front contour design) and some more.
Vehicle age (year of vehicle model’s market introduction):

The vehicle age contributes to consider constructive circumstances regarding pedestrian-protection-related activities. Therefore, it considers these influences neither separate nor in dependence of model or manufacturer but based on the (assumed or observed) development process in general.

**Advantage:**

- Consideration of pedestrian-protection-related activities or legal regulations and requirements of consumer protection of a certain time

**Disadvantage:**

- Assumption of a generalized development of the vehicle marked, independent from individual models or manufacturers

**Applicability as prediction variable:**

- Normalization or standardization is necessary to consider vehicle age directly as prediction variable (interval scaled parameter)
- Indirect consideration through exclusion of older vehicle models leads to a significant dataset reduction
- This effect especially occurs at real accident databases, since they show a tendency towards older vehicle models or an age difference to the current vehicle fleet

**Vehicle contour:**

Using the vehicle contour directly considers constructive pedestrian-protection-related activities. Contrary to the vehicle age parameter there is a link to individual models. However, it is hard to make a clear definition or distinction between different contours.

Furthermore, there is a large number of possible influencing parameters (e.g. bonnet angle, height of bonnet leading edge and front bumper, windscreen angle, A-pillar width, rigidity of the windscreen etc.).
Advantage:

- Consideration of pedestrian protection relevant activities

Disadvantages:

- Difficult to make a clear definition
- Metric measures like stiffness rare or unknown
- Link of generalized statements about certain vehicle contours without concrete evidence

Applicability as prediction variable:

- Important influencing parameter for the injury risk of pedestrians in principle
- Use as prediction variable hardly possible due to difficult classification and unavailable data
- Use of vehicle age as summarized parameter leads to better applicability

Vehicle mass:

With the vehicle mass, an altered kinetic energy and inertia of the vehicle can be considered at the same speed. However, the influence of the mass variance is relatively small due to the significantly lower masses of the pedestrians.

Advantage:

- Consideration of the influence of the vehicle mass on the kinetic energy and inertia of the vehicle

Disadvantages:

- Only low influence due to the large mass difference between vehicle and pedestrian
- Less influence (linear) as collision speed (square)

Applicability as prediction variable:

- Can contribute to model quality but has a low influence

Individual characteristics of the pedestrian are parameters, which are supposed to describe the pedestrian more accurately. This includes characteristical measures such as age, gender, height, body weight, medical history and some more. The shape of the influence of such variables is much discussed and often not exactly known.
Age of the person:

With the age of the person, age-dependent and physiological effects such as body height, bone mineral density, musculature, responsiveness and more can be summarily displayed in the model. Especially the change of the physiological constitution of people over their age is a phenomenon, which should be considered at the prediction of injury severities. In particular, the physiological influence of elderly of about 70 years and above is considered to be very high. The influence of the age is often assumed to be linear and rarely to be squared. There are also other modeling shapes possible in order to regard uneven physical development about person’s age. Despite some controversy discussions about age groups and their limitations, there is an agreement to consider children and adults separately. However, there are different opinions on the specification of age groups.

Advantages:

- Consideration of age-dependent, physiological effects
- Indirect consideration of different impact kinematics due to the lower height of children

Disadvantage:

- Development not exactly known (linear is insufficient), difficult to integrate as model parameter

Applicability as prediction variable:

- Very strong influence (usually at all types of road participants, but especially at pedestrians), therefore a consideration is strongly recommended
- Significant influencing factor especially when considering fatally injured persons
- Development of the impact not exactly known
- Linear influence is too inaccurate, therefore a clustered consideration seems reasonable
- Still relevant regarding demographic change
Gender:

Using the gender of the person enables to consider, physiological influences regarding anatomic variations of the different gender.

Advantage:

- Consideration of gender dependent, physiological influences

Disadvantages:

- Only binary expression
- Assumption of constant physiological influences of each gender
- Influence is very different, depending on the problem

Applicability as prediction variable:

- Only low gender specific influence at pedestrian accidents
- Use not meaningful due to low relevance
- Influences can be considered through other individual parameters (height, body weight)

Body height and mass of the person:

Using body height and the body mass of a person can consider individual differences regarding anatomy of the person. For pedestrians body height influences the impact point (especially of the head) on the vehicle in single cases and therefore influences the injury severity.

Advantage:

- Consideration of different impact kinematics due to varying body height and mass

Disadvantages:

- Only low influence in “normal” range due to large mass difference between vehicle and pedestrian
- Impact point along the longitudinal axis (x-direction) of the passenger car is usually less relevant than the impact point in lateral direction (y-direction)

Applicability as prediction variable:

- Consideration of the anatomy can be useful depending on the problem
- Only low influence for accidents between passenger cars and pedestrians
- Poor applicability as input variable for a system assessment
There are further influencing parameter such as the direction of the pedestrian in relation to the vehicle (specifies from which direction the pedestrian gets closer to the vehicle from vehicle perspective), pre-existing injuries of the pedestrian, thickness of the clothes etc., but which do not have a significant influence on the overall model.

An iterative process performs the selection of the independent parameter for the statistical model by considering the problem and the model quality influenced through inclusion and exclusion of single parameter (see described process of the model formulation above).

The created injury risk function models for the application of pedestrian protection relevant safety features at accidents involving passenger cars and pedestrians including the injury severities selected in section 4.2 consider the following parameter:

- Impact type
- Vehicle age
- Age of the person
- Collision speed $v_{coll}$

The kind of consideration of the mentioned parameters within the overall model varies.

The report considers only full frontal accidents regarding impact type. This is implemented by the exclusion of grazing accidents in the master-dataset. Thus, there will be no consideration of impacts, including a significant rotatory part in the resulting motion of the pedestrian or pedestrian having very small overlap with the vehicle front. For the distinction between full- and grazing impact, there are various approaches in literature (position of first contact at the vehicle, subdivision of the body in seven parts of equal size, hit of the body center of gravity and others).

This report considers only full impact accidents for the created injury risk functions. First, it deals with all collisions containing pedestrians, where the collision occurred at the vehicle front (definition according to the second digit of the "Collision Deformation Characteristics (CDC, SAE J224)"). Afterwards it excludes all cases where the point of contact (first contact) on the vehicle is more than 30 cm in x-direction (reference is the vehicle front with $x = 0$). All impacts at the edges of the vehicle (point of contact between 0 and 30cm from the vehicle outside edge) are examined individually and grazing impacts are excluded. Also specific cases (pedestrians kneeing, lying or jumping onto the windshield) are excluded. Additionally there is no consideration of cases, where pedestrian exclusively suffers from vehicle caused injuries through overrun of lower extremities.
Consideration of vehicle age is useful, due to significant improvement of pedestrian protective systems of various vehicle models over time. In addition, the introduction of pedestrian-tests within the Euro NCAP since 1997 strongly increased the interest of the automobile manufacturers regarding pedestrian protection. However, previously mentioned disadvantages need to be taken into account at the decision to use vehicle age as prediction variable. Therefore, the injury risk functions do not model vehicle age directly as influencing parameter but indirectly through filtering of the dataset.

Figure 4.5 shows the distribution of year of vehicle model’s market introduction (JDME) of each passenger car in a frontal collision with a pedestrian in GIDAS. There is a significant reduction of cases observing exclusively newer vehicles. The report considers only JDME of 2000 and above, to fulfill the current requirements of a preferably homogeneous group (sum few years to a category) on the one hand and a preferably large amount of cases on the other hand.
As already mentioned above, a consideration of person’s age for accidents, involving pedestrians is highly recommended. The function process of the age influence is often supposed to be linear, which is strongly questioned in literature. An exact function process is unknown. However, in general it is assumed that the influence at adults has a small increase in a wide range and greatly increases for elderly. Therefore, the report does not directly consider the person’s age as prediction variable for injury risk functions, but through the creation of preferably homogeneous groups³. Each group models its individual curves.

³ Due to the difficult modeling of the age influence (non-linear, but not known), an agreement was made within the project. The type of modeling and the group boundaries are based on the publications by Niebuhr, Junge, Achmus / Rosén.
Literature cites various boundaries for the separation to groups. The classification of the used age groups is based on publications of Niebuhr, Junge, Achmus/Rosén and includes the following:

- Children 0 – 14 years
- Adults 15 – 59 years
- Elderly 60+ years

Figure 4.6 shows the sample size of each age group of the master-dataset. Previously set limitations regarding vehicles with JDME of 2000 and above leads to partly small group sizes, which seem to be critical for the calculation of injury risk functions and might not be sufficient. Unknown MAIS values are excluded if used.

<table>
<thead>
<tr>
<th>Age group</th>
<th>MAIS1</th>
<th>MAIS2+</th>
<th>MAIS3+</th>
<th>fatal</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 14 years</td>
<td>53</td>
<td>7</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>15 - 59 years</td>
<td>90</td>
<td>43</td>
<td>19</td>
<td>4</td>
</tr>
<tr>
<td>60+ years</td>
<td>34</td>
<td>41</td>
<td>21</td>
<td>9</td>
</tr>
<tr>
<td>In total</td>
<td>177</td>
<td>91</td>
<td>42</td>
<td>14</td>
</tr>
</tbody>
</table>

Injury severity

Figure 4.6 – Sample sizes of defined age groups of the master-dataset for pedestrians

The collision speed $v_{\text{coll}}$ is identified as the most important influencing parameter regarding injury severity of pedestrians at accidents of passenger cars and pedestrians. It especially correlates with the injury severity at full frontal impacts; it is the most commonly available parameter in GIDAS and is very applicable for system assessment. It is stored as an integer value in km/h in the GIDAS-database and results of a reconstruction performed for each accident. Figure 4.7 shows the distribution of the collision speed in the master-dataset. The majority of accidents involving pedestrians (80 %) takes place at collision speed up to 40 km/h. Therefore, the probability models have a better prediction accuracy in this area.
In summary, the following parameters are directly or indirectly taken into account in the model:

- Collision speed of the vehicle
- Age of the pedestrian
- Year of vehicle model’s market introduction (vehicle age)
- Impact type

4.4 Selection of the statistical model

As mentioned earlier, there is no stated definition of an approach for the creation of injury risk functions and therefore no mandatory statistical Model. Chapter 3 explains various possible statistical models and discusses regarding advantages, disadvantages and applicability. For the selection of an appropriate approach, it is important to consider the properties of the used data and to proof the legitimacy of model assumptions. Conditions A1 – A5 need to be involved in this consideration.

The possibility of a multivariate contemplation is important to model such complex contexts in the model occurring of questions regarding vehicle safety. Since the applicability for system assessment is the main incentive reason for injury risk functions, the resulting function should be available in an analytical form.
From the perspective of traffic accident research, the applicability and adaption on empirical real accident data is of fundamental importance. The distribution of individual parameter within the data is not exactly known. The mentioned model of the survival analysis fulfills all requirements A1 – A5 and commonly approved. Logistic regression is very similar to survival analysis and is even identical assuming a logistic distribution. This assumption leads to a missing zero crossing (see condition A2). It meets all other conditions and is displays the “dose-response-relationship” very well.

Therefore, the statistical model of the binary logistic regression can be recommended for accidents between passenger cars and pedestrians and is used for the calculation of the following curves.

4.5 Computation of injury risk functions

Based on the described approach for the creation of injury risk functions, models for the probability for MAIS2+ injured, MAIS3+ injured or fatally injured pedestrians in full frontal collisions with passenger cars can be created. Grazing impacts and special cases as well as collisions with vehicle models with an JDME older than 2000 were not taken into account. The mentioned models are created for children, adults and elderly as a function of the collision speed \( f(v_{\text{coll}}) \). As already shown in Figure 4.6, the underlying limitations and the subdivisions partly lead to small group sizes. Therefore, the report displays exclusively functions of sufficient group size, as they are:

- Adults, MAIS2+ injured
- Adults, MAIS3+ injured
- Elderly, MAIS2+ injured
- Elderly, MAIS3+ injured

in order to obtain reliable curves.
Adults:

Figure 4.8 shows the curves for the probability of adults to be MAIS2+ or MAIS3+-injured. The related functions are:

\[
p_{MAIS2^+}(v_{coll}) = \frac{1}{1 + \exp(3.016 - 0.079 \cdot v_{coll})} \tag{4.4}
\]

and

\[
p_{MAIS3^+}(v_{coll}) = \frac{1}{1 + \exp(4.848 - 0.088 \cdot v_{coll})} \tag{4.5}
\]

Figure 4.8 – Injury risk function for adult pedestrians in frontal passenger car-pedestrian collisions

The curves show the typical sigmoid shape for the logistic regression. Both have an offset at \( v_{coll} = 0 \ km/h \) due to the model assumption of a logistic distribution. The offset value results from cases including injured pedestrians through collisions with vehicles at low speed and with stopped or parked vehicles.
In general, the MAIS3+ curve is located “below” the MAIS2+ curve (see condition A5) within the whole definition area, so there is no intersection of both curves.

This is necessarily true, because a person, who is MAIS3+ injured will always be MAIS2+ injured by definition. Based on the regression coefficients (or the logit), the influence of the collision speed \( v_{\text{coll}} \) can be assessed. As expected, it is recognizable at both curves that higher collision speed increases the probability of being injured. In comparison, the influence at the MAIS3+ curve is higher than the one at the MAIS2+ curve. This results in a slightly more increasing MAIS3+ curve. At very high collision speed, both curves have hardly any or only a small difference between each other and asymptotically approximate to the probability of 100%.

Table 4.1 gives an overview of the injury risk function models’ quality for adult pedestrians and lists the measures described in section 3.6 for the evaluation of the model quality.

The specification of the number of persons, who are more easily injured than the corresponding criterion (or completely uninjured) and the complementary number of persons, who are more severe or equally injured than the corresponding criterion shows the size of the used sample and their proportion. The total number of the model for fatally injured persons is greater than the total number for the models of MAIS2+ and MAIS3+ injured persons. Despite same filter criteria, this results from different treatment of the injury severity (deceased/not deceased by official definition, MAIS resulting from the various AIS regions) and the existence of persons with unknown MAIS values, but known official injury severities in the dataset.

The ratio between considered persons that meet the criterion and those that do not already gives an indication of the model quality. For very small groups, the creation of a curve is not useful, since the model cannot be considered as reliable. As shown in Table 4.1, there are only four fatally injured compared to 134 non-fatally injured persons (adult pedestrians) in the dataset. As a result, there is no valid model that can be generated for fatally injured adult pedestrians in the age group of 15-59 years. The results are highlighted grey and Figure 4.8 omits the curve.
Table 4.1 – Model quality IRF pedestrians, adults (15-59 years)

<table>
<thead>
<tr>
<th></th>
<th>MAIS2+</th>
<th>MAIS3+</th>
<th>fatal</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>number</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; criterion</td>
<td>90</td>
<td>114</td>
<td>134</td>
</tr>
<tr>
<td>≥ criterion</td>
<td>43</td>
<td>19</td>
<td>4</td>
</tr>
<tr>
<td>total</td>
<td>133</td>
<td>133</td>
<td>138</td>
</tr>
<tr>
<td><strong>model quality</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>true negative</td>
<td>92 %</td>
<td>97 %</td>
<td>99 %</td>
</tr>
<tr>
<td>true positive</td>
<td>52 %</td>
<td>40 %</td>
<td>25 %</td>
</tr>
<tr>
<td>total classification</td>
<td>79 %</td>
<td>89 %</td>
<td>97 %</td>
</tr>
<tr>
<td>Cox-Snell-R²</td>
<td>0.268</td>
<td>0.249</td>
<td>0.115</td>
</tr>
<tr>
<td>Nagelkerkes-R²</td>
<td>0.373</td>
<td>0.437</td>
<td>0.499</td>
</tr>
<tr>
<td>AUC</td>
<td>0.722</td>
<td>0.687</td>
<td>0.621</td>
</tr>
</tbody>
</table>

The classification results (classification according to formula (3.14) and threshold of $p_{thres} = 50\%$) show the prediction accuracy of the models in relation to the observed data. All three models result in a very accurate prediction rate for true-negative-cases (correct predicted as not MAIS2+/MAIS3+/fatal), but only an average (MAIS2+, MAIS3+) or bad (fatal) prediction rate for the true-positive cases (correct predicted as MAIS2+/MAIS3+/fatal). While the MAIS2+ model predicts slightly more than 50 % of the persons correctly as MAIS2+ injured, the rate for the model of fatally injured persons is only 25 %. This corresponds to only one in four persons observed in the dataset, who were fatally injured.

The values of Cox-Snell-R² 0.268 and 0.249 refer to an “acceptable” model quality for MAIS2+ and MAIS3+ and to a rather poor model quality of the curve for fatally injured pedestrians. The values of Nagelkerkes-R² show a different image comparing the three models. While the MAIS2+ model results in a value of 0.373, the model for fatally injured persons reaches a much higher value of 0.499. This is especially caused by very high true negative ratio, which results from the low proportion of fatally injured persons. These are rated higher at Nagelkerkes-R².
The AUC values of the ROC support the classification results and the Cox-Snell-$R^2$ values. The MAIS2+ model reaches 0.722, meaning a very high level for models of pedestrian injury severity prediction. Also the MAIS3+ model reaches a high level with 0.687. The model for fatally injured persons (0.621) is still in a suitable range but comes closer to a purely random event (e.g. coin toss, AUC = 0.5). It should be evaluated critically; also with regard to the low amount of cases.

Since all quality measures have their characteristic advantages and disadvantages, several criteria should be considered together for the evaluation of the model quality. The evaluation based on individual values has only a limited significance.

In summary, all model quality parameters certify reliable curves for the created models for the probability of a pedestrian being MAIS2+ or MAIS3+ injured at full frontal collisions with a passenger car. However, the model for the prediction of fatal injuries for pedestrians cannot be regarded as reliable. Especially the low amount of fatally injured persons in the sample is an exclusion criterion.
Elderly:

Figure 4.9 shows the injury risk functions for MAIS2+ or MAIS3+ injured pedestrians of elderly persons (59+ years). Related functions are:

\[
p_{\text{MAIS}2+}(v_{\text{coll}}) = \frac{1}{1 + \exp(2.233 - 0.113 \cdot v_{\text{coll}})}
\]  

(4.6)

\[
p_{\text{MAIS}3+}(v_{\text{coll}}) = \frac{1}{1 + \exp(2.976 - 0.076 \cdot v_{\text{coll}})}
\]  

(4.7)

Figure 4.9 – Injury risk function for elderly pedestrians in full frontal collisions with passenger cars

The curves in Figure 4.8 show the sigmoid shape and an offset from the point of origin. In comparison, the offset is much more significant than for adults (15-59 years). This is plausible, because real accidents show that elderly easily loose balance at very low collision speed, fall on the street and receive more severe injuries due to a lack of defense reactions. The physiological constitution, which becomes worse in old age, additionally intensifies such effect. Therefore, very low collision
speed or even stopped vehicles \(v_{coll} = 0 \text{ km/h}\) at the moment of collision can lead to (partly severe) injuries. The curves for MAIS2+ and MAIS3+ injured elderly also show no intersection.

The influence of the collision speed for the MAIS3+ model with a regression coefficient of 0.076 is located in a similar range as for the adult models (15-59 years). The coefficient for the MAIS2+ model at 0.113 is much higher, what leads to a steeper curve. This results to a probability for getting MAIS2+ injured of close to 100 % already for medium collision speed. This reflects the significantly increased vulnerability of elderly. It confirms the necessity of modeling the age influence or alternatively the subdivision of the sample into three homogeneous age groups.

Table 4.2 gives an overview of the quality of the models of the injury risk functions for elderly pedestrians and lists the measures for the evaluation of the model quality, described in section 3.6.

**Table 4.2 – Model quality IRF pedestrian, elderly (60+ years)**

<table>
<thead>
<tr>
<th></th>
<th>MAIS2+</th>
<th>MAIS3+</th>
<th>fatal</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; criterion</td>
<td>34</td>
<td>54</td>
<td>74</td>
</tr>
<tr>
<td>≥ criterion</td>
<td>41</td>
<td>21</td>
<td>9</td>
</tr>
<tr>
<td>total</td>
<td>75</td>
<td>75</td>
<td>83</td>
</tr>
<tr>
<td><strong>model quality</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>true negative</td>
<td>76 %</td>
<td>85 %</td>
<td>99 %</td>
</tr>
<tr>
<td>true positive</td>
<td>84 %</td>
<td>45 %</td>
<td>11 %</td>
</tr>
<tr>
<td>total classification</td>
<td>80 %</td>
<td>74 %</td>
<td>89 %</td>
</tr>
<tr>
<td>Cox-Snell-R²</td>
<td>0.315</td>
<td>0.201</td>
<td>0.101</td>
</tr>
<tr>
<td>Nagelkerkes-R²</td>
<td>0.422</td>
<td>0.287</td>
<td>0.204</td>
</tr>
<tr>
<td>AUC</td>
<td>0.797</td>
<td>0.653</td>
<td>0.549</td>
</tr>
</tbody>
</table>
The sample size of 75 (MAIS known) or 83 (official injury severity known) persons is smaller than the one for the above-mentioned models for adults. However, the proportion of persons, who meet the corresponding injury criterion, is much higher (proportion of MAIS3+ equals 28%, proportion of fatal injured equals 11%), which tends to have a positive impact on the model quality.

The MAIS2+ and the MAIS3+ model show good classification results. On the other hand, the model for fatally injured pedestrians meets a rather bad classification. These results arise as high true negative rate, but very low true positive rate, analogous to the model for fatally injured adults (15 – 59 years).

The MAIS2+ model shows very good Cox-Snell-R$^2$ and Nagelkerkes-R$^2$. Also the MAIS3+ model is still within an acceptable range. Both R$^2$ values for fatally injured pedestrians on the other hand seem to be very low and indicate a lower model quality.

The AUC of the ROC shows similar tendency between the three models compared to classification results and R$^2$ values. The models for MAIS2+ and MAIS3+ injured persons are very good respectively acceptable, whereas the model for fatally injured pedestrians is really close to the value of 0.5 for a random event. This suggests a very limited significance.

In summary, the common consideration of the several model quality measures for the created MAIS 2+ and MAIS3+ curves leads to a rating as valid and reliable models. These can be used for the calculation of the injury probability of an elderly pedestrian at a full frontal collision with a passenger car. The model for fatally injured pedestrians is based on a small number of fatally injured elderly, which results to a very low model quality. The use of this model for any probability statement of an elderly pedestrian getting fatally injured needs to be regarded as critical. Therefore, the model is not used and the report omits displaying the related injury risk function.

**Children:**

Table 4.3 shows the number of observed children (age of 0 – 14 years) per injury criterion in the sample and the measures for evaluating the MAIS2+ model quality. The models are based on less people than the former models (adults and elderly) as they refer to 60 or 67 cases in the dataset. Additionally, the sample contains only a very low number of children meeting the related criterion ($n_{\text{MAIS}2+} = 7, n_{\text{MAIS}3+} = 2, n_{\text{fatal}} = 1$). Therefore, the creation of a model for MAIS2+ injured children needs to be regarded as critical. For MAIS3+ injured and fatally injured children it is not useful at all.
The classification results for the MAIS2+ model show the maximum possible amount of all true negatives, but only a very low rate of true positives. That means, all observed children who are not MAIS2+ injured, but only 14 % of children who are MAIS2+ injured can be correctly predicted by the model (threshold $P_{\text{thres}} = 50 \%$). Using Cox-Snell-R$^2$ and Nagelkerkes-R$^2$, both show a very low model quality. In addition, the AUC value of the ROC is low and indicates that the model is close to a random model.

All model quality measures indicate that the calculated injury risk function does not allow reliable statements. This is primarily due to the low number of cases meeting the injury criterion. Therefore, the models for children are not used and the report omits all injury risk functions for children. The creation of valid models requires an increase of the dataset (e.g. through lowering the exclusion threshold for JDME).

**Table 4.3 – Model quality IRF pedestrians, children (0 - 14 years)**

<table>
<thead>
<tr>
<th></th>
<th>MAIS2+</th>
<th>MAIS3+</th>
<th>fatal</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; criterion</td>
<td>53</td>
<td>58</td>
<td>66</td>
</tr>
<tr>
<td>≥ criterion</td>
<td>7</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>total</td>
<td>60</td>
<td>60</td>
<td>67</td>
</tr>
<tr>
<td><strong>model quality</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>true negative</td>
<td>100 %</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>true positive</td>
<td>14 %</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>total classification</td>
<td>90 %</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Cox-Snell-R$^2$</td>
<td>0.081</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Nagelkerkes-R$^2$</td>
<td>0.158</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>AUC</td>
<td>0.571</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>
4.6 Discussion

Despite an increasing requirement of appropriate injury risk functions in the field of vehicle safety, there exists no predefined or standardized creation procedure. Additionally, there are only a few injury risk function related publications (mostly probability for fatally injured people in dependence to collision speed). And these come to different results even when using identical/similar datasets (as shown from Rosén and Sander [41] et al). Therefore, a critical reflection of the model, assumptions and the description of limitations is substantial.

Dataset / real accident data

The selection and characteristics of the dataset, as bases for the injury risk functions, is of high significance. The field of vehicle safety has a preference on real accident data to calculate injury severity predictions based on empirically investigated data. Treatment of the dataset is important to evaluate applicability of the data source in relation to the requested objective. Also important is the consideration of the influence of the case selection towards location and shape of the injury risk function.

The present report uses the real accident database GIDAS for the creation of injury risk functions. In general, this database is regarded as representative for the German accident scenario (applying a weighting procedure). Due to the GIDAS project specification of exclusive investigation of traffic accidents including personal injury, there is a lack of accidents with unharmed people (accident including only material damage). There is no knowledge about impact severity/collision speed threshold for uninjured pedestrians. There is also no reliable information on how many accidents occur between passenger cars and pedestrians, where the pedestrian remains uninjured.

The estimated number of unregistered traffic accidents of such impact constellation is presumably high, if no injuries occur and the vehicle does not receive any damage. Thus, there is usually a systematic overestimation of the injury probability with injury risk functions, based on GIDAS and exclusively injured persons. This overestimation especially takes place at very low collision speed (up to 10 km/h).

Dependent variable

At the beginning of the creation of the model, there is the question of which injury severity criterion should be used and its impact on the quality of the injury risk function. Thus, injury severity according to official definition involves the issue of hospitalization duration. If a person remains in the hospital for stationary observation, the classification is seriously injured regardless of the real injuries. The MAIS with six
classes of injuries is a better measure for the total injury severity. It results from the most severe AIS value of all single injuries of the person. However, in most cases the MAIS is not available in national traffic accident statistics (e.g. Fachserie 8 / Reihe 7 of the German Federal Statistical Office). As special characteristic, the AIS is not metric and thus the single values can hardly be set in direct relation to each other. In addition, the attempt to include the variety of injuries in a uniform scale level involves critical aspects. These are not undisputed among experts. For instance, a traumatic brain injury (TBI) of first degree (without unconsciousness) is classified as AIS value of 2 in the AIS version of 1990, Revision 1998. In contrast, in the more recent version of the AIS from 2005, Update 2008, assign an AIS value of 1. The downgrade attempts to encounter the too often, sometimes unjustified, rating of the TBI of the first degree. Since a TBI of the first degree occurs very frequently due to a head impact on the vehicle front or the road surface for pedestrians, and this is often the most serious single injury according to the AIS scale, there is a relevant mismatch between the two AIS versions. In particular when evaluating MAIS2+ injury probabilities, there might be a large difference in the group sizes depending on the choice on the AIS version. This results in differing injury risk functions.

Independent variable

Next to the injury severity criterion, there is the important question for appropriate parameters for the injury severity prediction. Therefore, section 4.3 addresses a variety of potentially possible parameters for this purpose. The more parameters the model includes, the more accurate might be the injury severity prediction. However, the effect of a dataset trained model occurs quickly by including many prediction variables. This is called “overfitting”, whereby it is very good predicting the model dataset, but no general statements can be made anymore. It is therefore important to consider the AIC or BIC (see section 3.6) for the selection of the included parameter. Beside the collision speed, the report identifies the age of the pedestrian and the impact type as relevant and uses them for the models. However, they are only considered indirectly through the creation of age groups and filtering the dataset on full impacts only.
Additional, the year of vehicle model’s market introduction is considered indirectly. Thereby, the injury risk functions refer to the current vehicle fleet as much as possible and therefore to the state of the art of current pedestrian protection activities. Type and function shape of these three parameters’ influence must not be modeled directly. This especially avoids the challenges of the pedestrian age, described in section 4.3.

The collision speed of the passenger car contains the highest power for the prediction of pedestrian's injury risk in a full frontal collision with a passenger car (see section 4.3). It includes a limited accuracy and tolerance range, due to the determination at the works of accident reconstruction. The GIDAS database lists the accuracy of the collision speed with 1 km/h and sets the tolerance range to approximately 5 km/h. The iterative determination during reconstruction and the tolerance range leads to accumulations at certain values (human preference of certain values, see Figure 4.7). The figure also shows that the distribution in the dataset is not uniform over the entire definition range. Lower collision speed occurs more frequently in real accidents (analogous to the entire traffic scenario). These characteristics influence the results of the models. The location and function shape of the logistic curve is mainly influenced by the large number of cases with low speed values. Statements on very high collision speed, which occur only rarely in real accident data, have lower confidence.

The age of the pedestrian is also a significantly influencing parameter for the injury severity of pedestrians. It includes several direct and indirect physiological influences, such as body height, bone density, muscular system or the ability to react. However, the function shape of the age influence is not exactly known and differs for each individual. The assumption of a linear, quadratic or similar approach needs to be seen critically, since such approach implies a continuous increase in the injury probability over the human lifetime. However, it is considered to be proven that children have a significantly different physical structure compared to adults. Older people up to a certain age show a noticeable decline in their physiological constitution. The influence of the age seems almost constant between these two points. For this reason, the influence of the age is taken into account by generating preferably homogeneous groups for the created models. They assume no age influence within these groups. The age thresholds cannot be clearly determined and orientate on Niebur, Junge, Achmus/Rosén in combination with the resulting case number in the sample within each group. The age limit for elderly is ≥ 60 years. A limit of ≥ 70 years might be the better choice regarding the start of physiological effects. But as it leads to very small case numbers, the resulting model quality for the injury risk functions become very low.
The creation of age groups produces a sharp delimitation. This does not correspond to the real influence. But it is a useful model assumption in consideration of the unknown real functional link.

The exclusion of grazing impacts intends the creation of a sample of accidents as homogenous as possible. There passenger cars hit pedestrians full frontal and picked them up completely or partly. However, this does not automatically mean to include all impacts addressing pedestrian protection relevant regions at the vehicle.

By exclusion of side- and grazing impacts also excludes occasional impacts in the area of the lower A-pillar or the windscreen area. This aspect is accepted in favor of an improved model quality. It also eliminates cases with a major part of the translational energy being converted into rotational energy where high a collision speed often causes minor injury severities.

The exclusive consideration of vehicles models with JDME of 2000 and above performs injury risk functions for the most current models with a certain amount of pedestrian protection systems (see section 4.3).

**Statistical model**

The selection of the statistical model is the central question in the modeling process of injury risk functions. In general, there is not one best model existing. Different models have different advantages and disadvantages depending on the application.

Survival analysis and logistic regression (corresponds to a survival analysis with a logistical distribution assumption) have often been used for the creation of injury risk functions in the past. Despite their limitations (see chapter 3.2 and 0) they have been proven as appropriate models (see [6], [21], [22], [25], [37], [38], [41], [40] among others). For both, the knowledge respectively the reasonable assumption of a distribution is decisive. However, the distribution of the independent variable for the injury risk based on real accident data is unknown. The selection of the logistic regression assumes a logistical distribution. In general, the verification of the applied distribution assumption should be considered.

An advantage among the logistic regression and others is the orientation on empirical data, which should always be preferred to phenomenological approaches.
Secondary impact

Pedestrian - passenger car accidents frequently result in an impact of the pedestrian with the road surface / curb or more rarely in a collision with other objects in the traffic or side area as a consequence of the primary collision with the passenger car. Such impact is called "secondary impact". It significantly affects the overall injury severity of pedestrians and cyclists [26]. The most common impact area is the road surface and most injuries occur in the area of the head and the upper and lower extremities. Compared to the primary impact, the secondary impact is less dependent on the collision speed. Even very low collision speed (particularly older pedestrians) often result in severe injuries because of the secondary impact.

Pedestrian protection relevant activities on the vehicles can only address the primary impact of the pedestrian on the vehicle. Most constructive parameters, such as the length of the bonnet, the angle of the windshield or the height of the bumper do not significantly affect the occurrence of a secondary impact as well as its injury severity (see Kühn et al. [23]).

Only height of the bonnet’s front edge is described as an influencing factor by Kühn [23]. This is plausible with regard to the kinematics (pick up versus throwing). This paragraph considers the importance of the secondary impact in a full frontal collision with a pedestrian and the dimension of the difference between vehicle induced injuries and secondary collision induced injuries. For this purpose it distinguishes all single injuries from the GIDAS database according to their cause of injury and defines two new parameters:

- \( \text{MAIS}_{\text{car}} \) = maximum single injury severity (according to AIS) of all injuries caused by the vehicle
- \( \text{MAIS}_{\text{sec}} \) = maximum single injury severity (according to AIS) of all injuries caused by the secondary collision
The assignment of injury severity caused by the passenger car or secondary impact is possible for 265 of 288 (whole sample) injured pedestrians with frontal impacts on passenger cars JDME 2000 and later. This estimation performs the interdisciplinary investigation team on the basis of all available information (impact points / damage to the vehicle, direction of movement of the pedestrian to the vehicle, type / location / severity of single injury, body height, pedestrian kinematics, etc.). However, it is subjected to some uncertainty. The 265 cases in the dataset are represented as follows:

- 46% - MAIS determined only from the vehicle 
  \( \text{MAIS} = \text{MAIS}_{\text{car}}, \text{MAIS} > \text{MAIS}_{\text{sec}} \)
- 21% - MAIS determined only from the secondary impact 
  \( \text{MAIS} > \text{MAIS}_{\text{car}}, \text{MAIS} = \text{MAIS}_{\text{sec}} \)
- 33% - MAIS determined by impact with vehicle and secondary impact 
  \( \text{MAIS} = \text{MAIS}_{\text{car}} = \text{MAIS}_{\text{sec}} \)

Thus in 79% of the cases entire MAIS equals the vehicle induced MAIS (MAIS_{car}). In 21% of the cases the secondary impact exclusively influenced the MAIS. Appendix D mentions respective specifications for further injury severity degrees (MAIS2+, MAIS3+).

A second injury risk function is created for the MAIS_{car}, which only represents the risk of injury by the impact on the vehicle. However, the occurrence and severity of a secondary impact cannot be used as a prediction variable because this is a consequence of the impact. It is strongly dependent on situational conditions (e.g. head impact on the road, speed and angle at the impact, etc.). In the time before the collision, it is not known whether a secondary impact will take place or how this impact will occur. This also explains the significantly lower values for the model quality measures of injury risk functions for pedestrians compared to injury risk functions for occupants of passenger cars. The secondary impact is definitely a very important factor, which affects the pedestrian injury severity. Unfortunately it cannot be statistically predicted. For this purpose, the simulation of single cases is necessary. Another rather theoretical possibility is to consider the information "secondary impact yes/no" as a binary independent parameter in the statistical model. However, this has practically no benefit for any system assessment.

Figure 4.10 shows the comparison of injury risk functions (injury probability for MAIS2+) with and without a secondary impact for adults (Figure 4.11 for elderly). The difference between the two curves is very small for adults, but becomes more apparent in the range of low collision speed. This is in accordance with the findings of Ashton [3].
For elderly the difference between the two curves is more pronounced. This shows that elderly persons are more likely to suffer severe injuries from secondary impacts. Reasons for that are the increased vulnerability in old age and the diminished or worse responsiveness. The benefit of passive pedestrian protection measures on the vehicle is therefore reduced for this age group against other age groups.
Figure 4.10 – Comparison of IRF (MAIS2+) with and without a secondary impact for adults

Figure 4.11 – Comparison of IRF (MAIS2+) with and without a secondary impact for elderly
4.7 Limitations and user references

Quantitative statements on the probability of injury of pedestrians with the help of the created functions are only useful in the context of statistical considerations and for large number of cases. The application for single cases as well as derivation of any specific injury severity ("At a collision speed of 40 km/h, a pedestrian suffers from an injury severity of MAIS3.") is out of question. Furthermore, the curves cannot be used to deduce consequences of speed reductions in any single case ("If the passenger car had hit the pedestrian with x km/h less, he would not have been died.")

In the context of evaluations, the previously made model assumptions and described limitations should be taken into account and cited if necessary. These include:

- The present functions do not apply to lying or kneeling pedestrians as well as other special cases (e.g. fall / jump onto the vehicle).
- The functions represent the injury probability for the full frontal impact. They do not apply to other collision constellations (side-, grazing-, rear impact).
- Even in the case of a full impact, there is certain dispersion over the vehicle width. Thus, the individual injury severity is significantly dependent on the impact point. The present injury risk functions do not consider such effect but represents an average injury probability.
- The created functions apply for passenger car models with JDME 2000 to 2013, with a large proportion of the vehicles from 2000 – 2005. Thus, they only apply to passenger car models without pop-up-bonnets and external airbags.

In addition, real accident data always has some restrictions regarding currency. The databases reflect the results of current and past vehicle generations. Thus, the benefit of new safety systems can only be determined retrospectively after a certain period of market penetration.

However, it is possible to assess potentials on the basis of simulation or by using certain assumptions and models. A pessimistic approach is always required from an engineer’s responsible perspective (see [9], [13], [12]).
5 Creation of injury risk functions for cyclists

The assessment of safety related vehicle systems is also very interested in injury risk models for cyclists in traffic accidents. Analogous to the procedure for passenger car - pedestrian accidents (see chapter 4), statistical models for injury risk functions are created based on the real accident database GIDAS (German In-Depth Accident Study) for passenger cars – cyclists accidents.

It is important to consider important differences such as:

- Modified impact kinematics (Usually greater throw distance, head impact at higher wrap distance, higher position (z-value) of the lower extremities)
- Higher mass of the cyclist (including bicycle)
- Higher speed of the cyclist
- Grazing impacts are more difficult to define
- Partially no contact between passenger car and bicycle rider
- Protection of the head due to helmet use

Selection and filtering of the dataset:

The available dataset for the calculation of injury risk functions is generated for accidents between passenger cars and cyclists as follows (n = sample size = number of accidents):

- GIDAS sample (07/2015): n = 27,051
- First collision car – cyclist: n = 5,019
- Only full frontal impact: n = 3,220
- Year of vehicle model’s market introduction off 2000: n = 911
- MAIS, age und $V_{rel}$ available: n = 775

Figure 5.1 – Sample for accidents between passenger cars and cyclists from GIDAS

For cyclists, no outlier diagnostics or single case checks is performed. In collisions between passenger cars and bicycles, there is much more variance in the impact constellations and the origin of injuries. The most important aspect is the partial
missing contact between the car and the cyclist when only the bike collides with the car (usually at the front or rear wheel).

Selection of the dependent variable

The creation of the injury risk functions uses the following dependent variables:

- **MAIS2+**
  \[
  y_{MAIS2+} = \begin{cases} 
  1 & \text{for } MAIS \geq 2 \\
  0 & \text{for } MAIS < 2 
  \end{cases}
  \]  \hspace{1cm} (5.1)

- **MAIS3+**
  \[
  y_{MAIS3+} = \begin{cases} 
  1 & \text{for } MAIS \geq 3 \\
  0 & \text{for } MAIS < 3 
  \end{cases}
  \]  \hspace{1cm} (5.2)

- **Deceased (official definition)**
  \[
  y_{fatal} = \begin{cases} 
  1 & \text{for } PVERL = "fatal" \\
  0 & \text{for } PVERL \neq "fatal" 
  \end{cases}
  \]  \hspace{1cm} (5.3)

The evaluation of the injury severity according to AIS uses the AIS catalog of 2005, update 2008.

Selection of the independent variable

Analogous procedure compared to pedestrians (see section 4.3) in terms of as most homogeneous groups:

- Relative speed
- Age of the cyclist
  - Children 0 – 14 years
  - Adults 15 – 59 years
  - Elderly 60+ years
- Age of the vehicle model: JDME ≥ 2000
- Impact type: full frontal

The model taken the higher speed of the cyclist compared to the pedestrian into account by using relative speed \( v_{rel} \) (see section 4.3) instead of the collision speed. Figure 5.2 shows the distribution of the relative speed in GIDAS. Most of the accidents involving cyclists take place at relative speed of up to 30 km/h. Thus, the probability models in this region have a better prediction accuracy.
Figure 5.2 – Distribution of the relative speed in the master dataset

Figure 5.3 shows case numbers of the divided age groups of the master dataset for accidents involving passenger cars and cyclists. The limitation of vehicles with JDME of 2000 and above leads to small group sizes, which are critical or not sufficient for the calculation of injury risk functions. Using the MAIS, unknown values are excluded.

<table>
<thead>
<tr>
<th>Age group</th>
<th>MAIS1</th>
<th>MAIS2+</th>
<th>MAIS3+</th>
<th>fatal</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 14 years</td>
<td>57</td>
<td>10</td>
<td>4</td>
<td>-</td>
</tr>
<tr>
<td>15 - 59 years</td>
<td>469</td>
<td>81</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>60+ years</td>
<td>116</td>
<td>42</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>In total</td>
<td>57</td>
<td>93</td>
<td>43</td>
<td>2</td>
</tr>
</tbody>
</table>

Figure 5.3 – Case numbers of the defined age groups in the master dataset for cyclists

A minor influence of the age of the vehicle model is measureable, but has less effect on the model quality than accidents involving passenger cars and pedestrians. This is plausible since many cyclists do not hit the passenger car during the collision phase, and therefore the injury severity is less dependent from the vehicle front contour design as well as possible pedestrian protection activities. Figure 5.4 shows the
distribution of JDME of all passenger cars in frontal cyclist impacts in GIDAS. The procedure for accidents involving passenger car and pedestrian (see section 4.3) for the consideration of the vehicle age also applies to passenger car-cyclist accidents. Only vehicle models introduced into the market from the year 2000, are considered.

Figure 5.4 – Distribution of the year of vehicle model’s market introduction (JDME) of passenger cars in frontal accidents with cyclists in GIDAS

**Selection of the statistical model:**

The statistical model of the binary logistic regression for the calculation of the injury risk functions for accidents involving passenger cars and cyclists is selected analogous to passenger cars - pedestrian accidents (see section 4.4).

**Calculation of the injury risk function:**

Based on the described procedure for the creation of injury risk functions, models for the probability of cyclists to be MAIS2+ injured, MAIS3+ injured or fatally injured in full frontal collisions with passenger cars are created. They do not take grazing impacts, special cases as well as collisions with passenger car models introduced into the market before 2000 into account. These models are created for children, adults and elderly as a function of the relative speed $f(v_{rel})$. The limitations and subdivi-
sions lead to partly small groups sizes as already shown in Figure 5.3 and Figure 4.6.

**Adult cyclists (15 – 59 years):**

Table 5.1 gives an overview over model quality of the injury risk functions for adult cyclists and lists the measures described in section 3.6 for assessing the model’s quality. The number of persons that meet the respective injury criterion differs significantly from those who are more slightly injured and serve as the basis for the creation of the models. Thus, 81 MAIS2+ injured cyclists compared to 469 more slightly injured persons constitute a good group size ratio, whereas 15 compared to 535 persons for the MAIS3+ model represents a poor group size ratio. This results to an expectedly low model quality. The fatally injured cyclists’ model includes only one deceased person (compared to 549 non-fatally injured) in the sample for the adult age group (15 – 59 years). For this reason, it is not reasonable creating a model.

The consideration of the model quality measures (Cox-Snell-R², Nagelkerken-R², AUC of the ROC) of the created models shows very low values and suggests that the models cannot sufficiently describe the injury mechanism or the injury criteria just based on the independent variable (relative speed). Due to the low values of the model quality measures, the report omits a representation of the functions.
Table 5.1 – Model quality IRF cyclists, adults (15 – 59 years)

<table>
<thead>
<tr>
<th></th>
<th>MAIS2+</th>
<th>MAIS3+</th>
<th>fatal</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>number</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; criterion</td>
<td>469</td>
<td>535</td>
<td>549</td>
</tr>
<tr>
<td>≥ criterion</td>
<td>81</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td>550</td>
<td>550</td>
<td>550</td>
</tr>
<tr>
<td><strong>model quality</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>true negative</td>
<td>100 %</td>
<td>100 %</td>
<td>---</td>
</tr>
<tr>
<td>true positive</td>
<td>1 %</td>
<td>0 %</td>
<td>---</td>
</tr>
<tr>
<td>total classification</td>
<td>85 %</td>
<td>97 %</td>
<td>---</td>
</tr>
<tr>
<td>Cox-Snell-$R^2$</td>
<td>0.024</td>
<td>0.023</td>
<td>---</td>
</tr>
<tr>
<td>Nagelkerkes-$R^2$</td>
<td>0.043</td>
<td>0.104</td>
<td>---</td>
</tr>
<tr>
<td>AUC</td>
<td>0.506</td>
<td>0.500</td>
<td>---</td>
</tr>
</tbody>
</table>

**Elderly cyclists:**

Table 5.2 gives an overview of the quality of the injury risk function models for elderly cyclists and lists the variables described in section 3.6 for assessing the model quality. These models also show a very low case number of MAIS3+ injured or fatally injured cyclists. Consideration of all model quality values (Cox-Snell-$R^2$, Nagelkerkes-$R^2$, AUC of the ROC) indicates low model quality. This confirms the indication of concerns regarding analogously transfer of the pedestrian model procedure. The report omits the presentation of these functions due to the low model quality values.
Table 5.2 – Model quality IRF cyclists, elderly (60+ years)

<table>
<thead>
<tr>
<th></th>
<th>MAIS2+</th>
<th>MAIS3+</th>
<th>fatal</th>
</tr>
</thead>
<tbody>
<tr>
<td>number</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; criterion</td>
<td>116</td>
<td>147</td>
<td>157</td>
</tr>
<tr>
<td>≥ criterion</td>
<td>42</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>total</td>
<td>158</td>
<td>158</td>
<td>158</td>
</tr>
<tr>
<td>model quality</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>true negative</td>
<td>98 %</td>
<td>99 %</td>
<td>---</td>
</tr>
<tr>
<td>true positive</td>
<td>7 %</td>
<td>9 %</td>
<td>---</td>
</tr>
<tr>
<td>total classification</td>
<td>74 %</td>
<td>93 %</td>
<td>---</td>
</tr>
<tr>
<td>Cox-Snell-$R^2$</td>
<td>0.043</td>
<td>0.054</td>
<td>---</td>
</tr>
<tr>
<td>Nagelkerkes-$R^2$</td>
<td>0.063</td>
<td>0.136</td>
<td>---</td>
</tr>
<tr>
<td>AUC</td>
<td>0.527</td>
<td>0.542</td>
<td>---</td>
</tr>
</tbody>
</table>

Children:

Table 5.3 gives an overview of the model qualities of injury risk functions for children and lists the variables described in section 3.6 for the evaluation of the model quality. The accident rates of children on bicycles in road traffic is generally much lower than in the previous models. The number of MAIS2+ and MAIS3+ injured cyclists of the age group 0 – 14 years is very small and there is no fatally injured child within the sample. The model quality of the MAIS2+ injured children shows better values than the models for adults and elderly. However, it is still not in an acceptable range for a reliable assessment of the injury risk of cyclists in collisions with passenger cars. Due to the low values of model quality measures the report omits a representation of the functions.
Table 5.3 – Model quality IRF cyclists, children (0 – 14 years)

<table>
<thead>
<tr>
<th></th>
<th>MAIS2+</th>
<th>MAIS3+</th>
<th>fatal</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>number</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; criterion</td>
<td>57</td>
<td>63</td>
<td>67</td>
</tr>
<tr>
<td>≥ criterion</td>
<td>10</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td>67</td>
<td>67</td>
<td>67</td>
</tr>
<tr>
<td><strong>model quality</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>true negative</td>
<td>98 %</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>true positive</td>
<td>20 %</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>total classification</td>
<td>87 %</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Cox-Snell-R²</td>
<td>0.178</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Nagelkerkes-R²</td>
<td>0.313</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td><strong>AUC</strong></td>
<td>0.591</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>

**Discussion:**

**Relative speed:**

The relative speed between passenger car and bicycle has a certain explanatory power for the injury severity of cyclists. Due to the determination within accident reconstructions, it is subjected to a limited accuracy and a tolerance range.

The accuracy of the relative speed is 1 km/h and the tolerance range is specified at approximately 5 km/h in the GIDAS database, which already corresponds to a very good reconstruction. The iterative determination during the reconstruction and the tolerance range result in accumulations of certain values (human preference of certain values, see Figure 5.2). The figure also shows that the distribution in the dataset is not uniform over the entire definition range. Smaller relative speed occurs more frequently in the event of a real accident (analogous to the entire traffic scenario). This creates a right-skewed distribution.
**Full frontal impact:**

At collisions between passenger cars and cyclists, there is considerably more variance in impact constellations and injury formations than in accidents with pedestrians. Therefore, the exact definition of a full frontal impact is difficult. This also results in the difficulty of predicting the injury severity of the cyclist. In contrast to the pedestrian, it is not sufficient to consider all impacts in the area of the vehicle front excluding the outside areas in order to define a full impact. A further distinction, at least to the groups "only impact of the bicycle on the passenger car" as well as "overlap of the cyclist with the vehicle front" seems reasonable at this point.

**Modelling / model quality:**

The calculated models for injury risk functions for cyclists in full frontal collisions with passenger cars show poor values of the model quality measures and cannot be used for reliable statements. It is not advisable to transfer the shown approach for passenger car - pedestrian accidents analogously to passenger car - bicycle accidents. The presented models give strong indications that they do not adequately describe the occurring injury mechanisms, since the heterogeneity of the impact constellations is still too large even in the case of exclusive frontal impacts (on the passenger car). Further investigations of the modeling are mandatory in order to determine additional influences and to improve the models. This is, for example, the use of a bicycle helmet, although this only addresses a specific body region of the cyclist.

Two real examples (see appendix C) show the possible wide range of passenger car – cyclist accidents (with frontal impact on the passenger car). This demonstrates the difficulty for a reliable prediction of the injury severity. One approach for a model improvement is the creation of are more homogeneous subgroups regarding collision position, impact point, overlap (exact definition of full frontal impact required) and others. This should be done with regard to the remaining case numbers.

The influence of the JDME is significantly lower for accidents between passenger cars and cyclists than in the case of passenger cars and pedestrians. A restriction could possibly be renounced in favor of larger case numbers.

The modeling of the injury severity by the relative speed does not appear to be sufficient for all accidents between passenger cars and bicycles, since high relative speed can also result in low injury severities (see example 1 in appendix C) and low relative speed can also result in high injury severities (see example 2 in appendix C). Further possibilities for a more accurate representation of the input kinetic energies are given by the distinction in the ego-speed of the vehicle \( v_{car} \) and the speed of the
bike $v_{cyc}$ or the distinction of the relative speed in longitudinal $v_{rel,x}$ and lateral direction $v_{rel,y}$.

**Secondary impact:**

The secondary impact has an even greater impact on accidents between passenger cars and bicycles than on accidents between passenger cars and pedestrians (see section 4.6).

In the selected sample, the information for 745 injured cyclists exists, which maximum AIS was a result of a collision with the vehicle or a secondary impact. With regard to the contribution of the vehicle and secondary impact to the MAIS of the person, these 745 cases show the following distribution:

- 23\% - MAIS only determined by vehicle impact  
  \(\text{MAIS} = \text{MAIS}_{\text{car}}, \text{MAIS} > \text{MAIS}_{\text{sec}}\)
- 44\% - MAIS only determined by secondary impact  
  \(\text{MAIS} > \text{MAIS}_{\text{car}}, \text{MAIS} = \text{MAIS}_{\text{sec}}\)
- 33\% - MAIS determined by vehicle impact and secondary impact  
  \(\text{MAIS} = \text{MAIS}_{\text{car}} = \text{MAIS}_{\text{sec}}\)

Compared to pedestrians, where the MAIS determines from secondary impact in approximately every fifth case, the proportion of cyclists is twice as high. The proportion of cyclists who have not been injured by the passenger car (\(\text{MAIS} = \text{MAIS}_{\text{sec}}, \text{MAIS}_{\text{car}} = 0\)) is 42\%. This means that two out of five cyclists, who collide (with their bikes) with the vehicle front of a passenger car suffer their injuries only from the impact with the road / road surface. Therefore, the improvement of cyclists protection in traffic accidents with purely vehicle-related activities can only be achieved to a limited extent.
6 Summary and outlook

Injury risk functions gain importance in the field of vehicle safety. Whether for the understanding of injury mechanisms, the definition of protection criteria or the transfer of simulation results on real accident scenario. It often requires statements about the correlation between technical accident parameters, such as loads or accident severity, and the predicted injury severity. So far, there are no established, standardized and generally accepted definitions for the creation of statistical models.

This report examines the principles of injury risk functions, researches several statistical models and assesses their advantages and disadvantages as well as the applicability to real-accident data. Logistic regression proves as a well-suited and appropriate model for the creation of injury risk functions based on real accident data. Consequently, it applies the procedure for the creation of injury risk functions to full frontal pedestrian and cyclist - passenger car accidents and creates models for MAIS2+, MAIS3+ and fatally injured persons.

Regarding pedestrian - passenger cars accidents collision speed has the greatest impact on the injury severity. The age of the pedestrian is another important influencing factor. The separation of the dataset into three most widely homogenous age groups considers this influence. The exclusion of older vehicles (based on the year of vehicle model’s market introduction) and grazing impacts seem necessary in the sample. The report uses real world accidents out of the German In-Depth Accident Study (GIDAS) as database. Generally, GIDAS contains a sufficient amount of cases available for creating injury risk functions for pedestrians and cyclists. However, some restrictions, in particular to certain years of vehicle model’s market introduction, considerably reduces the sample.

Valid probability models are created for MAIS2+ and MAIS3+ injured pedestrians (adults and elderly). Figure 4.8 and Figure 4.9 show associated IRFs. Fatally injured pedestrian model quality ratings turn out to be quite low for both age groups. Therefore, the report omits the presentation of associated IRFs. Probability models for MAIS2+, MAIS3+ and fatally injured children cannot be seen as reliable. This is because of very small case numbers. Additionally, it must be indicated that a verification of the distribution assumption of the independent variables is recommended to ensure a correct representation of the characteristics of the dataset by the generated models. Confidence intervals provide the opportunity to assess the reliability of probability values.
The created functions can evaluate the effect on pedestrian accidents of such active safety systems whose activation leads to a collision speed reduction. A statistical comparison of the injury probability of the original accident with the modified one (simulated) assess the benefit of such a system when considering sufficient number of accidents.

Whenever using injury risk functions, applicability has to be taken into account. Listed functions do not apply to single case statements, any declaration of specific injury severities and special cases (here: lying, kneeling, jumping pedestrians; grazing, side, rear impacts; and rollover).

Analogously to the probability models for passenger car - pedestrian accidents, the study creates models for injury risk functions for passenger car - cyclist accidents. Instead of collision speed, they are using the more appropriate relative speed.

Despite the large case number, the quality of probability models for full frontal accidents between cyclists and passenger cars is not sufficient for reliable statements. None of the examined parameters proves to be a significant prediction variable for all full frontal passenger car-bicycle accidents. The reason for this is the heterogeneity of the group. Therefore, adoption of the models for passenger car-pedestrian accidents is not sufficient. Nevertheless, the approach is transferable methodically. A closer look regarding collision constellation and additional prediction variables (e.g. helmet usage) promise more reliable probability models.
7 References

References


[27] Lohninger, Hans: Grundlagen der Statistik. Pressbaum, Austria: Epina, 2010


[34] Niues, Tobias; Junge, Mirko; Rosen, Erik: Pedestrian injury risk and the effect of age. In: Accident Analysis and Prevention Bd. 86 (2015), S. 121–128


[36] O'Halloran: Logit/Probit


[41] ROSÉN, ERIK; SANDER, ULRICH: Pedestrian fatality risk as a function of car impact speed. In: Accident Analysis & Prevention Bd. 41 (2009), Nr. 3, S. 536–542


[46] VUFO: German In-Depth Accident Study. URL http://www.gidas.org
8 List of figures

Figure 2.1 – Examples of injury risk functions including reading example .................. 7

Figure 3.1 – Example for relative frequencies of MAIS2+ injuries over collision speed from GIDAS .................................................... 11

Figure 3.2 – Observed shares (relative frequencies) from GIDAS ...................... 12

Figure 3.3 – Determination of a regression function ..................................... 13

Figure 3.4 – Example of a binary logistic regression curve ............................. 15

Figure 3.5 – Comparison of the density functions of the logit model and the normal distribution of the probit model [36] ................................. 18

Figure 3.6 – injury risk functions for pedestrians with head impact (Contour lines of equal injury severity) [32] ................................................. 21

Figure 3.7 – Scatter plot of injury risk functions, sample size 10 [38] .............. 25

Figure 3.8 – Scatter plot of injury risk functions, sample size 160 [38] .......... 25

Figure 3.9 – Model quality consideration: example of classification results ........ 27

Figure 3.10 – Model quality consideration: example for ROC AUC ................ 29

Figure 4.1 – Structure of the GIDAS project ............................................... 33

Figure 4.2 – Sample of accidents involving pedestrians and passenger cars out of GIDAS ................................................................. 34

Figure 4.3 – Definition of the relative speed vrel in GIDAS ............................ 39

Figure 4.4 – Definition of the difference of speed ∆v in GIDAS ..................... 39

Figure 4.5 – Distribution of the year of vehicle model’s market introduction (JDME) involved in accidents with pedestrians in GIDAS ............... 48

Figure 4.6 – Sample sizes of defined age groups of the master-dataset for pedestrians ................................................................. 49

Figure 4.7 – Distribution of the collision speed in the master-dataset ............. 50
Figure 4.8 – Injury risk function for adult pedestrians in frontal passenger car-pedestrian collisions…………………………………………………………………………52

Figure 4.9 – Injury risk function for elderly pedestrians in full frontal collisions with passenger cars ………………………………………………………………………56

Figure 4.10 – Comparison of IRF (MAIS2+) with and without a secondary impact for adults ……………………………………………………………………………………..67

Figure 4.11 – Comparison of IRF (MAIS2+) with and without a secondary impact for elderly ……………………………………………………………………………………..67

Figure 5.1 – Sample for accidents between passenger cars and cyclists from GIDAS ……………………………………………………………………………………..69

Figure 5.2 – Distribution of the relative speed in the master dataset…………………..71

Figure 5.3 – Case numbers of the defined age groups in the master dataset for cyclists ……………………………………………………………………………………..71

Figure 5.4 – Distribution of the year of vehicle model’s market introduction (JDME) of passenger cars in frontal accidents with cyclists in GIDAS…………………..72

Figure 8.1 – Risk of a polytrauma for pedestrians with head impact………………..XI

Figure 8.2 – Extended injury risk function for a polytrauma ($ISS_x > 5$) for pedestrians with head impact (Contour lines of equal injury severity) …………………..XII

Figure 8.3 – Sketch and deformation at passenger car – cyclist accident: example 1 ………………………………………………………………………………………XIII

Figure 8.4 – Sketch and deformation at passenger car – cyclist accident: example 2 ………………………………………………………………………………………XIV
A Censored data

Censored (or truncated) data describes data, which is not completely known. For example, one knows the function value $y(x)$ above a certain argument $x$ without knowing the exact limit. Censored data is generated if not all necessary observations or measurements can be made (see [16], [27], [38]). For example, a traffic accident with an injury criterion (e.g. MAIS2+ injured driver) only happens at a certain prediction value (e.g. collision speed $v_{coll}$) and cannot be repeated with a changed prediction variable at the same boundary conditions. That means it is known that the injury criterion occurred at the real collision speed. However, it is not known whether it would also have occurred at a smaller collision speed.

The status of the data regarding their censorship should be known for further work with data for injury risk functions. It is distinguished into right censored, left censored, interval censored and uncensored or exact data. The four data types are explained in the following.

**Right censored data:**

A data point $(x; y)$ which is defined by the definition (2.2) $y(x) = 0$ (for example, a slightly injured person with MAIS = 1 in the case of the injury criterion MAIS2+) is called right censored, if the limit $x_{thres}$ is within the interval $[x; +\infty]$. The exact value of the limit is unknown.

An example from biomechanics:

A biomechanical model is assumed, in which the injury criterion $y$ is intended to indicate the occurrence of a fracture (0 – no fracture, 1 – fracture) as a result of an acting force (prediction variable $x$). A force $x_t$ is applied to the model and as a reaction, no fracture occurs. According to equation (2.2) follows:

$$y(x = x_t) = 0$$  \hspace{1cm} (8.1)

Under constant boundary conditions, any force $x \leq x_t$ will also not result in a fracture due to the lower load. The behavior of the model as a result of an acting force less than or equal to the tested force is known:
\[ y(x \leq x_l) = 0 \] (8.2)

On the other hand, the reaction to a force \( x > x_l \) is unknown, since the exact threshold \( x_{thres} \), up to which no fracture emerges, is not known from the measurement. According to equation (2.2) follows:

\[ y(x > x_l) = \begin{cases} 0 & \text{for } x < x_{thres} \\ 1 & \text{for } x \geq x_{thres} \end{cases} \] (8.3)

The behavior of the model above, in a graph "right", of the test value \( x_l \) is unknown or censored. It is said that the measured value is "right censored".

**Left censored data:**

A data point \((x; y)\) which is defined by definition (2.2) \( y(x) = 1 \) (for example, an injured person with MAIS \( \geq 2 \) in the case of the injury criterion MAIS2+) is designated as left censored if the limit value \( x_{thres} \) lies within the interval \([-\infty; x]\). The exact value of the limit is unknown.

An example from biomechanics:

The previously described biomechanical model is assumed to have an acting force \( x_u \). This time, a fracture occurs as a reaction. According to equation (2.2) follows:

\[ y(x = x_u) = 1 \] (8.4)

Under constant boundary conditions, each force \( x \geq x_u \) will also result in a fracture due to the greater load. The behavior of the model as a result of an acting force greater than or equal to the force tested is known:

\[ y(x \geq x_u) = 1 \] (8.5)
On the other hand, the reaction to a force $x < x_u$ is unknown, since the exact limit value $x_{thres}$, from which a fracture emerges, is not known from the measurement. According to equation (2.2) follows:

$$y(x < x_u) = \begin{cases} 
0 & \text{for } x < x_{thres} \\ 
1 & \text{for } x \geq x_{thres} 
\end{cases}$$

The behavior of the model below, in a graph "left", of the test value $x_u$ is unknown or censored. It is said that the measured value is "left censored".

**Interval censored data:**

A data point $(x; y)$ is called interval censored if the limit $x_{thres}$ is within the interval $[x_l; x_u]$. In this process, $x_l$ is the lower threshold at which the person or model is uninjured and $x_u$ is the upper threshold at which the person or model is injured. The exact value of the limit $x_{thres}$ is unknown.

An example from biomechanics:

The previously described biomechanical model is assumed one more time. The model is first loaded with a force $x_l$ and no fracture occurs as a reaction. According to equation (2.2) follows:

$$y(x = x_l) = 0$$

(8.7)

Subsequently, a force $x_u$ is applied to the model and a fracture occurs as a reaction. According to equation (2.2) follows:

$$y(x = x_u) = 1$$

(8.8)

Under constant boundary conditions, any force $x \leq x_l$ will not result in a fracture and any force $x \geq x_u$ will result in a fracture. The behavior of the model as a result of an acting force smaller than $x_l$ and greater than $x_u$ is known:
\[
y(x) = \begin{cases} 0 & \text{for } x \leq x_l \\ 1 & \text{for } x \geq x_u \end{cases}
\] (8.9)

On the other hand, the reaction to a force \( x_l < x < x_u \) is unknown, since the exact limit value \( x_{\text{thres}} \), from which a fracture emerges, is not known from the measurement. According to equation (2.2) follows:

\[
y(x_l < x < x_u) = \begin{cases} 0 & \text{for } x < x_{\text{thres}} \\ 1 & \text{for } x \geq x_{\text{thres}} \end{cases}
\] (8.10)

The behavior of the model within the interval of the test values \([x_l; x_u]\) is unknown or censored. It is said that the measured value is "interval censored".

**Uncensored or exact data:**

A data point \((x; y)\) is designated as uncensored or exact if the limit \( x_{\text{thres}} \), at which the injury criterion \( y \) occurs, exactly corresponds to the examined prediction variable \( x \).

\[
y(x) = \begin{cases} 0 & \text{for } x < x_{\text{thres}} \\ 1 & \text{for } x \geq x_{\text{thres}} \end{cases}
\] (8.11)

In the equations (8.1) – (8.11), the prediction variable \( x \) can, analogously to (2.3), also be shown in vector notation.

Real accident data is always left and right censored, since this data is almost non-reproducible and always represents multivariate single events.
B Extension of the “Contour Lines of Equal injury Severity”

Distinction of the impact type:

For missed collision no risk of injury:

\[ p_{ISSx} = 0 \]  (8.12)

For grazing impact linear increase from missed collision to full-frontal impact:

\[ p_{ISSx}(v_{coll}) = a(v_{coll}) \cdot x + b(v_{coll}) \]  (8.13)

For full-frontal impact: (exponential) function in dependence of the collision speed

\[ p_{ISSx}(v_{coll}) = \min_{v \geq 0} \left( \frac{v_{coll}}{v_{crit}} \right)^{ISSx}, 1 \]  (8.14)

Figure 8.1 – Risk of a polytrauma for pedestrians with head impact
“Both pedestrian injury risk functions, the one depending on the collision speed (fig. 5) and the one depending on the pedestrian’s engagement with the vehicle (fig. 4) can be summarized in one common higher-dimensional function. The resulting risk function is a function of the collision speed, i.e. the speed of the car, and the location where, in relation to the vehicle’s width, the pedestrian is hit.” [32]

Figure 8.2 shows the composite function for a polytrauma (ISSx > 5), based on the equations (8.12) to (8.14):

Figure 8.2 – Extended injury risk function for a polytrauma (ISSx > 5) for pedestrians with head impact (Contour lines of equal injury severity)
C Examples for remarkable car – cyclist accidents

Example 1:

Figure 8.3 shows the sketch and the impact constellation of a passenger car – cyclist accident from the sample used in chapter 5.

![Sketch and deformation at passenger car – cyclist accident: example 1](image)

The following speed values were determined or reconstructed:

Cyclist with the speed: \( v_{cyc} = 15 \text{ km/h} \)

Passenger car with the speed: \( v_{car} = 87 \text{ km/h} \)

Reconstructed relative speed: \( v_{car} = 73 \text{ km/h} \)

The cyclist was female and 82 years old. Despite of the high relative speed and the woman's quite advanced age, she only suffered from a \( MAIS = 1 \) (single injuries: first degree TBI, fracture of the thumb, head laceration, contused laceration). The images of the deformations at the bicycle in Figure 8.3 show that the cyclist was hit by the car front at the rear wheel, but was not fully captured by the vehicle.
Example 2:

Figure 8.4 shows the sketch and the impact constellation of another passenger car – cyclist accident from the sample used in chapter 5.

![Image of accident scene]

Figure 8.4 – Sketch and deformation at passenger car – cyclist accident: example 2

The following speed values were determined or reconstructed:

- Cyclist with the speed: $v_{cyc} = 15 \text{ km/h}$
- Passenger car with the speed: $v_{car} = 0 \text{ km/h}$
- Reconstructed relative speed: $v_{car} = 15 \text{ km/h}$

The cyclist was male and 43 years old. At the moment of the collision, the passenger car driver had already stopped his vehicle. Thus, the speed of the passenger car was 0 km/h. Despite the relatively low relative speed and the middle age of the cyclist, he suffered from a relatively high injury severity $MAIS = 3$ (femur fracture).
D Influence of the secondary impact on the injury severity

<table>
<thead>
<tr>
<th></th>
<th>MAIS caused only by vehicle impact</th>
<th>MAIS caused only by secondary impact</th>
<th>MAIS caused by vehicle and secondary impact</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pedestrians</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAIS1+</td>
<td>46%</td>
<td>21%</td>
<td>33%</td>
</tr>
<tr>
<td>MAIS2+</td>
<td>66%</td>
<td>16%</td>
<td>19%</td>
</tr>
<tr>
<td>MAIS3+</td>
<td>62%</td>
<td>19%</td>
<td>19%</td>
</tr>
<tr>
<td><strong>Cyclists</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAIS1+</td>
<td>23%</td>
<td>44%</td>
<td>33%</td>
</tr>
<tr>
<td>MAIS2+</td>
<td>39%</td>
<td>48%</td>
<td>13%</td>
</tr>
<tr>
<td>MAIS3+</td>
<td>59%</td>
<td>31%</td>
<td>10%</td>
</tr>
</tbody>
</table>