CESifo WORKING PAPERS

7144
2018

July 2018

Adolescents on the Road: A Case Study of Determinants of Risky Behaviors

Filippo Elba, Fiammetta Cosci, Anna Pettini, Federico M. Stefanini

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editors: Clemens Fuest, Oliver Falck, Jasmin Gröschl
www.cesifo-group.org/wp

An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: www.CESifo-group.org/wp

CESifo *Working Paper No. 7144*
Category 3: Social Protection

Adolescents on the Road: A Case Study of Determinants of Risky Behaviors

Abstract

The 2016 report of the European Transport Safety Council claims that EU safety progress has come to a standstill. This study aims at deepening the knowledge of factors that influence adolescents' risky behavior on the road. Bayesian Networks offer a promising new way to looking at the issue. In the analysis of a dataset collected in Tuscany, Italy, called EDIT, we found evidence that the use of alcohol and illegal substances explain only part of the probability of having an accident, and that other observable variables, like the level of distress or the type of school attended are significantly related to the probability of incurring in a road crash. New and close attention should be given to a systemic approach and to a plethora of environmental and individual variables that may rise the probability of road accidents for very young drivers.

JEL-Codes: C110, D910, I110.

Keywords: Bayesian Networks, structural learning, road accidents, distress factors, risky behavior, adolescence, youth, novice drivers.

Filippo Elba
Department of Economics and Management (DISEI)
University of Florence / Italy
filippo.elba@unifi.it

Fiammetta Cosci
Department of Health Sciences (DSS)
University of Florence / Italy
fiammetta.cosci@unifi.it

*Anna Pettini**
Department of Economics and Management (DISEI)
University of Florence / Italy
anna.pettini@unifi.it

Federico M. Stefanini
Department of Statistics, Computer Science, Application (DiSIA)
University of Florence / Italy
federico.stefanini@unifi.it

*corresponding author

1. Introduction

Road safety is one of the UN's Sustainable Development Goals and a global issue. The great part of all road traffic deaths and injuries occur in low-income and middle-income countries, though absolute numbers in the higher-income countries of the OECD are still high. Since 1970s, along with the increasing number of circulating vehicles and good transports, and the removal of barriers to empower the European market yielded the need of harmonizing the heterogeneous regulatory approaches of the EU member states (Threlfall, 2003). Road safety became the main issue and great progress has been gained over the last two decades as a result of the combined effort of international institutions and EU policies (Castillo et al, 2014B). In 2001, the European Union set the target to halve road death by 2010 and renewed its commitment to improve road safety by setting a target of reducing road deaths by 50% by 2020 compared to 2010.

Indeed, safety for car occupants has increased greatly in many countries through improvements of road systems, prevention campaigns, and vehicle design. The adoption of new laws and regulations and the enforcement of these laws to improve compliance with speed limits, seat belt use, and drink-driving rules have had a major impact on reducing the burden of road transport accidents (OECD, 2015).

Still, road traffic crashes represent a serious health problem, both at social and individual levels. A recent work by Wijnen and Stipdonk (2016) provided an international overview of the most recent estimates of the social costs of road crashes. They showed that the economic and social costs of road crashes fatalities, serious injuries, and minor injuries in High Income Countries ranged from 0.5% to 6.0% of the GDP with an average of 2.7%. They also pointed out that the majority of costs is related to injuries, with an average share of 50%, while human costs, which seem to cover the great part of the total costs, are still underestimated.

According to the international guidelines, general costs include direct, indirect, and immaterial costs. Direct costs are those of health care services, rehabilitation, as well as administrative costs and property damages. Indirect costs value the lost household productivity and lost earnings and consumption of victims, survivors, caregivers, and families. Road traffic injuries, indeed, may imply severe financial stress on a family, who might have to absorb a part of the direct medical costs in addition to the indirect costs: the victim's inability to continue normal activities may result in lost income for parents and caregivers, for example through the reallocation of the work (Castillo et al, 2014A). Immaterial (human) costs involve pain, psychological damage, and loss of quality life in those implicated in the accident as well as in their relatives. In addition, people who involved in road traffic collisions may have long-term injuries or disabilities. The extent of disability and loss in quality of life range from minor or short-term reduced functioning, which may affect daily functioning, to severe or permanent disabilities (Hours et al, 2013). Direct costs may not present substantial differences across age groups while both indirect and human costs are higher for young people, due to a statistical value of a life which is, on average, higher (Johansson-Stenman, Martinsson, 2008).

The specific focus on youth and safety on the road come along with the last wave of the European Safety Policy, since road traffic crashes are the leading cause of death among 15-24-year-olds. The enforcement of highly cost-effective policies like seat belt use, alcohol restrictions, motorcycle helmet use, speed enforcement, graduated licensing schemes, and safer infrastructure have produced great results in terms of road fatalities decrease.[5] Among the determinants of risky behavior, impaired driving has been considered a central issue, and the zero-tolerance policy for novice drivers has led to positive results (Podda, 2012).

The decreasing trends in general accidentality are uniformly distributed across age groups, and the relative relevance of the phenomenon for young drivers remains steady over the time, thus there is still room for increasing attention. About 4,000 young people (aged 18-24) are killed each year on EU roads, the is the first single cause of death in this age group. Young drivers are still overrepresented in road traffic death statistics by a factor between 1.2 and 3.9[6] when compared with the proportion of this age group in the general population, signaling that they need a special attention within the broader field on road safety policies. (WHO 2007)

Young people face a unique mixture of factors that leads to a higher rate of collisions and deaths. A lack of experience on the road make them vulnerable and worse at anticipating and reacting to hazards. Inexperienced drivers need large safety margins to compensate their lack of experience but they usually choose small safety margins (Hakkert, Gitelman, 2014). In addition, a combination of physical and developmental immaturity, as well as youth-related lifestyles, increase the risk of young road users to road traffic collisions (European Transport Safety Council, 2017).

The long list of factors explaining the overrepresentation of youth in the statistics is a perfect stimulus to find out which other determinants may be of interest and prevention policies remain the most promising ground being the great part of road accidents preventable as they relate to behavior. The opportunity to deepen the research on the determinants of young drivers' risky behavior came from data collected to this aim in Tuscany (Italy). The EDIT[7] Dataset was created by the Epidemiology Observatory of the Tuscan Regional Health Board from the Survey "Epidemiology of determinants of road accidents in Tuscany - 2015". The Survey was conducted within the framework of a project that aims at becoming a reference for the production of data analyses and knowledge on road accidents among adolescents. The survey includes several areas of adolescents' life and we believe that analyzing them as a whole can be of interest to explore which intervention model might be more adequate.

The paper is organized as follows: section 2 focuses on the background of our analysis by reviewing the main issues on risk factors for young drivers; in section 3 the role of psychological distress in young drivers is analyzed. Section 3 also describes the dataset, the specific tool used to assess psychological distress, and the empirical model. Section 4 draws

[5] For an extensive review of the road safety literature, mainly in high-income countries, see the appendix to Sheehan 2014 (p.57).

[6] As a share of all driver fatalities within the EU, the proportion of fatalities for young drivers ranges from 18 % in Denmark to 32% in Germany. In contrast, the share of this age group in the total population ranges from 8% in Denmark to 13% in Ireland (SafetyNet (2009) Novice Drivers, p.5).

[7] Epidemiologia dei Determinanti dell'Infortunistica stradale in Toscana.

the basics on the statistical analysis run to study how risky behavior of young drivers is related to the different aspects of their life. Section 5 shows and comments the results. Section 6 is based on conclusions.

2. Risk factors for accidentally in adolescents

The risk for accidentally is high for adolescent drivers. The logical first places to look in explaining this phenomenon are age and inexperience, but it is difficult to distinguish the relative effect of each because they are very highly correlated (Institute of Medicine (US) and National Research Council (US) Committee on the Science of Adolescence, 2011). Both come into play in making drivers more likely to take risks and less able to detect and respond to hazards. Studies in countries where it is common to license drivers at age 18 suggest that inexperience is a greater risk factor than chronological age, but it is likely that they interact (Blum & Blum, 2009). Observational studies of crashes and violations have shown that adolescent drivers are more likely to speed, tailgate, and leave too small gaps between their vehicle and the one in front, for example. They also lack the experience that helps older drivers perceive that their speed is too great for conditions or take note of a situation in the middle distance that may require responsive action. Two conditions that exacerbate the already heightened risk for young and inexperienced drivers are driving at night, where there is limited light and the tiredness is higher, and driving with peers (Institute of Medicine (US) and National Research Council (US) Committee on the Science of Adolescence, 2011). For teenage drivers, particularly males, peer passengers are a distraction and perhaps a motivation to drive too fast or take other risks.

Adolescents are also at risk of driving under the influence of alcohol or substances (Institute of Medicine (US) and National Research Council (US) Committee on the Science of Adolescence, 2011). Overall, males use more substances than females; adolescents from families with low socioeconomic status (SES) are more likely to smoke cigarettes than pairs; adolescents who live in poverty or in affluence seem to have higher rates of substance use than those within intermediate income groups.

The mental health status of adolescents relates in various ways to this issue since psychological symptoms or disorders may increase the attraction to risk of young people and increase the rate of risky behaviors (Institute of Medicine (US) and National Research Council (US) Committee on the Science of Adolescence, 2011). In a prospective epidemiological study from the United States on depressed boys and girls, the depression rates increased across the puberty span , particularly among girls (Gleid and Pine, 2002). The same was observed for anxiety disorders. By contrast, conduct problems were found more prevalent among boys; conduct problems were found to be associated with smoking and substance use, vehicle crashes and other impulsive behaviors, and risky sexual behavior (Institute of Medicine (US) and National Research Council (US) Committee on the Science of Adolescence, 2011).

3. Psychological distress in adolescents

Psychological distress has been largely defined as a state of emotional suffering characterized by symptoms of depression (e.g., lost of interest, sadness, hopelessness) and anxiety (e.g., restlessness, feeling tense) (Mirowsky & Ross, 2002). These symptoms may be tied in with somatic symptoms (e.g., insomnia, headaches, lack of energy) that are likely to vary across cultures (Kleinman, 1991). Psychological distress is usually described as a non-specific mental health problem (Dohrenwend & Dohrenwend, 1982) although it is clearly characterized by depressive and anxious symptoms, nevertheless depressive and anxiety disorders are phenomena distinct from psychological distress although interrelated to it (Payton, 2009).

The status of psychological distress in the psychiatric nosology is ambiguous and has been debated at length. On one hand, it is viewed as an emotional disturbance that may affect the social functioning and day-to-day living of individuals (Wheaton, 2007). On the other hand, it is a diagnostic criterion for some psychiatric disorders and, together with impairment in daily functioning, a marker of severity of symptoms of other psychiatric disorders. Otherwise, in line with the stress-distress model, it is viewed as a transient phenomenon consistent with a normal emotional reaction to a stressor. Horwitz (2007) illustrated this point by quoting a series of studies conducted among adolescents and showing the high fluctuation of depressive symptoms over interval as short as one month. Horwitz argued that this fluctuation reflects the relatively brief sorrows that follows from failing a test, loosing a sporting match, breaking up with a boyfriend or a girlfriend.

The prevalence of psychological distress is difficult to pinpoint due to the variety of the scale assessing it, of the time windows used in data collection, and the cut-off point applied to dichotomize the score of distress and identify individuals with pathological distress. It roughly ranges between 5% and 27% in the general population (Drapeau et al., 2012). Two characteristics of the prevalence of psychological distress are noteworthy: the widespread sex difference and the variation over life. The prevalence of psychological distress is higher among females than in males in all age groups (Caron & Liu, 2011). This sex difference has been explained according to three different hypotheses: the first hypothesis is that psychological distress may be partly attributable to sex-related personality traits or biological components; the second hypothesis is that, in most societies, females are more exposed or more vulnerable to the socio-cultural risk factors associated with psychological distress (e.g., parental stress, marital stress) (Clearly & Mechanic, 1983); the third hypothesis suggests that, in most cultures, the expression of emotions differs across sex (Drapeau et al., 2010). The prevalence of psychological distress also varies with age, tending to decrease over the lifespan from late adolescence (Caron & Liu, 2011). The decreasing is usually attributed to differential exposure to risk factors and to survival bias.

Since the literature does not report on this topic, the aim of the present study was to verify whether accidentality in adolescents can be related to psychological distress.

4. Methodology

A Bayesian network (BN, Friedman and Koller, 2009, for a comprehensive account) is a type of probabilistic graphical model that can be used to build models from data and/or expert opinion. It can be used for a wide range of tasks, including prediction, anomaly detection, diagnostics, automated insight, reasoning, time series prediction and decision making under uncertainty. More technically, a BN is a statistical model in which the joint distribution of discrete or discretized random variables included into the study is represented through the product of conditional distributions defined after inferring conditional independence relationships (CIRs). The qualitative component of a BN consists of a Directed Acyclic Graph (DAG) in which nodes represent random variables and where a missing arrow implies a CIR. For example, $X -> Y -> R$ is a DAG made by three random variables and R and X are conditionally independent given Y: here the missing arrow is $X -> R$. Separation theorems make possible to obtain implied CIRs given observed variables.

The quantitative component of a BN is made by conditional probability values. At each node, say Y, a conditional probability table (CPT) is defined where elements are probability values of event $Y = y_i$ given the configuration taken by its conditioning random variables, say $X = x_i$ that is those variables associated to nodes in the DAG that originate arrows reaching node X (parents of Y in the DAG).

In the first step of the analysis, DAG's structure (nodes and arrows) is specified given expert's prior beliefs and/or by using algorithms of structural learning that exploit collected data. Then, model parameters (elements of all CPTs) are estimated. After a BN is fully specified, marginalization and conditioning can be performed by fast and effective algorithms even in very large networks of hundred variables, thus diagnostic and predictive reasoning are performed following Bayes' rule. Monte Carlo simulations expand the inferential task making possible to account for uncertainty about model parameters and, eventually, structure.

Structural learning is performed by maximization of a Bayesian objective function (details in Friedman and Koller, 2009, section 18.3.2). A "blacklist" can be compiled to forbid algorithm optimization steps that produced candidate arrows not compatible with background information. The score function $BDe(G) = log(P(D \mid G)) + log(P(G))$ is defined by the integrated likelihood function $P(D \mid G)$ of all data D given a structure G and an initial distribution $P(G)$ representing the expert degree of belief about structure G (i.e. a DAG).

BDe stands for Bayesian Dirichlet (Likelihood) Equivalent metric, with the following main features : (1) parameters of CPTs are marginally independent and each initial distribution belongs to the Dirichlet family of distributions; (2) two different structures that represent the same set of CIRs have equal BDe values; (3) the elicitation of initial distributions is simplified into the selection of the size of a hypothetical sample representing the strength of our prior beliefs, here set to 10 virtual observations.

The point estimate of model parameters for the top-scoring structure (DAG) is calculated as expected value of model parameters with respect to the final distribution.

Given an estimated BN, it is possible to read CIRs using separation theorems and building the Markov blanket for a node A (Friedman and Koller, 2009): this is a set of nodes $MB(A)$ composed of A's parents, its children, and its children's other parents: it can be shown that A is conditionally independent of all other nodes outside $MB(A)$ given its MB.

5. The analysis

5.1 Data (EDIT)

The Database "EDIT 2015" has been created from the Survey "Epidemiology of determinants of road accidents in Tuscany - 2015"[8]. The Survey was conducted by the Epidemiology Observatory of the Tuscan Regional Health Board within the framework of a project that aimed to become a reference point for the production of analysis and knowledge about road accidents among adolescents.

The Regional Board of Education compiled the school classes list from which the statistical sample was extracted. The sample was stratified according to school type and Local Health Board (LHB, in italian: *Azienda sanitaria locale* - Asl). Four hundred individuals were selected for each Asl (four schools per territory), with the exception of Florence's territory, where eleven schools were selected because of demographic reasons. For each LHB, schools were sampled with routine sampling, with a sampling probability proportional to the number of students of each institute. Schools were previously sorted by typology. For each school selected, five classes were extracted (from the first to the last year) from different sections.

The 2015 Survey was conducted from February to May 2017. It involved 5,077 students from fifty-seven secondary schools. Respondents were 14-19 years old: the most represented age group was 16, with 1,007 kids; the smaller group was composed of 553 subjects of 19 years of age. With regards to sex, 54.2% were males, 45.2% females.

The 2015 Survey dealt with a wide range of issues: driving behavior; relationships with equals and relatives; educational performance; sport activities; eating habits; consumption of alcoholic drinks and tobacco; use of substances; sexual behavior; bullying; propensity to gambling; sleep quality; emotional state. Concerning this last point, the EDIT Survey, in line with the literature about adolescents' psychological problems, adopted the "Kessler Psychological Distress Scale" (K6). This tool measures mental condition starting from a list of self-reported symptoms allowing the identification of the level of distress.

K6 is a quantifier of non-specific psychological distress (Kessler et al., 2002; Kessler et al., 2003), it is the most widely-used screening scale for mental illness in community epidemiological surveys. The K6 questions originate from the Item Response Theory and were initially developed from pilot survey results.[9] It has demonstrated excellent internal consistency and reliability (Cronbach's $\alpha = 0.89$) (Kessler et al., 2002; Kessler et al., 2003). It has shown consistent psychometric properties across major socio-demographic sub samples as determined by the areas under the Receiver Operating Characteristic (ROC) curve (Kessler et al., 2002). Each of the six items on the questionnaire ask about feelings that might have occurred during the previous 30 days and are rated by the respondent on a 5-point scale. The K6 is scored using the unweighted sum of answer responses, where responses of "none

[8] This Survey was first run in 2005, and then repeated in 2008, 2011, 2018.

[9] see: http://qcmhr.uq.edu.au/worc/measures.htm

of the time" were zero to "all of the time being" yielding a score of four. Thus, the range of responses is 0-24.

5.2 The Empirical model

The original "EDIT 2015" database contains 200 variables for 5,077 observations. A first selection isolated the 60 most important variables for the purpose of our research. Some of them were transformed to reduce the number of distinct values by aggregation (for examples, hours of sleep, kind of relationship with equals or relatives, assessment of schoolwork, K6 score), while continuous variables were transformed into interval variables, after partitioning their sample space trough suitable intervals (for example, body mass index). The final number of working variables is 36 (Table 1).

For variables connected by strong "conceptual" associations, new variables were created crossing and in some cases reducing classes' number, e.g. of parents' age and professional status. In the original dataset, information about each parent was provided, nevertheless after data cleaning and formatting a few variables were recoded into a new one. For example, parents' age was transformed into a new unique variable which sums up the concordance between parents' ages according to three levels: both under 50 years of age, both over 50 years of age, one under and one over 50 years of age. Parents' professional status was changed into a new unique variable divided into seven levels: both employed; both retired and/or unemployable; both unemployed and/or inactive; one employed and the other one retired or unemployable; one employed and the other one unemployed; one unemployed and the other one unemployable or retired; "I don't know".

In other cases, new variables were created after "vertical" and/or "horizontal" aggregations (with respect to an individual by variable table). For example, this solution was adopted for adolescents' consumption of substances and alcohol, sexual behaviors, gambling propensity, eating habits. Variables that describe consumption of different substances with similar effects or belonging to the same class were horizontally aggregated in: consumption of stimulants (smart drugs and amphetamines); of hallucinogens (LSD, ecstasy and magic mushrooms); of cannabinoids (cannabinoids and synthetic cannabinoids). On the other side, vertical aggregation referred to answers to drop-down questions, e.g., variables that describe cannabinoids consumption.

After manipulations and the removal of observations with NA values, a dataset of 3,647 units for 36 variables was obtained and used to learn the structure of our BN model.

The software used to infer network's structure was R. The R-packages used are *bnlearn* (Scutari, 2010), *gRain* (Højsgaard, 2012), *Rgraphviz* (Hansen et al., 2017), *ggm* (Marchetti et al., 2015). The list of "forced" root-nodes was made by: respondents' sex; respondents' age, parents' age concordance. With regard to parents' qualification concordance, we let the algorithm chose whether it was an external factor or a child node of any other variable of the model, except parents' age concordance.

A BN based on an Unsupervised Machine Learning algorithm[10] is here presented. Thus, no pre-structure among variables was used (apart from the above exceptions). In other words we did not make a-priori assumptions about which variables are dependent or independent.

[10] The algorithm chosen for our purpose is the Hill Climbing. For details: Russel and Norving (2003).

The BN model used aims at investigating which is the most appropriate possible causal model supported by the data and investigates the relationships among all variables in the Survey.

5.3 Results

The use of a Bayesian Network approach on the EDIT dataset has been built without *a priori* hypotheses on the determinants of young drivers risky behavior. Figure 1 shows the top-scoring DAG (structure) for variables listed in Table 1. It gives some interesting hints to the analysis. The first is the importance that should be given to a systemic approach when adolescents are the target population: their attitude toward risk when driving is part of a complex system of variables that are related to and that influence each other.

Evaluating variables that the DAG places right before the node AT (accidents which have required the ER): they all show a propensity to risky behaviors, such as consuming cannabis daily (node CA), smoking (X), using psychotropic substances (BX), illegal substances (AS), drinking too much (BK), taking fines (AK).

The conditional probability tables (CPTs) show the numbers.[11] Taking for example AK and AT, among all students who have never driven under the effect of substances and had no fines, the 4% has had an accident while among students who never experienced impaired driving but had fines, the 14.6% had an accident. A remarkable difference, which persists looking at students who had an accident and admitted to have driven under the effect of substances: 13% of them had no fines before while 27.8% had. The differences show that fines have some predictive capacity on accidentality.

The network, together with the CPT, confirms the important role of impaired driving: it rises the probability of incurring in an accident by 11.3%.[12] The risk rises to 28% for those who have taken fines and have driven under the influence of illegal substances. Current preventive European policies, like those ensuring high levels of enforcement of drink–driving legislation, are mainly based on this well known fact.[13]

Alcohol and illegal substances, however, do not explain the whole phenomenon, as many other factors concur to increase the risk on the road. The DAG produced by working on the EDIT database, as well as the associated CPTs, clearly show that the probability for an adolescent to use the road improperly is affected by a plethora of environmental and individual conditions, spanning from family background and the type of school attended to the level of distress and the relationships with peers. The network yields reasonable connections among variables that belong to the same sphere of life. Thus, within the network, different subsets of variables outline the socio-family environment (BE,BD,BF,BG,BH,M), the relationship with the body (BC,BM,BO,AP), with the school (BA,Q,Tt,P), the attitude toward addiction (CA,BZ,BY,X,BK) as well as the subset of variables belonging to the psychological

[11] CPTs of our analysis are available at the link: https://docs.google.com/document/d/1BA5iGL5R_orqO76hSDLkmhqYy4qwtqb-PRC__swFf94/edit?usp=sharing

[12] Note that this number does not include minor as well as fatal accidents.

[13] Exceeding speed limits, drink or distracted driving and failure to wear a seat belt are the leading causes of death and serious injuries on European roads (PIN Annual Report 2016).

status of the adolescents (BB,O,U,AL,AM,R,S). Not surprisingly, the network shows that the probability of having an accident changes according to the sex and, to some extent, the age.[14] These two variables are obviously exogenous, and this is the reason why we forced them in the role of 'parents' in the network. Males are, on average, at higher risk of making accidents than females because they drive more regularly and use more frequently substances (24.9% vs 17.4%). If we take a male and a female who regularly use cannabis and attend a professional school with low school performance, their probability to have an accident differ by 2.5% (12.5 vs 10%).

Other *profiles* show more striking differences in terms of probability of road crash. The average probability for a young person, taken from the EDIT dataset, of having an accident is 5.6%. The probability drops at 4.85% for a male attending a high school, who does not use cannabis and who has good scores at school: a significant difference of 12.5% for a male with opposite characteristics (i.e., a regular cannabis user, attending a technical or professional school with poor achievements). A male attending a professional school, using cannabis daily, with low scores at school and who has never read a book has a probability of 12.5% while a female who attends a high school, who does not use cannabis, has high scores at school and read books has probability of 4.8%. The difference among males and females in terms of probability to have an accident disappear (6.8% for both sexes) if we consider students with a high level of distress, low performance at a professional school, and no habit to read books. Students affected by a high level of distress, but able to have good scores at a high school, on the other hand, have a lower probability (5.7% for males and 5.2 for females). Thus, *ceteris paribus*, poor grades at school act also as predictor of the risk on the road. School performance, in turn, is highly related to the level of distress: among students with low score at school, 29% of males and 48% of females had a high level of distress. When, instead, the school achievements are good, only the 3% of male and the 15% of females present high levels of psychological distress.

If we outline other profiles by comparing males and females with opposite habits in terms of cigarettes and cannabis use, and drinking a lot (also occasionally), the probability goes from 12.7% (high risk profile) to 4.7% (low risk profile).

To sum up, the attitude to risky behaviors is clearly related to the probability of having an accident, but even the highest risk profile explain only a part of the cases of crashes (AT). Chance plays obviously an additional role, but nobody would argue that it explains the greatest part of accidents (Hakkert et al., 2014). A network like the one presented here helps to make guesses about which group of variables should be taken into account. A d-separation test[15] indicates which ones are in need to be studied.

[14] The impact of age is here related to the particular age group included in the EDIT database (age 14-19): the number of young with a drive license is smaller in the first adolescence.

[15] According to the Bayes assumption: "each variable is conditionally independent of its non-descendants, given its parents." It is possible to reason about independence using this statement, but we use the d-separation test as a more formal procedure for determining independence. Therefore, if the set of conditioning variables is empty, we can investigate if two variables are marginally independent; if the set of conditioning variables is full, we can investigate if two variables are conditionally independent given the ones in the set.

The level of psychological distress (BB), the quality of relationship with peers (O), the number of hours slept per night (U), are not d-separated from AT, namely the psychological subset variables of a young are connected to his or her attitude toward risk. In addition, the level of psychological distress, measured via the K6 index, contributes to the probability to have an accident independently from the group of the risk-on-the-road variables (CA, X, BK). In other words, when the observed attitude toward risk is very low, the risk while driving may be significant if psychological area variables occur.

The model of the DAG in fig. 1 can be compared to other models obtained stratifying the observations by a given variable[16]. The second and the third DAGs were obtained stratifying for school type, in particular for high school versus professional schools. In both cases BE (the education degree of parents) was no more linked to the other variables, and for the high school students also age concordance and working condition of parents did not predictive the rest of the variables. The most striking difference is the role played by school performance: for high school students it directly predicted the psychological distress and the length in time of use of pc, the relationship with peers and the use of illegal substances, which is in turn directly predicted road accidentally. In addition, for these students the level of distress was directly related to eating habits. Regarding professional school students, connections among variables, and in particular the Markov Blankets for AT, are more alike the ones commented on the DAG in fig.1. The stratification for the type of school, thus, suggests that different models should be thought and assessed for students involved in different carrier paths.

[16] Stratification based on one variable classes is possible if collected data are statistically representative on that score.

Table 1: Labels and meaning of variables in the network

A – Gender;	AU – Social or alone drug user;
B – Age;	AV – Age of first drug use (in classes);
M – Broken family;	BA – Type of school;
N – Quality of family relationships;	BB – Distress index (K6 in classes);
O – Quality of relationships with own age people;	BC – Body Mass Index (in classes);
P – School achievements;	BD – Concordance of ages of parents;
Q – Repeating of school years;	BE – Concordance of qualification of parents;
R – Use of PC;	BF – Concordance of activity status of parents;
S – PC hours per day;	BG – Parental alcohol consumption;
Tt – Number of books red (in classes);	BH – Parental cigarettes consumption;
U – Hours of sleep (in classes);	BK – Drinking too much in the last month;
X – Number of smoked cigarettes per day (in classes);	BL – Gambling propensity;
AK – Traffic fines (in classes);	BM – Sport propensity and motivations;
AL – Victim of bullying;	BO – Diet and/or pills consumption to lose weight in the last month;
AM – Bullying acted;	BX – Use of prescription drugs;
AP – Threw up to lose weight in the last month;	BY – Use of energy drinks;
AS – Driving while impaired (number of times, in classes);	BZ – Use of cocaine;

AT – Number of road accidents with use of Emergency Room (in classes);	CA – Use of cannabinoids.

Figure 1: Top-scoring structure (DAG) obtained by maximizing the BDe score through hill climbing. Names and meaning of variables are listed in Table 1.

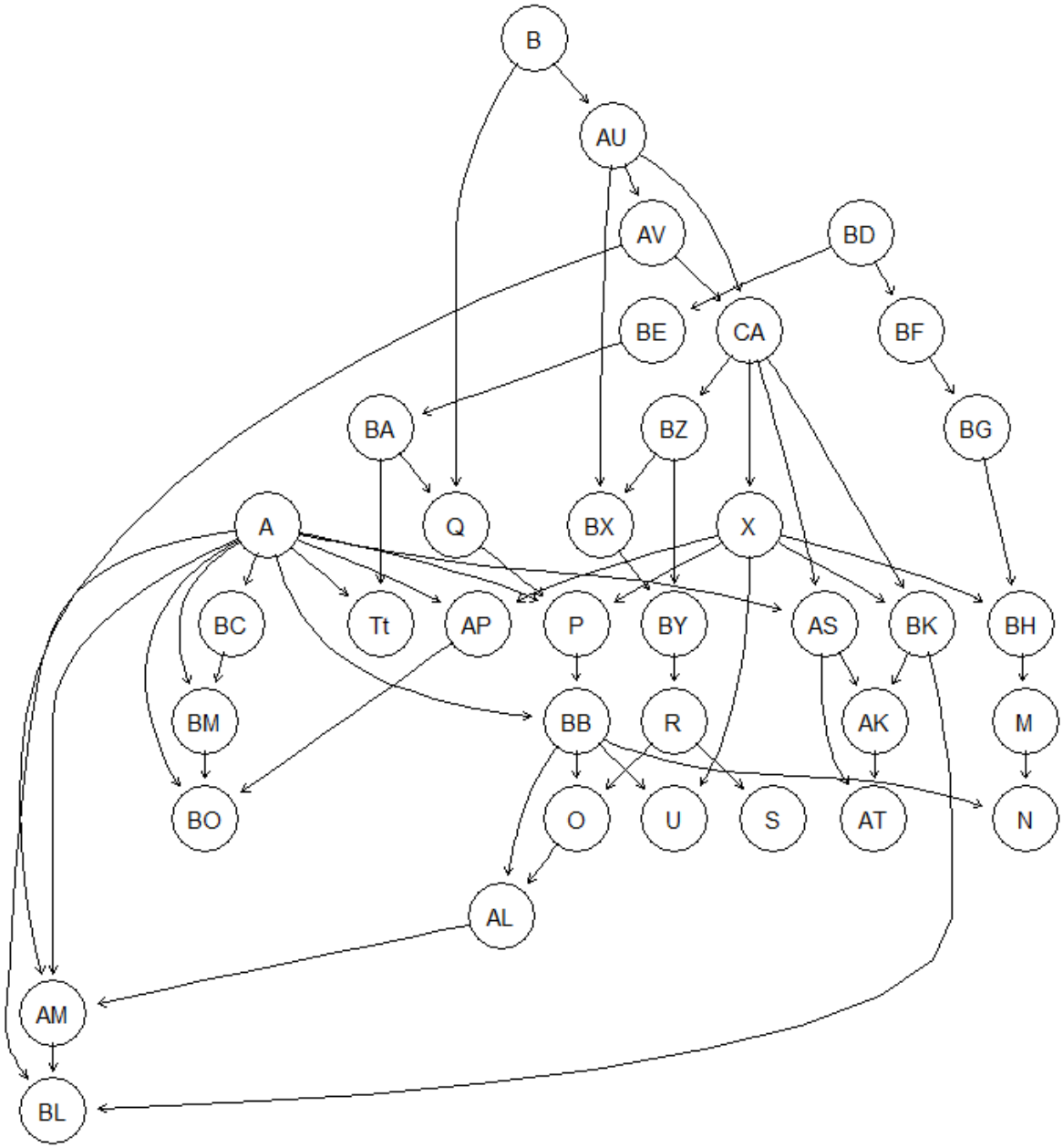


Figure 2: Bayesian Network for high schools students

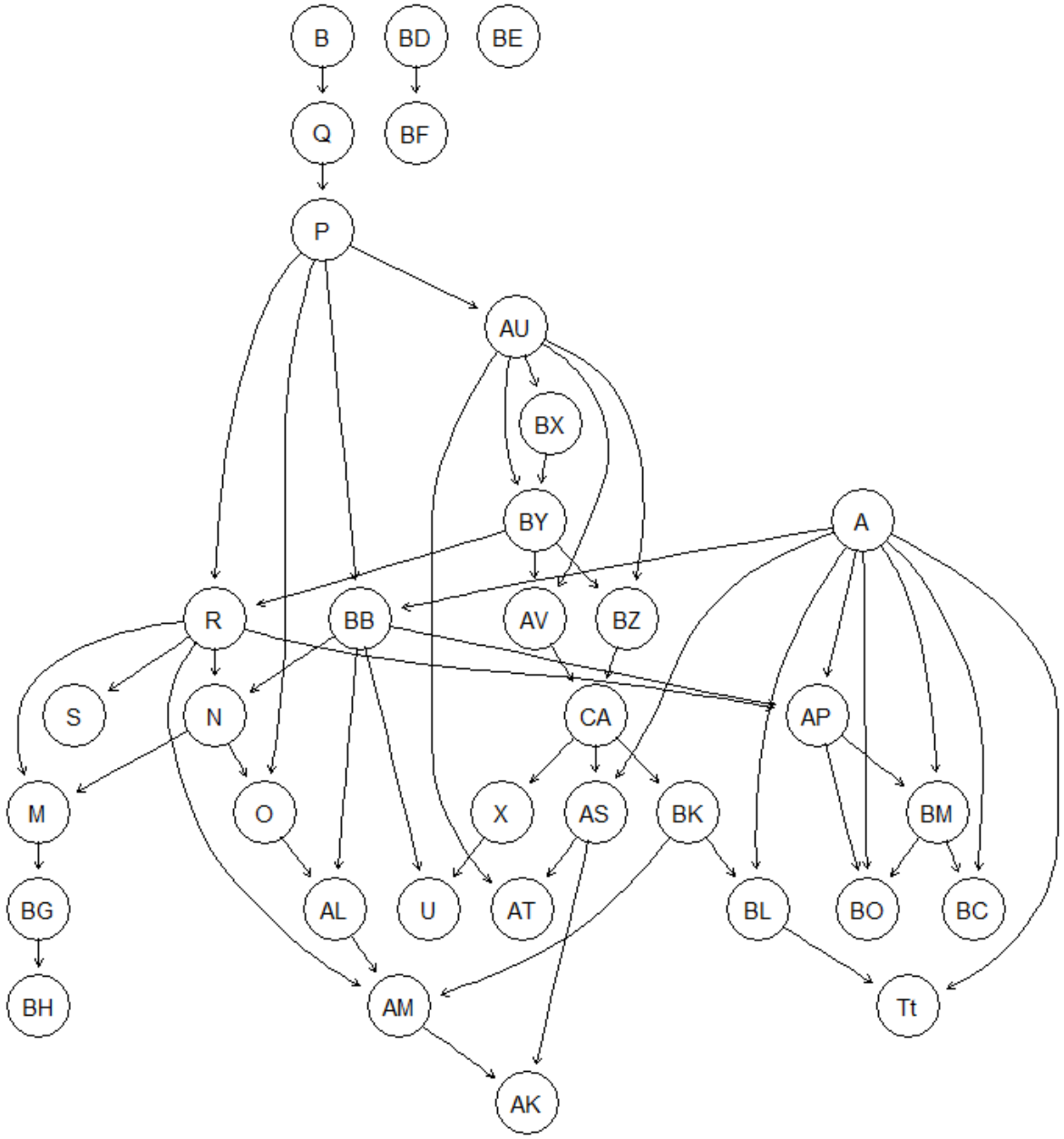
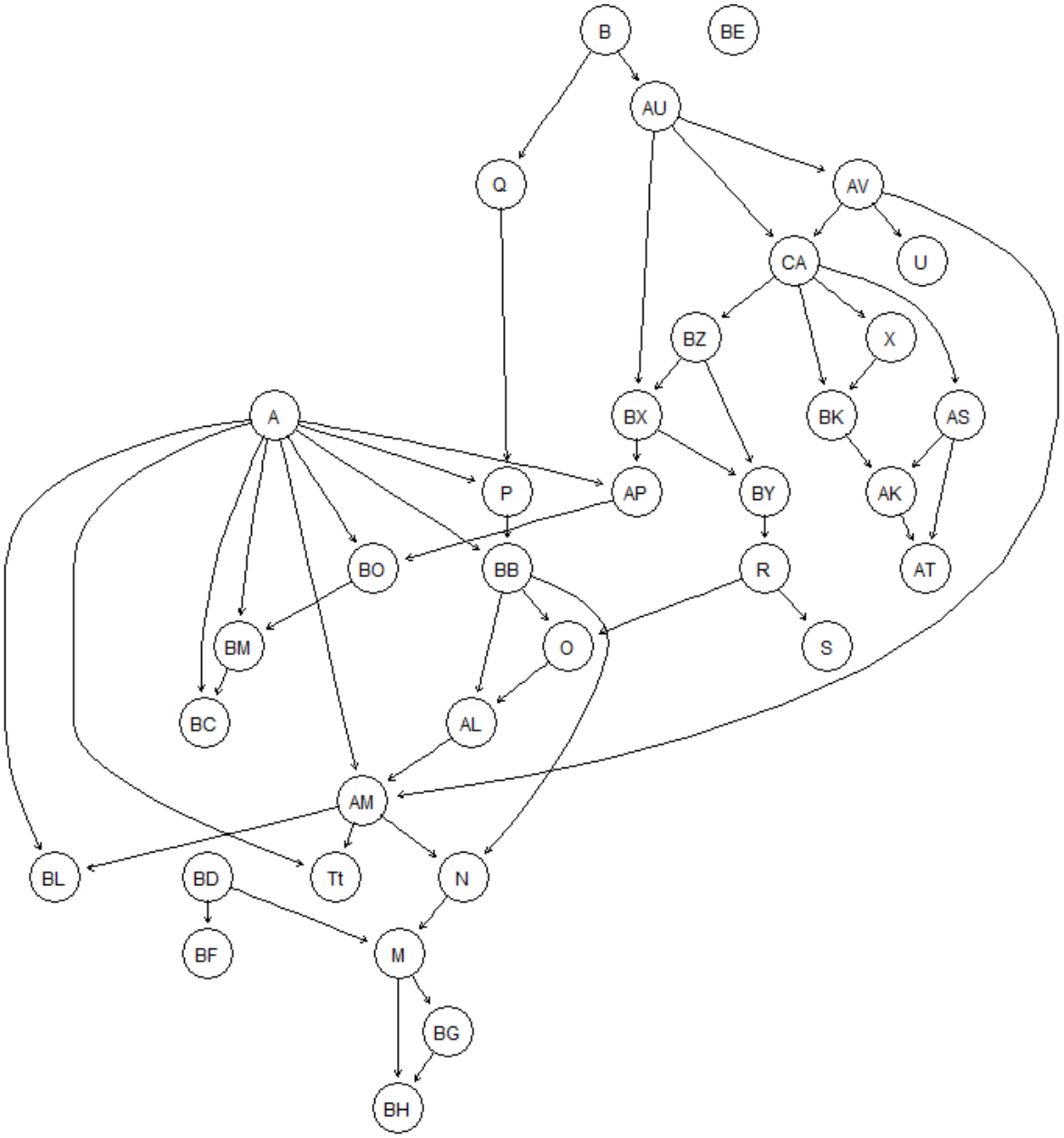


Figure 3: Bayesian Network for technical schools students



6. Final remarks

The 2016 report of the European Transport Safety Council claims that EU safety progress has come to a standstill. In addition, the decreasing trends in general accidentality are uniformly distributed across age groups, signaling that the relative relevance of the phenomenon for young drivers - overrepresented in road traffic death statistics by a factor between 1.2 and 3.9 - remains steady over the time.

This study aims at deepening the knowledge of factors that influence adolescents' risky behavior on the road. Bayesian Networks offer a promising new way to looking at the issue. In the analysis of a dataset collected in Tuscany, Italy, called EDIT, we found evidence that the use of alcohol and illegal substances explain only part of the probability of having an accident, and that other observable variables, like the level of distress or the type of school attended are significantly related to the probability of incurring in a road crash.

In line with the idea that *"Coordinated investments in adolescent health and wellbeing provide high economic and social returns and are among the best investments that can be made by the human community to achieve the UN's Sustainable development Goals and the Global Strategy for Women's, Children and Adolescents' Health" (The Lancet, 2017)*, we claim that when the probability of road accidents for very young drivers is under study, new and close attention should be given to a systemic approach and to a plethora of environmental and individual variables.

It is important to underline that the networks under scrutiny in this paper should not be considered as causal networks of general applicability. Indeed they are useful both to calculate the probability of future observations given values of variables describing a scenery of interest (predictive inference) or to perform probabilistic imputation of missing values for a given student of interest (diagnostic reasoning). This is particularly true if future observations are exchangeable with the EDIT dataset. If sudden changes happen in the population of students, for example as regards their habits due to viral advertisings or tough law enforcements, past data could become of limited use, at least before building proper model extensions. Similarly, the context of EDIT is Tuscany, thus the transportability of inference towards other regions or nations should not be assumed without inspection (external validity).

Policy formulation and intervention are further potential very important uses for a Bayesian network, for example with the aim of lowering novice drivers' high accidentally. It is well known that observational studies are exposed to the possibility of spurious associations between variables, for example for the exclusion of a variable from the study which is the common cause of two or more variables included into the study. Further work should be addressed to the extension of our networks towards full causal models, starting from the consideration of omitted potential confounding variables. Here we notice that EDIT was not designed with this primary goal thus we envisage the need of further sources of information to achieve this goal.

Acknowledgements

The authors would like to express their gratitude to the Health Regional Agency in Tuscany (ARS Toscana) for sharing part of the EDIT dataset, and in particular to Dr. Fabio Voeller and Drs. Lisa Gnaulati.

References

Adminaite, D. Jost, Stipdonk, G., & H & Ward, H. (2017). Ranking EU progress on road safety: 10th road safety Performance Index (PIN) report.

Atchison, L. (2017). Reducing casualties involving young drivers and riders in Europe.

Caron, J., & Liu, A. (2011). Factors associated with psychological distress in the Canadian population: a comparison of low-income and non low-income sub-groups. *Community Mental Health Journal*, *47*(3), 318-330.

Castillo-Manzano, J. I., Castro-Nuno, M., & Fageda, X. (2014). Could being in the European Union save lives? An econometric analysis of the Common Road Safety Policy for the EU-27. *Journal of European Public Policy*, *21*(2), 211-229.

Castillo-Manzano, J. I., Castro-Nuño, M., & Fageda, X. (2014). Can health public expenditure reduce the tragic consequences of road traffic accidents? The EU-27 experience. *The European Journal of Health Economics*, *15*(6), 645-652.

Castillo-Manzano, J. I., Castro-Nuño, M., & Pedregal, D. J. (2014). The trend towards convergence in road accident fatality rates in Europe: The contributions of non-economic variables. *Transport Policy*, *35*, 229-240.

Cleary, P. D., & Mechanic, D. (1983). Sex differences in psychological distress among married people. *Journal of Health and Social Behavior*, 111-121.

Di Clemente, R. J., Santelli, J. S., & Crosby, R. A. (Eds.). (2009). *Adolescent health: Understanding and preventing risk behaviors*. John Wiley & Sons.

Dohrenwend, B. P., & Dohrenwend, B. S. (1982). Perspectives on the past and future of psychiatric epidemiology. The 1981 Rema Lapouse Lecture. *American Journal of Public Health*, *72*(11), 1271-1279.

Drapeau, A., Beaulieu-Prévost, D., Marchand, A., Boyer, R., Préville, M., & Kairouz, S. (2010). A life-course and time perspective on the construct validity of psychological distress in women and men. Measurement invariance of the K6 across gender. *BMC Medical Research Methodology*, *10*(1), 68.

Drapeau, A., Marchand, A., & Beaulieu-Prévost, D. (2012). Epidemiology of psychological distress. In *Mental illnesses-understanding, prediction and control*. InTech.

Elba, F., Gnaulati, L., and Voeller, F. (2017). Reti bayesiane per lo studio del fenomeno degli incidenti stradali tra i giovani in Toscana. arXiv preprint arXiv:1710.07066.

Elvik, R., & Lund, J. (2015). Contributing factors to traffic injuries in adolescentsJohan Lund. *European Journal of Public Health*, *25*(suppl_3).

EC–European Commission. (2010). Towards a European road safety area: policy orientations on road safety 2011-2020. *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, COM (2010)*, *389*.

European Commission, SafetyNet (2009) Novice Drivers, retrieved <October 25, 2017>.

Gentry, J., Long, L., Gentleman, R., Falcon, S., Hahne, F., Sarkar, D., & Rgraphviz, K. H. (2009). Provides plotting capabilities for R graph objects. *R package version*, *2*(0).

Green, F. (2000). *Youth Road Safety* (No. AP-R159/00).

Grimm, M., & Treibich, C. (2013). Determinants of road traffic crash fatalities across Indian states. *Health economics*, *22*(8), 915-930.

Glied, S., & Pine, D. S. (2002). Consequences and correlates of adolescent depression. *Archives of pediatrics & adolescent medicine*, *156*(10), 1009-1014.

Hakkert, A. S., & Gitelman, V. (2014). Thinking about the history of road safety research: Past achievements and future challenges. *Transportation research part F: traffic psychology and behaviour*, *25*, 137-149.

Horwitz, A. V. (2002). Selecting Outcomes for the Sociology of Mental Health: Issues of Measurement and Dimensionality. *Journal of Health and Social Behavior*, *43*(2), 143-253.

Horwitz, A. V. (2007). Distinguishing distress from disorder as psychological outcomes of stressful social arrangements. *Health:*, *11*(3), 273-289.

Højsgaard, S. (2012). Graphical independence networks with the gRain package for R. *Journal of Statistical Software*, *46*(10), 1-26.

Hours, M., Chossegros, L., Charnay, P., Tardy, H., Nhac-Vu, H. T., Boisson, D., & Laumon, B. (2013). Outcomes one year after a road accident: results from the ESPARR cohort. *Accident Analysis & Prevention*, *50*, 92-102.

Kessler, R. C., Andrews, G., Colpe, L. J., Hiripi, E., Mroczek, D. K., Normand, S. L., ... & Zaslavsky, A. M. (2002). Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychological medicine*, *32*(6), 959-976.

Kessler, R. C., Barker, P. R., Colpe, L. J., Epstein, J. F., Gfroerer, J. C., Hiripi, E., ... & Zaslavsky, A. M. (2003). Screening for serious mental illness in the general population. *Archives of general psychiatry*, *60*(2), 184-189.

Johansson-Stenman, O., & Martinsson, P. (2008). Are some lives more valuable? An ethical preferences approach. *Journal of Health Economics*, *27*(3), 739-752.

Hughes, C. C. (1990). Rethinking Psychiatry: From Cultural Category to Personal Experience. *The Journal of Nervous and Mental Disease*, *178*(3), 210-211.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Marchetti, G. M., Drton, M., & Sadeghi, K. (2015). ggm: Functions for graphical Markov models. *R package version*, *2*.

OECD, Health at a glance, 2015

Payton, A. R. (2009). Mental health, mental illness, and psychological distress: same continuum or distinct phenomena?. *Journal of health and Social Behavior*, *50*(2), 213-227.

Podda, F. (2012). Drink driving: Towards zero tolerance.

Russel, J., Norving, P., 2003. Artificial Intelligence: A Modern Approach (2nd ed.), Upper Saddle River, New Jersey: Prentice Hall.

Scott-Parker, B., Goode, N., & Salmon, P. (2015). The driver, the road, the rules… and the rest? A systems-based approach to young driver road safety. *Accident Analysis & Prevention*, *74*, 297-305.

Scutari, M. (2009). Learning Bayesian networks with the bnlearn R package. *arXiv preprint arXiv:0908.3817*.

Sheehan, P., Sweeny, K., Rasmussen, B., Wils, A., Friedman, H. S., Mahon, J., & Stenberg, K. (2017). Building the foundations for sustainable development: a case for global investment in the capabilities of adolescents. *The Lancet*, *390* (10104), 1792-1806.

Wheaton, B. (2007). The twain meet: distress, disorder and the continuing conundrum of categories (comment on Horwitz). *Health:*, *11*(3), 303-319.

World Health Organization. (2007). Youth and road safety.

World Health Organization. (2009). Global status report on road safety: time for action.

World Health Organization. (2013). Global status report on road safety 2013: supporting a decade of action: summary.

World Health Organization. (2015). *Global status report on road safety 2015*. World Health Organization.