A Bayesian Decision-Support Tool for Child Sexual Abuse Assessment and Investigation

Tadei, Alessandro

2019-06


http://hdl.handle.net/10138/240786
https://doi.org/10.1177/1079063217732791

unspecified
acceptedVersion

Downloaded from Helda, University of Helsinki institutional repository.

This is an electronic reprint of the original article.

This reprint may differ from the original in pagination and typographic detail.

Please cite the original version.
A Bayesian Decision-Support Tool for Child Sexual Abuse Assessment and Investigation

Alessandro Tadei1*, Johan Pensar2, Jukka Corander3, Katarina Finnilä1, Pekka Santtila1, Jan Antfolk1

1 Department of Psychology and Logopedics, Åbo Akademi University, Turku, Finland
2 Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland
3 Department of Biostatistics, University of Oslo, Oslo, Norway.

* Corresponding author

Alessandro Tadei, E-mail: atadei@abo.fi. Address: Fabriksgatan 2, 20500, Turku, Finland.
Abstract

In assessments of child sexual abuse (CSA) allegations, informative background information is often overlooked or not used properly. We therefore created and tested an instrument that uses accessible background information to calculate the probability of a child being a CSA victim that can be used as a starting point in the following investigation. Studying 903 demographic and socio-economic variables from over 11,000 Finnish children, we identified 42 features related to CSA. Using Bayesian logic to calculate the probability of abuse, our instrument - the Finnish Investigative Instrument of Child Sexual Abuse (FICSA) - has two separate profiles for boys and girls. A cross-validation procedure suggested excellent diagnostic utility ($AUC = .97$ for boys and $AUC = .88$ for girls). We conclude that the presented method can be useful in forensic assessments of CSA allegations by adding a reliable statistical approach to considering background information, and to support clinical decision making and guide investigative efforts.

Keywords: Child Sexual Abuse, Multimodal assessments, Bayesian logic, CSA investigation, Decision making
Since the beginning of the 1980’s, the number of Child Sexual Abuse (CSA) allegations has increased in many Western countries (Ceci & Bruck, 1995; Kauppinen, Sariola, & Taskinen, 2000). Research suggests that both under- and over-reporting of CSA is a concern (Ceci & Bruck, 1995; Lipian, Mills, & Brantman, 2004; London, Bruck, Ceci, & Shuman, 2005; Svedin & Back, 2003). Underreporting refers to the fact that many cases of CSA never come to the authorities' attention. One reason for this is that some abused children do not disclose the abuse to adults and/or that adults fail to report such disclosures (London et al., 2005; Svedin & Back, 2003). Naturally, the actual extent of this problem is difficult to reliably estimate. On the other hand, at least 5%-35% of CSA allegations are likely unfounded, suggesting over-reporting to also be very common (Ceci & Bruck, 1995). A recent study suggests that approximately 40% of CSA allegations in Finland are unfounded (Korkman, Antfolk, Fagerlund, & Santtila, n.d.). Unfounded allegations can, for example, stem from misinterpretations of physical, psychological or behavioral symptoms displayed by the child, or from attention seeking and mental illness (O’Neal, Spohn, Tellis, & White, 2014).

When investigating CSA allegations, the police frequently turn to experts for forensic interviewing of the alleged victim and for evaluating behavioral or psychological symptoms. Often, these experts also appear before court to provide expert testimony of relevance to the final judicial decision regarding abuse (Gratz & Orsillo, 2006). Considering the importance of this expert testimony, expert evaluations need to be of the highest possible quality. Investigations of CSA allegations are, however, very challenging. Only a minority of allegations can be validated on the basis of physical evidence, such as pregnancy or the presence of semen (Muram, 2001). Approximately 70% of CSA allegations lack strong corroborating evidence (i.e., medical or other material evidence such as photographs and...
BAYESIAN MODEL FOR CHILD SEXUAL ABUSE INVESTIGATION

Contrary to what is commonly believed, unique behavioral symptoms are not valid indicators of CSA (Drach, Wientzen, & Ricci, 2001). Another widely held belief is that the child’s report is a very reliable source of information for deciding whether abuse has taken place or not (Berliner & Conte, 1993; Lamb, 1994; Poole & Lamb, 1998). The rationale behind this belief is that deliberate fabrication by children is rare and that false allegations are more likely to emanate from adults. Nevertheless, children are suggestible and may provide inaccurate information, especially when young and/or when considerable time has passed after the alleged event. While most experts agree that children can provide accurate reports, there is little doubt that a child’s account can be distorted both by improper interviewing or by normal memory decay (Ceci & Bruck, 1995).

Unsurprisingly, there is some controversy regarding how well experts perform as decision makers in allegations of CSA. Research suggests that expert evaluations are of alarmingly poor quality both in Finland (Korkman, Santtila, Westeråker, & Sandnabba, 2008; Korkman, Santtila, & Sandnabba, 2006) and in other countries (Cederborg, Orbach, Sternberg, & Lamb, 2000; Davies, Westcott, & Horan, 2000), and that expert decisions vary to such an extent that they cannot be considered reliable (Dror & Cole, 2010; Horner, Guyer, & Kalter, 1993a, 1993b). Moreover, legal experts’ understanding of expert evaluations has also been investigated in two recent studies, and it is apparent that these assessments are laden with serious problems (Tadei, Finnilä, Korkman, Salo, & Santtila, 2014; Tadei, Finnilä, Reite, Antfolk, & Santtila, 2016).

Bayesian Reasoning in Assessments of Allegations of Child Sexual Abuse

In CSA assessments, the expert needs the ability to evaluate and integrate complex and sometimes contradictory information to arrive at a conclusion. Research indicates that the reliability of clinical decisions is lower than the reliability of decisions based on actuarial (i.e., statistical) data (Dawes, 1994; Dawes, Faust, & Meehl, 1989; Goldberg, Faust, Kleinmuntz, 1989; Goldberg, Faust, Kleinmuntz, 1989).
& Dawes, 1991; Janus & Prentky, 2003). Because human cognitive capacity of using statistical information is limited (Edwards & von Winterfeldt, 1986; Tversky & Kahneman, 1974), methods to deal with complex actuarial information might prove helpful in forensic decision making (Kochenderfer, 2015). Uncertainty about whether an event has occurred or not should be represented in terms of probabilities, and these probabilities should be adjusted based on new information (Baron, 1994; Dammeyer, 1998; Kuehnle, 1998). For example, information retrieved during the assessment should update the baseline probability of abuse. But the baseline probability should also consider the particular features of the child under investigation, and this probability could be used as the best possible starting point for further investigation. Statistical models based on Bayes’ Theorem are particularly well suited for this (Herman, 2005; Wood, 1996). Bayes’ Theorem assumes the baseline probability of an event is updated as new information becomes available. For example, let us assume the probability that a randomly picked 12-year-old Finnish girl has been sexually abused. According to the most recent population-based victimization (used in the present study), the baseline likelihood of her having been abused, when nothing else is known, is approximately .03. This is then the starting probability that has to be updated for each new piece of information that becomes available. Let us further imagine that the girl exhibits symptoms of anxiety. If we know that anxiety is exhibited by 60% of abused children and by 30% of non-abused children being assessed, how likely is it that she has been abused in light of her anxiety symptoms? To understand how the probability that she has been abused is updated given this new information, we can consider a pool of 1000 girls of her age. Of these girls, 30 (3.0%) would have been abused and, of these girls, 18 (60%) would show symptoms of anxiety. Of the 1000 girls, 970 (97%) would not have been abused and, of these girls, 291 (30%) would show anxiety. Hence, 309 (18 + 291) girls show anxiety but only 18 (5.8%) have been abused. This means that, compared to the baseline probability (.03), the observation of anxiety almost
doubles the probability of the girl having been abused. With \( H \) meaning hypothesis (the girl has been abused), and \( E \) meaning evidence (the girl shows anxiety symptoms), the reasoning can be expressed as follows

\[
p(H|E) = \frac{p(E|H)p(H)}{p(E)} = \frac{0.6 \times 0.03}{0.018 + 0.291} \approx 0.058
\]

Nevertheless, .058 is still a low probability and the option that she has not been abused is much more probable (.942). For each new piece of information, the probability of abuse can be updated.

**The Current Study**

Many studies have demonstrated that not only laypeople, but also people with a background in behavioral sciences, often find it difficult to use statistical information correctly (Lehman, Lempert, & Nisbett, 1988; Levett, Danielsen, Kovera, & Cutler, 2005; Vidmar, 2005). Some authors suggest this is due to the use of formulas and abstract probabilities (Cosmides & Tooby, 1996; Gigerenzer, 2002), while other authors think the underlying theoretical concepts are difficult to understand for non-mathematicians (Fenton & Neil, 2011). Multi-modal assessments, that is, assessments using multiple sources of information, necessitate a meticulous overview and integration information from available (and missing) evidence. To deal with the problems described and improve the reliability in assessments of CSA allegations, we aimed to develop a computerized statistical-model that calculates the starting probability of a suspicion of CSA being true (i.e., that the child in question actually has been abused). To do this we used available victimization data from a large representative sample in Finland. Data retrieved from these victimization data were analyzed using Bayesian statistics.

**Methods**

**Context and Sample**
All data used in our research were part of a larger project focusing on different forms of violence against children and adolescents in Finland. The project was managed by the Police University College in collaboration with the Finnish Youth Research Society (Ellonen, Fagerlund, Kääriäinen, Peltola, & Sariola, 2013). The most recent data collection was carried out in 2013, and is the third of its kind. Two other data collections were carried out in 1988 (Sariola, 1988) and in 2008 (Ellonen, Kääriäinen, Salmi, & Sariola, 2008). The project aimed to quantify the prevalence of several phenomena, such as crime experiences, peer victimization, cyber-bullying, and sexual abuse. Alongside these phenomena, the data also include demographic data of the child and its family, and personal matters (e.g., health status, family employment situation, use of drugs or alcohol). Because research shows that the prevalence and nature of child sexual abuse has changed over the years (Laaksonen et al., 2011), we decided to use only data from the 2013 collection.

Participants

The 2013 data collection included reports from 11,364 respondents in total. The participants attended either the 6th or the 9th grade in different schools around Finland. To gather a representative sample of 6th and 9th graders, a stratified cluster sampling procedure with the school class as sample unit, was employed. The sample stratification was made according to province, municipality, and school size. Because the samples were drawn using the class information of the year before the data collections, the authors were unable to calculate the exact response rate (for more details about the dataset, see Ellonen et al., 2013). The response rate is instead known for the two previous versions of the questionnaire (89% in 1988; 88% of 6th and 64% of 9th grade pupils in 2008).

From the original dataset, we excluded 566 respondents because they had not provided answers for key questions related to sexual abuse, 33 respondents because they did not answer the questions related to their and/or the offender’s age at the time of the abuse, and 100
respondents because the participant either (i) declared to have been abused but they reported to be older at the time of abuse than they were at the time of the survey, or (ii) declared to have been abused, but the age difference between them and the offender was below 5 years, although this minimum age difference was explicitly formulated in the question. Thus, our final dataset consisted of reports from 10,665 children. Of the participants, 51.3% \((n = 5451)\) were girls and 48.7% \((n = 5184)\) were boys (30 children did not report their gender).

**Measures and Definitions**

We defined child sexual abuse as the occurrence of one or more experiences of the events listed in Table 1 before age 17, along with at least a 5-year age difference between the victim and the offender. We acknowledge that this operationalization also defined consensual sexual relation with a 5-years older person as abuse. This could have been avoided by also considering answers to the question: “Did you see the situation as sexual abuse?” The negative consequence of this would have been not considering as victims the children, who, in a juridical meaning, would have been abused, without perceiving the event as sexual abuse. Subsequently, the final decision was to avoid false negatives (i.e., not to define a real CSA event as such) at the risk of false positives (i.e., define the event as CSA when it is not). We also want to point out that this decision does not invalidate the logic of the subsequent analyses. All subsequent analyses could be conducted with different operational definitions of abuse, depending on the legal context in which they are intended to be used.

**Statistical Analysis**

**Classifier.** The aim of a classifier is to determine to which category a designated class variable \(H\) belongs given some evidence provided by a set of feature variables \(E_1, ..., E_d\). For
a classifier to be accurate, the features need to be informative in discriminating between the
different classes of $H$. A classifier is typically trained in a supervised manner on labeled data
containing joint observations of the features and class variable. In our study, we wanted to
investigate if CSA can reliably be predicted given self-reported information about the child’s
family, relations with peers, past experiences with crime events, use of drugs or alcohol and
other life events unrelated to the possible CSA. The data were extracted from questionnaires
and represented as categorical variables.

In our procedure we considered a probabilistic classifier, which assigns the class label
to the category with the highest conditional probability given the values of the feature
variables. In addition to specifying the most probable class, the updated class distribution also
provides valuable information regarding the uncertainty of the proposed class label. Using
Bayes’ theorem, the classification criterion of assigning $H = h$ given the feature values $E_1 =
e_1, \ldots, E_d = e_d$ can be inferred from a generative model and formulated as
\[
\arg \max_{h \in \mathcal{H}} \{p(E_1 = e_1, \ldots, E_d = e_d | H = h) p(H = h)\},
\]
(1)
where $\mathcal{H}$ is the set of class categories. Without additional assumptions, a practical limitation
of (1) is that the number of parameters required to model the joint feature distribution grows
exponentially with the number of features. The most common way of dealing with this is to
assume a naïve Bayes structure. A naïve Bayes classifier assumes that all feature variables are
conditionally independent given the class variable, allowing a factorization of the joint feature
distribution according to
\[
p(E_1, \ldots, E_d | H) = \prod_{i=1}^{d} p(E_i | H),
\]
(2)
In other words, the high-dimensional multivariate distribution in (1) can be modeled as
a product of $d$ low-dimensional univariate distributions. In a broader perspective, the naïve
Bayes classifier falls in the graphical model family of Bayesian networks (directed graphical
networks).
models), which have been developed as a means to model multivariate distributions (see Koller & Friedman, 2009, for an in-depth overview of the theory of Bayesian networks). One of the key features of Bayesian networks is the use of directed acyclic graphs to compactly represent and visualize complex dependence structures over collections of variables. In particular, the dependence structure assumed by the naïve Bayes classifier is depicted in Figure 1(a).

In addition to being highly scalable, the naïve Bayes model is very efficient in terms of handling missing data. In fact, summing out a non-observed feature variable is equivalent to simply omitting the corresponding conditional distribution from (2). This particular property is very useful in terms of updating the class distribution and performing feature selection, especially in our case where the training data has been collected through questionnaires containing many non-answered questions.

Rather than using a single naïve Bayes classifier, we propose using a collection of profile-specific classifiers. In this approach, there is a different model for each context specified by a profile variable. This enables modeling situations where the feature distributions are expected to differ between the profiles. When taking a profile variable into account, the classification criterion can be formulated as

$$\arg \max_{h \in \mathcal{H}} \left[ p(S = s)p(H = h|S = s) \prod_{i=1}^{d} p(E_i = e_i|H = h, S = s) \right].$$

Note that the features are still conditionally independent given the class and profile variables. In Figure 1(b), the original naïve Bayes structure in Figure 1(a) has been updated to illustrate the additional profile variable.
**Feature selection.** Feature selection is the process of selecting an informative subset of non-redundant features from a set candidate features. An identified feature is informative if it is useful for discriminating between the class categories. In addition to being marginally informative, a good feature should also be non-redundant in the sense that it should remain informative given the other features. In particular, the joint effect of several strongly correlated features is easily over-emphasized in a naïve Bayes classifier, since the joint effect is approximated as the sum of the individual effects. In this work, we take inspiration from the structure learning field of Bayesian networks and propose a feature selection technique that is designed to select features that are both informative and non-redundant.

Let $bf(E; H|E')$ denote the log Bayes factor that measures the support of a dependence between $E$ and $H$ in the presence of $E'$. In other words,

$$bf(E; H|E') = m(E|E', H) - m(E|E'),$$

where $m(E|E', H)$ is the log marginal likelihood of $E$ given $E'$ and $H$, and $m(E|E')$ is the log marginal likelihood of $E$ given $E'$ (see Heckerman, Geiger, & Chickering, 1995). Values above zero support $E$ and $H$ being dependent given $E'$, and the higher the value the stronger is the support. Analogously, values below zero support conditional independence between $E$ and $H$ given $E'$.

Furthermore, let $E_{\text{cand}}$ denote a set of candidate features and let $E_{\text{curr}} \subseteq E_{\text{cand}}$ denote a subset of features already selected from the candidate set. The score of a candidate feature $E \in E_{\text{cand}} \setminus E_{\text{curr}}$ is then defined as

$$\text{score}(E|E_{\text{curr}}) = \min_{E' \in E_{\text{curr}}}[bf(E; H|E')].$$

The score will thus favor features that are informative also in the presence of each of the already added features. If no feature has been added, the above score is reduced to measuring marginal dependence between $E$ and $H$ as quantified by $bf(E; H)$. 
Starting from an empty set of features, we use the following greedy search strategy for selecting a set of features:

1. Select $E^* = \arg \max_{E \in E_{\text{cand}} \setminus E_{\text{curr}}} [\text{score}(E|E_{\text{curr}})]$.

2. IF $\text{score}(E^*|E_{\text{curr}}) > 0$: add $E^*$ to $E_{\text{curr}}$ and repeat 1-2,

ELSE return $E_{\text{curr}}$.

To allow for the set of informative features being different in each profile, the data is split according to the different profiles and the feature selection algorithm is then run separately for each profile. Taking the profile-specific feature sets into account, our classification criterion can finally be formulated as

$$\arg \max_{h \in H} \left[ p(S = s)p(H = h|S = s) \prod_{i=1}^{d} p(E_i = e_i|H = h, S = s) I(E_i \in E(s)) \right],$$

where $E(s)$ is the feature set associated with profile $S = s$ and $I(E_i \in E(s))$ is an indicator function taking on value 1 if $E_i \in E(s)$ and value 0 if $E_i \notin E(s)$. A useful property of the above distribution is that it can still be captured by a single Bayesian network using the concept of context-specific independence (Boutilier, Friedman, Goldszmidt, & Koller, 1996; Pensar, Nyman, Koski, & Corander, 2015).

**Classifier validation.** To evaluate the performance of the classifier, we computed the Receiver Operating Characteristics (ROC) curves (Fawcett, 2004). A ROC curve is a graphical plot that illustrates the performance of a binary classifier by plotting the true positive rate against the false positive rate. By varying the threshold value for assigning the class label we obtain different points in ROC space. The complete curve can be obtained by letting the threshold value incrementally increase in small steps from 0 to 1, since the considered classifier returns a probability. To quantify the ROC curve in a single numerical measure we also calculated the Area Under the Curve (AUC). Since the AUC is a proportion of the area of the unit square, it will be between 0 and 1, however, no realistic classifier
should have an AUC lower than 0.5, which corresponds to random guessing. The AUC can be interpreted as the probability that a random positive instance is ranked higher than a random negative instance (assuming that positives rank higher than negatives). For a more robust assessment of the out-of-sample performance of the classifier, we used cross-validation where the data were randomly split into a training set and a test set of equal size. The classifier parameters were learned using the training data and the classifier was then evaluated on the holdout test data. To reduce variability, the cross-validation procedure was repeated 100 times, that is, the final results are averages over 100 ROC curves and corresponding AUCs.

**Instruments.** All numerically demanding procedures were implemented and run in MATLAB.

**Graphical Model.** The final model was also manually constructed in AgenaRisk (Fenton & Neil, 2012), which provides a more user-friendly graphical interface. Figure 2 shows an example of the Bayesian classifier in AgenaRisk and how data are entered. The classifier starts from the probability that the CSA suspicion is true given the base rate prevalence in the general population. After this, this probability is updated as more evidence is added.

We assume a starting situation where no observations have been made. Panel A gives the baseline probability of .03. In panel B, this probability is updated given that the girl replied *Yes* to the question “Have you ever tried drugs (e.g., hashish or ecstasy)?”. Here, the probability of having tried drugs is 0.194 if the child has been abused and 0.019 if the child has not been abused. In panel C, the probability is updated given that the child has dinner with the adults she lives with at least twice a week, and that she has not been victim of any
physical attack in the last 12 months. After considering the use of drugs, the frequency of
dinner with the family, and the absence of any recent physical attack, the probability of sexual
abuse is approximately .15.

After having selected the profile variable and the corresponding sets of feature
variables, we built a classifier in form of a Bayesian network in AgenaRisk. The network,
which follows the structure illustrated in Figure 1(b), is composed of the following variables,
which are known as nodes in a graphical model: a profile node, a set of feature nodes, and a
result (or class) node. The result node gives the probability that the assessed child has been
sexually abused. This node is linked to the profile node and to all the features in the classifier.
After specifying the profile, only nodes that are included in the feature set for that specific
profile can influence the distribution of the result node, while nodes outside the feature set
will have no influence. In practice, this is achieved by imposing regularities within the feature
distributions according to the concept of context-specific independence.

If the information is available, all nodes linked to a profile can be defined as True or
False or by choosing among the possible answers for the given classifier. For example, the
question “How often do you eat dinner with one or both of your parents (or those adults who
you live with)?” has two possible answers: “Several times a week or more” and “Once a week
or less”). When information is lacking a node can be left undefined. When new evidence is
inserted into the classifier, by assigning a value to a node, the classifier automatically updates
the probability of sexual abuse by performing inference in the model.

Results

We first calculated the CSA prevalence in Finland, dividing the sample into girls and
boys. The analyses indicate that the CSA base rate for boys (age 0-16) was .007 and 0.03 for
girls. We also experimented dividing the sample according both the gender and the age (0-12
and 13-16) but the analyses showed that using gender was sufficient (i.e., the age information did not increase the accuracy of the classifier).

After this, we conducted the feature selection. In total we identified 42 significant features that could be used to assess the probability that an assessed child has been sexually abused. We identified 28 girl-specific features, 17 boy-specific features, and 3 general features valid for both girls and boys. See Appendix for each feature, the name used in the classifier, the full question asked in the survey, and the Bayes factor, which is a measure of the marginal dependence between the feature and the class variable. The order of the features in the table is the same as the order in which they were selected by the feature selection method. The order a rough estimate of the importance of a feature in the presence of the other features.

As a final evaluation step, we calculated the ROC curves. These were calculated for both genders separately. The AUC values indicate that the FICSA shows excellent diagnostic performance (Figure 3). The performance is somewhat better for boys (AUC = .97) than for girls (AUC = .88).

The FICSA Graphical Model

Finally, Figure 4 shows the classifier built in AgenaRisk. The classifier is visually divided in three main parts: on the left the features valid only for girls, on the right the features valid only for boys, in the middle the features valid for both genders. The squared node at the top permits the user to select the gender of the assessed individual. After this selection, all features that are irrelevant for the chosen gender are disregarded by the software. The target node gives the final outcome probability for sexual abuse. Features are grouped
according to themes. For girls these are (in clockwise order) Friends, Family, Psychological status, Substance abuse, Cyber-violence, and Violent experiences. For boys these are (in clockwise order) Friends, Cyber-violence, Family, Sibling or peer violence, and Violent experiences. The feature theme shared between girls and boys is Sexual experiences.

To further clarify how the FICSA works, we decided to create two different fictive scenarios. Both scenarios describe a CSA allegation regarding 11-years old girls. Here, we use the features in FICSA to arrive at a probability that the girl has been abused. Table 2 shows the features of relevance for CSA allegations with girls as the possible victim. The full question related to each variable can be found in the Appendix.

Case 1 describes a girl who is severely bullied by schoolmates, both in person and through social networks. She is shy and has never reported her problems to anyone but her parents. She once kissed a boy from her class and is, since then, object of her female classmates’ jealousy. They bully her both in person and by phone messages and have threatened to hurt her physically. During the last month she has been bullied online by people both known and unknown to her. We know that her family eats dinner together every day and that she has never been victim of physical or psychological violence. The parents are considering to send her to another school, because the situation is affecting the girl’s wellbeing. She has difficulties sleeping and cannot focus on her studies. She never smoked or used drugs, but her parents let her have a glass of sparkling wine on special occasions.
Case 2 describes a girl who lives only with her father. He often leaves her alone during evenings and nights. When the father is not home, she usually hangs out with a group of slightly older boys who gather in the main square or in the park close to her home. The father frequently complains that she is not doing her household chores, and that her mother would be as disappointed in her as he is. Sometimes he calls her names too. She smokes, and she drinks almost weekly. She does not use the Internet. Finally, it has become known that she has satisfied some of her classmates sexual requests, and that she was victim of violent aggression 3 months ago.

In both cases the baseline probability of sexual abuse is .03, but the FICSA gives two different probabilities for abuse after updating the prior using case-specific features. In Case 1 the probability is .27. This is considerably higher than the baseline probability, but, conversely, the probability of no abuse is .73. In Case 2 the probability of abuse is .84 and the probability of no abuse is .16.

Discussion

The aim of this study was to develop and test a method using Bayesian logic that uses accessible background information to assist clinical decision making in CSA allegations. To do this we used demographical information from a large, representative population-based victimization study. Our final model included 42 features in total and showed excellent diagnostic utility. There are, however, some concerns that currently restrict its feasibility. Some of these concerns pertain to the model assumptions, while other concerns pertain to the validity and use of the estimated abuse probability.

Model Assumptions

Although the FICSA is based on sound logic, the outcome is also dependent on the validity of the premises, that is, the validity and reliability of the data entered into the
classifier. Here, a first concern is the time interval between the experienced CSA and participation in the questionnaire. If participation took place immediately after the event, we could be sure that all of a child’s answers described the situation before the abuse. In such a case, variables related to CSA, would be actual predictors. In the current study, instead, some of the variables that were included in the FICSA could also describe the situation after the event. For example, the feature “Drugs”, that refers to the question “Have you ever tried drugs (e.g. hashish or ecstasy)?” can also be a consequence of abuse (Hornor, 2010). This might appear to undermine the validity of the current method. We argue, however, that the chronological order of events has no impact in the decision making in assessments of CSA suspicions. If at the time of the assessment, a variable is known and this variable is reliably associated with experiencing abuse in the past (or future) it will contribute valid information. An example that clarifies this is the presence memories of the abuse, which, invariably would appear after the event and which are useful evidence in the investigation.

Another concern pertains to the operational definition of CSA, which in the current study includes a minimum age difference between victim and offender. To reflect Finnish legal practice, this age difference was set at 5 years in the victimization study. This definition excludes all sexual relations, in which the age difference is lower, and involves, for example, sexual relations between a 12-year old and a 14-year old. In such cases the law itself is less precise, and courts have to evaluate the nature of the event, the type of relationship between the individuals and so on. Similarly, all sexual relations involving a minor, where the age difference between the two parts is over 5 years are considered CSA, even if the parties themselves might view them as consensual. We also considered using the question: “Did you see the situation as sexual abuse?” as a criterion. In this way it would have been possible to exclude sexual acts perceived to be consensual and get a stricter definition of CSA. Because some children may not perceive an abusive act as sexual abuse due to, for example, cognitive
difficulties or being victims of manipulation, we refrained from using this criterion. Furthermore, an allegedly consensual relation could relatively easily be demonstrated during the investigation and the possible trial, and therefore it poses few practical problems.

**The Validity and Use of the Estimated Abuse Probability**

The FICSA gives a value ranging between 0 (abuse very unlikely) and 100 (abuse very likely) that expresses the probability that the CSA has actually taken place and the case can be substantiated. Because it is improbable that the result will be either 0 or 100, and in most cases the model will give a value somewhere in between these extremes, it is important to consider the interpretation of the value in attempting to reach a final decision. For this purpose, the ROC curves in Figure 3 provide valuable information by showing the estimated trade-offs between true and false positives that are obtained for different threshold class probabilities. Another important next step will be to compare the outcome of validated real cases (i.e., cases where an offender has credibly confessed or cases where the alleged victim has credibly retracted the allegation) with the results of the method, and assess the trade-off between false and true positives in actual assessed CSA allegations. Even in the cases where the classifier gives a very high value, it is vital to also consider the possibility that CSA has not yet taken place (and, of course, potentially never will). Because the FICSA rests on demographic and background variables, it is very important to give full consideration to other evidence. Only using demographic and background variables, our method demonstrated excellent capacity to discriminate between abused and not abused children. Depending on the gender, the AUC values ranged between .88 and .97. We conclude that the FICSA might contribute important information in assessments and investigations of child sexual abuse suspicions.

**Conclusions and Future Directions**
If the limitations described can be addressed and the method’s validity can be demonstrated, we argue that Bayesian classifiers provide a powerful tool that can consider multiple pieces of evidence used to providing a starting probability for CSA assessments, and support decision making in assessment of CSA allegations. It could also guide the investigative process towards also looking for and using hitherto overlooked background information. Importantly, we do not think FICSA should in any way replace the gathering and evaluation of traditional evidence. Rather, we think that FICSA can be used to provide a best-possible starting point for further investigation. As the relationship between information modeled here and other forms of information (e.g., medical assessments, interview-based evidence) becomes known, the FICSA could be updated to also take this information into account.

Another important step to take is to cross-validate the selection of variables (and their weights) in other representative samples. Gathering local data would also be necessary for the classifier to be useful in other populations. This is because it is likely that several indicators vary over time and place, and that indicators that are currently valid in Finland are not necessarily valid elsewhere.
References


BAYESIAN MODEL FOR CHILD SEXUAL ABUSE INVESTIGATION

York: Free Press.


BAYESIAN MODEL FOR CHILD SEXUAL ABUSE INVESTIGATION


## Table 1

### Events Considered Sexual Abuse

- Receive a proposal to do anything sexual
- Offender shows genitals
- Victim shows genitals
- Offender touches victim’s genitals over the clothes
- Victim touches offender’s genitals over the clothes
- Offender touches victim’s genitals under the clothes
- Victim touches offender’s genitals under the clothes
- Petting
- Imitation of sexual intercourse, without penetration
- Sexual intercourse, with penetration
Table 2

Two Examples of CSA Allegations Regarding an 11-Year Old Girl

<table>
<thead>
<tr>
<th>Variable</th>
<th>Case 1</th>
<th>Case 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Friends</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Friends age</td>
<td>Same age</td>
<td>Older but still &lt;18</td>
</tr>
<tr>
<td>Getting on with adults more than peers (6M)</td>
<td>Not true</td>
<td>Not true</td>
</tr>
<tr>
<td>Sexual experience with peer, no touch</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Public places 10pm-12pm</td>
<td>Never</td>
<td>Fairly often</td>
</tr>
<tr>
<td><strong>Family</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insulted by father (&gt;12M)</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Pushed or shaken by father (&gt;12M)</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Dinner in family</td>
<td>Several times a week or more</td>
<td>Once a week or less</td>
</tr>
<tr>
<td><strong>Psychological Status</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Worrying a lot (6M)</td>
<td>Certainly true</td>
<td>Somewhat true</td>
</tr>
<tr>
<td><strong>Addictions</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drugs</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Alcohol</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Smoke cigarettes</td>
<td>Have never smoked</td>
<td>Smoke, but not daily</td>
</tr>
<tr>
<td><strong>Cyber-Violence</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bullied or insulted by text messages (12M)</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Threatening messages by phone (12M)</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Sexually harassing messages by phone (12M)</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Bad rumors on Web (12M)</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Rude behavior from unknown on Web (12M)</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Sexy photos or videos requested on Web (12M)</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td><strong>Violent Experiences</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attack threat (12M)</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Event</td>
<td>12M</td>
<td>&gt;12M</td>
</tr>
<tr>
<td>--------------------------------------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>Attack threat (12M)</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Theft (12M)</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Theft (&gt;12M)</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Hit or attacked (12M)</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Hit or attacked (&gt;12M)</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Drunk or high when victim of theft</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Incident described to mother (12M)</td>
<td>No</td>
<td>/</td>
</tr>
<tr>
<td><strong>Sexual Experiences</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex with Bf or Gf</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Sex, anal or oral, real or attempted, with peers</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Sexual proposal on Web by unknown (12M)</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>
Figure 1. (a) The naïve Bayes dependence structure represented by a directed acyclic graph; (b) The profile-specific naïve Bayes structure represented by a directed acyclic graph.
Figure 2. Example of a simple Bayesian classifier in AgenaRisk.
Figure 3. ROC curves and Area Under the Curve values (mean ± standard deviation).
Figure 4. FICSA: A Bayesian network classifier for estimating the probability of Child Sexual Abuse.
Appendix

List of Significant Predictors of CSA. The Bayes factor value measures the marginal dependence between the feature and class variable. The features are ordered in the same order in which they were selected by the feature selection method.

<table>
<thead>
<tr>
<th>Feature name</th>
<th>Original question translated in English</th>
<th>Bayes Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alcohol</td>
<td>Have you ever consumed alcohol, for example half a bottle of beer, a glass of wine or a glass of spirits?</td>
<td>94.28</td>
</tr>
<tr>
<td>Friends Age</td>
<td>Are the friends you spend most time with...</td>
<td>53.02</td>
</tr>
<tr>
<td>Sexually harassing messages by phone 12M</td>
<td>Have you experienced any of the following in the past 12 months: Someone has sent you sexually harassing messages by phone</td>
<td>54.59</td>
</tr>
<tr>
<td>Sexy photos or videos requested on Web 12M</td>
<td>Have you experienced any of the following in the past 12 months: Has an unknown person asked you to send sexy photos or videos of yourself to him/her on the Internet?</td>
<td>63.08</td>
</tr>
<tr>
<td>Public places 10pm _ 12am</td>
<td>At what hours do you spend time in these public spaces: Between 10 p.m. and midnight</td>
<td>70.70</td>
</tr>
<tr>
<td>Attack threat 12M</td>
<td>In addition to incidents you have possibly mentioned above, has anyone only threatened to hit you or attack you in the past 12 months?</td>
<td>33.89</td>
</tr>
<tr>
<td>Sex, anal or oral, real or attempted, with peers</td>
<td>Sexual activity, intercourse with a peer - Have you ever had, or have any people of your age attempted to have intercourse or anal or oral sex with you?</td>
<td>25.78</td>
</tr>
<tr>
<td>Drugs</td>
<td>Have you ever tried drugs (e.g. hashish or ecstasy)?</td>
<td>40.73</td>
</tr>
<tr>
<td>Rude behavior from unknown on Web 12M</td>
<td>Have you experienced any of the following in the past 12 months: Has an unknown person behaved rudely towards you or used obscene language when you have talked to him/her on the Internet?</td>
<td>47.89</td>
</tr>
<tr>
<td>Smoke cigarettes</td>
<td>Do you smoke cigarettes?</td>
<td>77.37</td>
</tr>
<tr>
<td>Attack threat &gt;12M</td>
<td>Has anyone only threatened you with violence before this?</td>
<td>34.66</td>
</tr>
<tr>
<td>Bad rumors on Web 12M</td>
<td>Have you experienced any of the following in the past 12 months: Someone has spread rumours or written bad things about you on the Internet</td>
<td>31.93</td>
</tr>
</tbody>
</table>
### BAYESIAN MODEL FOR CHILD SEXUAL ABUSE INVESTIGATION

| **Theft >12M** | Has anyone stolen something from you without using force before this? | 22.52 |
| **Hit _ attacked 12M** | Has anyone hit you or attacked you in the past 12 months? Here you can also mention offenses that occurred during violent robberies. | 24.32 |
| **Hit _ attacked >12M** | Has anyone hit you or attacked you before this? | 26.85 |
| **Dinner in family** | How often do you eat with one or both of your parents (or those adults who you live with) in the evening? | 26.83 |
| **Sexual proposal on Web by unknown 12M** | Have you experienced any of the following in the past 12 months: Has an unknown person sexually propositioned you on the Internet? | 51.26 |
| **Worrying a lot 6M** | If you think about the past 6 months, how do the following apply to you: I worry a lot | 21.47 |
| **Insulted by father >12M** | When you had rows with your father, did he: Mock you, call you names, swear or otherwise hurt you emotionally but did not physically hurt you? | 21.02 |
| **Incident described to mother 12M** | Have you told someone about the most serious incident in the past 12 months: Mother | 21.27 |
| **Sexual experience with peer, no touch** | Sexual experiences with a peer - Have you ever experienced sexual things that did not involve actual physical touch with someone your age? | 18.61 |
| **Sex with Bf or Gf** | Sexual experiences with a peer - Have you ever had sex with your boyfriend or girlfriend? | 14.82 |
| **Getting on with adults more than peers 6M** | If you think about the past 6 months, how do the following apply to you: I get on better with adults than with people my own age | 19.34 |
| **Bullied or insulted by text messages 12M** | Have you experienced any of the following in the past 12 months: Someone has bullied you or called you names by text messages | 18.18 |
| **Theft 12M** | Has anyone stolen something from you without using force in the past 12 months? | 21.67 |
| **Drunk or high when victim of theft** | Theft - Were you under the influence of alcohol or other substances? Depending on "Has anyone ever stolen something from you without using force?" | 19.20 |
| **Pushed or shaken by father >12M** | When you had rows with your father, did he: Push, shove or shake you violently? | 13.36 |
| **Threatening messages by phone 12M** | Have you experienced any of the following in the past 12 months: Someone has sent you threatening messages by phone | 18.82 |

*Valid for boys*
<table>
<thead>
<tr>
<th>Question / Scenario</th>
<th>Response / Details</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Public places after 12am</td>
<td>At what hours do you spend time in these public spaces: After midnight</td>
<td>35.35</td>
</tr>
<tr>
<td>Sex, anal or oral, real or attempted, with peers</td>
<td>Sexual activity, intercourse with a peer - Have you ever had, or have any people of your age attempted to have intercourse or anal or oral sex with you?</td>
<td>24.92</td>
</tr>
<tr>
<td>Stealing 6M</td>
<td>If you think about the past 6 months, how do the following apply to you: I take things that are not mine from home, school or elsewhere</td>
<td>21.41</td>
</tr>
<tr>
<td>With how many peers sexual touching?</td>
<td>Sexual touching with a peer - With how many people have you had experiences like this? Depending on &quot;Have you ever experienced sexual touching with someone your age?&quot;</td>
<td>9.15</td>
</tr>
<tr>
<td>Perpetrator age for property damage</td>
<td>Property damage - How old was the perpetrator? Depending on &quot;Has anyone ever broken or ruined any of your things on purpose?&quot;</td>
<td>7.13</td>
</tr>
<tr>
<td>Encouraged to go to authorities If bullied on Web or phone</td>
<td>Harassment on the Internet or by phone - How did the person(s) you told react or what did they do: Encouraged me to seek help from authorities</td>
<td>10.20</td>
</tr>
<tr>
<td>Chocked or assaulted with knife or gun 12M</td>
<td>Battery/assault - What kind of violence was used: You were choked or assaulted with a knife or gun. Depending on &quot;Has anyone ever hit you or attacked you?&quot;</td>
<td>9.72</td>
</tr>
<tr>
<td>Robbery 12M</td>
<td>Has anyone used force to steal something from you in the past 12 months?</td>
<td>17.09</td>
</tr>
<tr>
<td>Sexual proposal on Web by unknown 12M</td>
<td>Have you experienced any of the following in the past 12 months: Has an unknown person sexually propositioned you on the Internet?</td>
<td>13.58</td>
</tr>
<tr>
<td>Parents drunk 12M</td>
<td>How often in the past 12 months have you seen your parents visibly drunk?</td>
<td>15.28</td>
</tr>
<tr>
<td>Handsome for others?</td>
<td>How well do the following apply to your experiences relating to your appearance: I often hear that I am beautiful or handsome</td>
<td>10.78</td>
</tr>
<tr>
<td>Sex with Bf or Gf</td>
<td>Sexual experiences with a peer - Have you ever had sex with your boyfriend or girlfriend?</td>
<td>17.23</td>
</tr>
<tr>
<td>Drunk or high when mocked by siblings_peers</td>
<td>Mocking by siblings/peers - Were you under the influence of alcohol or other substances? Depending on &quot;Has your sister, brother, or a peer ever called you names, said mean things to you, or said that they didn't want you around, which scared you or made you feel really bad?&quot;</td>
<td>10.64</td>
</tr>
<tr>
<td>Perpetrator gender if mocked by siblings_peers</td>
<td>Mocking by siblings/peers - Was the perpetrator a girl or a boy? Depending on &quot;Has your sister, brother, or a peer ever called you names, said mean things to you, or said that they didn't want you around, which scared you or made you feel really bad?&quot;</td>
<td>3.45</td>
</tr>
<tr>
<td>Event</td>
<td>Question</td>
<td>Frequency</td>
</tr>
<tr>
<td>------------------------------------------</td>
<td>--------------------------------------------------------------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>Physically injured if assaulted</td>
<td>Battery/assault - Were you hurt, were you physically injured? Depending on &quot;Has anyone ever hit you or attacked you?&quot;</td>
<td>11.59</td>
</tr>
<tr>
<td>Where attacked by siblings or peers 12M</td>
<td>Peer/sibling violence - Where were you hit or attacked by your sister, brother or a peer? Depending on &quot;Has your sister, brother or a peer ever hit you?&quot;</td>
<td>9.69</td>
</tr>
<tr>
<td>Where mocked by siblings_peers</td>
<td>Mocking by siblings/peers - Where did the mocking take place? Depending on &quot;Has your sister, brother, or a peer ever called you names, said mean things to you, or said that they didn’t want you around, which scared you or made you feel really bad?&quot;</td>
<td>10.31</td>
</tr>
</tbody>
</table>