UNIVERSITÉ DE MONTRÉAL

ESTIMATION QUANTITATIVE DU RISQUE LIÉ AUX MACHINES EN EXPLOITANT DES RAPPORTS D'ENQUÊTE D'ACCIDENT ET L'ANALYSE LOGIQUE DE DONNÉES

SABRINA JOCELYN

DÉPARTEMENT DE MATHÉMATIQUES ET DE GÉNIE INDUSTRIEL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

THÈSE PRÉSENTÉE EN VUE DE L'OBTENTION

DU DIPLÔME DE PHILOSOPHIAE DOCTOR

(GÉNIE INDUSTRIEL)

JUIN 2017

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Cette thèse intitulée :

ESTIMATION QUANTITATIVE DU RISQUE LIÉ AUX MACHINES EN EXPLOITANT DES RAPPORTS D'ENQUÊTE D'ACCIDENT ET L'ANALYSE LOGIQUE DE DONNÉES

présentée par : JOCELYN Sabrina

en vue de l'obtention du diplôme de : Philosophiae Doctor

a été dûment acceptée par le jury d'examen constitué de :

Mme DE MARCELLIS-WARIN Nathalie, Doctorat, présidente

M. CHINNIAH Yuvin, Ph. D., membre et directeur de recherche

M. OUALI Mohamed-Salah, Doctorat, membre et codirecteur de recherche

M. ADJENGUE Luc-Désiré, Ph. D., membre

M. LAMY Pascal, Docteur, membre externe

REMERCIEMENTS

Ce projet fut réalisable grâce à la contribution exceptionnelle de l'Institut de recherche Robert-Sauvé en santé et en sécurité du travail (IRSST) et à la bourse No. 141111 du Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG). Je remercie grandement ces deux organismes.

Mon parcours doctoral fut une aventure qui allait au-delà de la conduite et de la réalisation d'un projet de recherche original apportant une contribution scientifique. Ce fut un bel exercice d'apprentissage de travail dans un contexte interdisciplinaire. Ce fut également une expérience de développement personnel et une belle occasion de discuter avec des professeurs et des étudiants qui me faisaient découvrir des façons de penser ou procéder différentes et enrichissantes. Parmi ces professeurs, je tiens à remercier pour leur confiance à mon égard :

- mon directeur de thèse, M. Yuvin CHINNIAH et mon codirecteur de thèse, M. Mohamed-Salah OUALI. Leur encadrement et le partage de leurs expériences professionnelles m'ont outillée pour mon doctorat et m'outilleront dans ma carrière en recherche;
- Mme Soumaya YACOUT, professeur titulaire à Polytechnique Montréal. Sa précieuse collaboration a permis la réalisation d'une étape décisive de mon projet de thèse.

Parmi les étudiants, je tiens à remercier spécialement : Mlle Afrooz MOATARI-KAZEROUNI, Mlle Diana LÓPEZ-SOTO, M. Bannour SOUILAH et M. Taha BELMEKKI. Leur promptitude à m'apporter des éclaircissements relatifs à des domaines techniques rattachés à ma thèse a contribué à l'avancement de mes recherches.

Un merci spécial à Chantal TELLIER de la Direction de la recherche et de l'expertise de l'IRSST qui, au tout début de mon parcours doctoral, m'a consacré plusieurs heures pour m'expliquer l'organisation de la base de données dépersonnalisée d'accidents réalisée à l'IRSST.

Finalement, je remercie mon mari, mes parents et ma sœur pour leur support inestimable, leurs encouragements continus et leurs nombreux conseils tout au long du doctorat. Merci du fond du cœur.

RÉSUMÉ

Les préventionnistes en sécurité des machines utilisent différents outils, dont des rapports d'enquêtes d'accidents, pour les aider dans l'identification des risques en milieu de travail. L'information est alors extraite ponctuellement, en lisant un rapport à la fois. Par la suite, les rapports consultés risquent de sombrer dans l'oubli.

En matière de gestion du risque, l'identification du risque est succédée par l'estimation du risque. Les préventionnistes en sécurité des machines utilisent généralement des outils qualitatifs pour estimer le risque. Cet aspect qualitatif crée de la subjectivité dans les prises de décision quant aux moyens de réduction du risque. De plus, la nature statique de ces outils contraint son utilisateur à des paramètres du risque prédéterminés. Si d'autres paramètres sont requis pour mieux définir le risque qui a évolué, ces outils seront incapables de les intégrer. Cela peut conduire à des décisions inadaptées en matière de réduction du risque.

Pour pallier ces inconvénients, cette thèse vise à proposer une démarche d'identification et d'estimation du risque qui facilite le suivi des risques liés aux machines, ainsi qu'à leur environnement physique et organisationnel en milieu de travail. La démarche utilise le retour d'expérience (REX) dynamique pour exploiter efficacement et durablement les rapports d'enquête d'accident. Le REX dynamique est à la fois un processus de remontée d'information et d'inférence de connaissances. La connaissance essentielle est extraite à partir de l'information contenue dans les rapports, après que l'information ait été formalisée dans une base de données. Cette connaissance peut être actualisée au fur et à mesure de la mise à jour de la base de données par la remontée d'information. La connaissance est inférée sous la forme de règles pertinentes générées par un algorithme de fouille de données. Une règle est une combinaison de conditions décrivant des accidents appartenant à un même ensemble, appelé « classe ». Chaque condition se compose d'un indicateur auquel une valeur ou une plage de valeurs est affectée. Un indicateur est un facteur de risque ou une cause potentielle d'accident. Ainsi, avec le REX dynamique utilisant une base de données pouvant être mise à jour régulièrement, les connaissances issues des rapports seront continuellement mises à profit et évolueront avec le contexte.

Un algorithme d'apprentissage automatique nommé « Analyse logique de données (ALD) » (*Logical Analysis of Data* : LAD) est intégré au REX dynamique pour assurer que la démarche proposée fonctionne même pour un échantillon restreint de données. En effet, cette thèse a

démontré que, pour un petit échantillon de 23 accidents liés à des convoyeurs à courroie, l'ALD est capable de générer des règles avec une précision de classification adéquate : entre 72% et 74%. Le choix des convoyeurs à courroie s'appuie sur deux constats. Premièrement, de tous les types de convoyeurs, ceux à courroie ont provoqué le plus d'accidents (16,8%) entre 1990 et 2011, d'après 137 rapports d'accidents de la Commission des normes, de l'équité, de la santé et de la sécurité du travail (CNESST) liés à des convoyeurs. Deuxièmement, ce type de convoyeurs représente la plus grande proportion (8,5%) des accidents graves et mortels, toutes machines confondues, entre 1999 et 2007, d'après une base de données de l'Institut de recherche Robert-Sauvé en santé et en sécurité du travail (IRSST).

Les 23 rapports d'accidents traités dans cette thèse, proviennent du Centre de documentation de la CNESST. Une analyse de l'information de chaque rapport a permis de tirer les éléments décrivant le contexte accidentel. Ce traitement d'information a donné naissance à une base de données à partir de laquelle l'ALD a généré deux séries de règles. D'abord, l'une pour une version de la base de données divisant les 23 accidents en une classe d'accidents en maintenance et une classe d'accidents en production. Ensuite, l'autre pour une version de la base de données partageant les 23 accidents en deux classes : « Non mortel » et « Mortel ».

Certaines des règles générées ont montré qu'un accident peut survenir en raison de conditions dangereuses (ex., un environnement de travail encombré), mais aussi en présence de conditions d'apparence sécuritaire (ex., l'existence d'un programme de prévention). Dans ce dernier cas, il faut investiguer pour comprendre les dessous d'une condition qui semble sécuritaire. Par exemple, pour 60% des accidents en maintenance survenus en dépit de l'existence d'un programme de prévention, l'omission de sa mise à jour pourrait expliquer l'accident. D'autres règles ont montré que les accidents analysés s'expliquent principalement par des facteurs de risque ou causes rattachées à l'équipement, l'organisation, l'individu, ou le moment.

Des paramètres quantitatifs associés aux règles, tels que leurs couvertures et la fréquence de leurs indicateurs, ont permis d'entamer la hiérarchisation des règles et des facteurs de risques (la couverture est le nombre d'accidents que décrit la règle). Une méthode développée pour estimer la probabilité du dommage associé à chaque règle a permis de compléter la hiérarchisation des règles de couvertures identiques. Cette hiérarchie, établie sur une base quantitative, aide les préventionnistes à déterminer de manière objective les facteurs de risque ou causes possibles

d'accident à prioriser. La méthode exploite les fonctions de masse des indicateurs composant la règle. L'étude a montré que la probabilité des règles caractérisant les accidents mortels analysés est supérieure à celle des règles décrivant les accidents non mortels étudiés. Constat surprenant puisque, dans la réalité, les accidents non mortels (graves et non graves) sont plus fréquents que ceux mortels. Ce constat s'explique par le fait que les accidents analysés proviennent du Centre de documentation de la CNESST qui publie des rapports d'enquête uniquement d'accidents graves ou mortels. Puisque dans la thèse, les accidents avec la plus grande gravité du dommage (mortels) sont aussi les plus probables, il est suggéré que les préventionnistes des entreprises concernées par les accidents analysés entament le processus de réduction du risque en s'attaquant d'abord à la prévention de dommages mortels.

La probabilité du dommage calculée permettra d'avoir un référentiel de comparaison permettant de suivre l'évolution du risque. Par exemple, à la suite de la mise en œuvre d'un moyen de réduction du risque, il sera possible d'en évaluer l'impact sur la probabilité du dommage initialement calculée.

La démarche proposée est transposable à des équipements industriels autres que les convoyeurs à courroie. Elle peut être utilisée pour l'estimation de la probabilité d'occurrence d'un événement dangereux de nature diverse. Cette probabilité calculée pourra être intégrée à des outils qualitatifs, dans le but de préciser leurs niveaux de probabilité d'occurrence d'un événement dangereux. Cette intégration rendra le processus d'estimation du risque plus objectif.

Le succès de la démarche proposée repose sur la bonne volonté des intervenants à faire remonter l'information concernant les risques liés aux machines. Si aucun intervenant ne révèle d'information relative à un nouvel état d'un moyen de réduction du risque ou à un nouvel accident ou incident, l'information ne sera jamais enregistrée dans la base de données. Alors, les règles décrivant le risque ne seront jamais actualisées. Conséquemment, il en sera autant pour les facteurs de risques et les causes potentielles d'accidents, ainsi que les probabilités associées. Dans pareil contexte, des décisions dépassées risquent d'être prises pour réduire le risque.

Une culture de sécurité et une confiance mutuelle dans l'entreprise sont primordiales afin d'encourager la remontée d'information pour brosser un portrait plus juste du risque et améliorer l'efficacité des moyens de réduction du risque.

ABSTRACT

In machinery safety, safety practitioners use different sources as accident investigation reports to help them identify the risks in the workplace. In that case, they retrieve the knowledge from those reports one at a time, then may forget about them later.

Risk identification is followed by risk estimation in risk management. Safety practitioners in machinery safety generally use qualitative tools to estimate the risk. The qualitative aspect entails subjective decision-making regarding risk reduction measures. Moreover, the static nature of those tools forces its users to work with predetermined risk parameters. If new parameters are required to better describe the changing risk, those tools will be unable to consider them, which will lead to outdated decisions in risk reduction.

To overcome these issues, this thesis aims at suggesting a risk identification and risk estimation method that facilitates tracking of machinery-related risk in the workplace as well as their physical and organizational environment. That method exploits dynamic experience feedback (ExF) to make the most out of the reports in an efficient and sustainable way. Dynamic ExF is a process consisting of reporting information as well as inferring knowledge at the same time. The essential knowledge is extracted from the information contained in the reports after that information has been formalized in a database. That knowledge can be updated gradually as new information is reported. The knowledge is inferred in the form of relevant patterns generated by a data mining algorithm. A pattern is a combination of conditions describing accidents pertaining to a same set called “class”. Every condition is made of an indicator respecting a specific value or range of values. The indicator is a risk factor or a potential cause of accident. All in all, with a dynamic ExF using a database that can be updated on a regular basis, the reports will not go to waste after being read. Instead, they will continually contribute to the knowledge inference which will progress in the context.

A machine learning algorithm called Logical Analysis of Data (LAD) is integrated with the dynamic experience feedback process to ensure that the method is also suited for scarce data. Indeed, LAD proved to be efficient since the classification accuracy of the patterns generated from a 23-belt-conveyor-related accident database was adequate: between 72% and 74%. Two facts explain the choice of belt conveyors for the thesis:

- among all types of conveyors, they are the ones responsible of the biggest proportion of accidents (16.8%) between 1990 and 2011, according to 137 accident investigation reports from the *Commission des normes, de l'équité, de la santé et de la sécurité du travail* (CNESST) owing to conveyors;
- belt conveyors have the biggest ratio (8.5%) of serious and fatal accidents related to all kinds of machines, between 1999 and 2007, according to a database of the *Institut de recherche Robert-Sauvé en santé et en sécurité du travail* (IRSST).

The 23 accident investigation reports dealt with in this thesis come from the CNESST's Documentation Center. Analyzing the information in every report allowed for the identification of the elements describing the accidental context. Processing that information lead to a database that LAD used to generate two kinds of patterns:

- one for a version of the database splitting the 23 accidents into two classes: maintenance-related accidents and production-related ones;
- the other for a version of the 23-accident database comprising "Non fatal" and "Fatal" classes.

Some of the patterns generated showed that an accident can happen due to dangerous conditions (e.g. a poor environment in the workplace), but also because of an apparently-safe condition (e.g. an existing prevention program). In that case, one should investigate the unsafe sub-factors underlying to the apparently-safe condition in order to understand the occurrence of the accident. For example, 60% of the maintenance-related accidents happened despite the presence of a prevention program. Not updating that program could be a reason why the accident happened. Other patterns showed that risk factors or causes related to the equipment, the organization, the individual or the moment explain mainly the accidents analyzed.

Quantitative parameters related to the patterns, such as their coverage and their indicators frequency, enabled to start ranking the patterns as well as their indicators according to their importance (the coverage is the number of accidents a pattern characterizes). A probability of occurrence of harm estimation method associated with each pattern was developed to complete that hierarchy among the patterns with identical coverage. Such hierarchy with quantitative basis objectively guides the safety practitioner with the risks factors or accident potential causes

needing to be taken care of in priority. The probability of occurrence of harm estimation uses the mass functions related to the indicators included in the pattern. It is found that the patterns representing the “Fatal” class have a higher probability compared with the ones describing the “Non fatal” class. Surprising fact because in reality, non fatal accidents (serious and non serious ones) are more frequent than fatal accidents. Since the CNESST publishes accident investigation reports only regarding serious or fatal injuries, such difference is understandable. Nevertheless, considering the sample studied for the thesis, the most severe type of accident (fatal) is also the most likely. Therefore, it is suggested that the safety practitioners from the enterprises concerned by the accidents analyzed perform the risk reduction process preventing fatalities first.

The probability of occurrence of harm calculated has the potential to serve as a basis for comparison that enables to track the risk evolution. For instance, after implementing a risk reduction measure, one will be able to evaluate the effect of that measure on the probability of occurrence of harm previously calculated.

The method suggested is transposable to industrial equipment other than belt conveyors. The same approach can be adopted to estimate the probability of occurrence of a hazardous event of different nature. In such case, the probability calculated can be integrated to qualitative tools to specify their labels describing the probability of occurrence of a hazardous event. That integration adds objectivity to risk estimation process.

The success of that method relies on the good will of the stakeholders to bring feedback on the machinery-related risk portrait. If no stakeholder reveals information about a new state of a risk reduction measure or about a new circumstantial event, that information will never be registered in the database. Accordingly, the patterns defining the risk will never be updated, and so will not be the essential risk factors and accident potential causes, as well as the probabilities related. Consequently, outdated decision-making might be performed.

A safety culture as well as a mutual trust in the enterprises is important to encourage feedback in order to improve the risk portrait and the efficiency of the risk reduction measures.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

ALD	Analyse logique de données
CNAMTS	Caisse nationale de l'assurance maladie des travailleurs salariés (France)
CNESST	Commission des normes, de l'équité, de la santé et de la sécurité du travail (Québec)
CPSST	Centre patronal de santé et sécurité du travail du Québec
CSST	Commission de la santé et de la sécurité du travail (Québec)
DRP	Direction des risques professionnels (France)
EPI	Équipement de protection individuelle
ExF	*Experience feedback*
$G_{dommage}$	Gravité du dommage
GSC	*Greedy Set-Covering*
HAC	Classification agglomérante hiérarchique (*Hierarchical Agglomerative Clustering*)
HSE	*Health and Safety Executive* (Royaume-Uni)
HSL	*Health and Safety Laboratory* (Royaume-Uni)
IRSST	Institut de recherche Robert-Sauvé en santé et en sécurité du travail (Québec)
IS	*IteratedSampling*
ISO	Organisation internationale de normalisation (*International Organization for Standardization*)
JICOSH	*Japan International Center for Occupational Safety and Health* (Japon)
LAD	*Logical Analysis of Data*
MELITO	Moment, Équipement, Lieu, Individu, Tâche, Organisation
MILP	Programmation linéaire en nombres entiers 0-1 (*Mixed 0-1 Integer and Linear Programming*)
$\mathcal{P}$	Probabilité

$\mathcal{P}_{dommage}$	Probabilité du dommage
P_i^+	Règle positive (*positive pattern*) numérotée *i*
P_i^-	Règle négative (*negative pattern*) numérotée *i*
REX	Retour d'expérience
RSST	Règlement sur la santé et la sécurité du travail (Québec)
s.d.	Sans date
SST	Santé et sécurité du travail

LISTE DES ANNEXES

AVANT-PROPOS

Ce document est une thèse par articles. Les articles, au nombre de trois, se trouvent aux annexes A, B et C. La doctorante les a rédigés sous la supervision de son directeur de recherche et de son codirecteur de recherche. Les deux premiers articles ont été publiés dans des revues scientifiques. Le troisième a été soumis à une revue scientifique à des fins de publication. Contrairement à une thèse classique où le corps du document détaille la méthodologie et les résultats, le corps d'une thèse par articles résume l'ensemble du travail de la thèse, tout en expliquant le fil conducteur, les liens entre les articles et leurs contributions scientifiques.

Ce document comporte cinq chapitres. Le premier introduit le contexte et la problématique de recherche de la thèse. Le second synthétise la recension des écrits. Le troisième présente le travail réalisé, détaillé auparavant dans les trois articles. Le quatrième présente une discussion générale sur l'ensemble du travail. Le cinquième conclut la thèse tout en émettant quelques recommandations.

CHAPITRE 1 INTRODUCTION

Il existe trois grandes catégories de projets en recherche : 1) développement de connaissances, 2) développement technologique et 3) innovation (Langhame, communication personnelle, 16 mai 2013). Cette thèse par articles s'inscrit dans la première catégorie. Elle concerne le domaine de la sécurité du travail, plus particulièrement la sécurité des machines en milieu de travail. Dans ce document, le mot « machine » désigne tout « ensemble équipé ou destiné à être équipé d'un système d'entraînement, composé de pièces ou d'organes liés entre eux dont au moins un est mobile et qui sont réunis de façon solidaire en vue d'une application définie » (ISO, 2010).

1.1 Contexte et problématique de recherche

1.1.1 Les accidents du travail liés aux machines

Les machines peuvent être à l'origine d'accidents du travail engendrant des dommages réversibles (ex., ecchymose) ou irréversibles (ex., amputation), voire des décès. Comme toute lésion professionnelle, ces accidents occasionnent également des coûts. Ces derniers sont humains (ex., la valeur du changement de la qualité de la vie du travailleur accidenté), financiers directs (ex., frais médicaux) ou financiers indirects (ex., coûts de productivité perdue) (Lebeau et Duguay, 2011). La valeur subjective d'une vie humaine peut fluctuer de 0,306 à 7,8 M $ CA 2002 selon l'agence et le pays (Zhang et al., 2004 ; Lebeau et Duguay, 2011).

À titre d'exemple, de 2000 à 2007, les États-Unis comptaient 230 incapacités permanentes et décès liés à des machines (convoyeurs notamment) dans le secteur minier (Ruff et al., 2011). En 2010, le Japon comptait 12257 accidents du travail (décès inclus) liés à des convoyeurs et des machines de transport (JICOSH, s.d.). En 2014, la France dénombrait 15244 accidents, toutes machines et équipements fixes confondus (DRP / CNAMTS, 2015). Ces accidents ont occasionné 674744 journées de travail perdues. De ce nombre d'accidents, le pays comptait 996 incapacités permanentes et 1 décès. Toujours en 2014, l'Australie assistait à 10 décès par coincement d'une partie corporelle d'un travailleur dans une machine en mouvement (Safe Work Australia, 2015); chacune des victimes a dû s'absenter du travail au moins 4 jours. Au Canada, plus particulièrement dans la province du Québec, la Commission des normes, de l'équité, de la santé et de la sécurité du travail (CNESST) a recensé 3244 accidents du travail en 2014, engendrés par

des machines (CSST, 2015a). De ce total, 694 s'expliquent notamment par l'accès du travailleur à des pièces en mouvement (CSST, 2015a).

Les causes suivantes expliquent, en totalité ou en partie, les accidents du travail liés aux machines : l'accès à des pièces en mouvement, le manque ou le contournement de moyens de protection, l'absence de cadenassage lors des interventions de maintenance, le manque de formation ou d'expérience du travailleur, l'usage de méthodes de travail inadaptées, le manque de supervision du travailleur, les modifications erronées de la machine ou de son système de commande relatif à la sécurité, le manque d'appréciation du risque, la gestion déficiente de la santé et de la sécurité du travail (Gardner et al., 1999; Lindquist, 2011; Caputo et al., 2013; Chinniah, 2015).

Des résumés et rapports d'enquête d'accident du travail lié aux machines relatent des causes similaires. Ces documents sont consultables à partir de bases de données, telles que celle du :

- *United States Department of labor* aux États-Unis (https://www.osha.gov/pls/imis/AccidentSearch.search?);
- EPICEA en France (http://epicea.inrs.fr/servlet/public_request);
- Centre de documentation de la CNESST au Québec (https://www.centredoc.cnesst.gouv.qc.ca/in/fr).

Par exemple, dans la base de données du *United States Department of labor*, le résumé de l'accident « 200375871 -- Report ID: 0454510 » relate des fractures à la main gauche d'un travailleur. Cet événement est survenu sur une presse, notamment en raison du contournement de dispositifs de verrouillage et du démarrage d'un mécanisme par un tiers, à l'insu de la victime.

Dans EPICEA, le résumé d'accident correspondant au numéro de dossier 22346 explique l'écrasement du bras d'un travailleur entre la partie fixe d'une machine et un vitrage en mouvement. Le vitrage s'est déplacé intempestivement, via un mécanisme d'entraînement.

Dans la base de données du Centre de documentation de la CNESST, le rapport d'enquête EN003876 décrit l'accident mortel suivant :

> *[...] l'employeur actionne la levée de la benne basculante de sa camionnette afin de vider son contenu par gravité, mais cette dernière cesse de monter à environ 60 cm. Le*

> *travailleur qui l'accompagne s'introduit sous la benne pour ajouter de l'huile dans le réservoir du système de levage hydraulique. Le travailleur actionne accidentellement la commande de descente de la benne et il est alors écrasé entre le châssis de la camionnette et la structure de la benne.*

L'arbre des causes de la figure 1.1 résume le contexte accidentel décrit dans le rapport d'enquête de 44 pages de cet accident. Les principaux éléments suivants constituent ce rapport, à l'instar de la plupart des rapports d'enquête de la CNESST :

- un résumé de l'accident;
- une description de l'organisation du travail dans l'entreprise impliquée dans l'accident;
- une description du travail en cours au moment de l'accident;
- les faits et analyses de l'accident;
- une conclusion;
- des annexes explicatives.

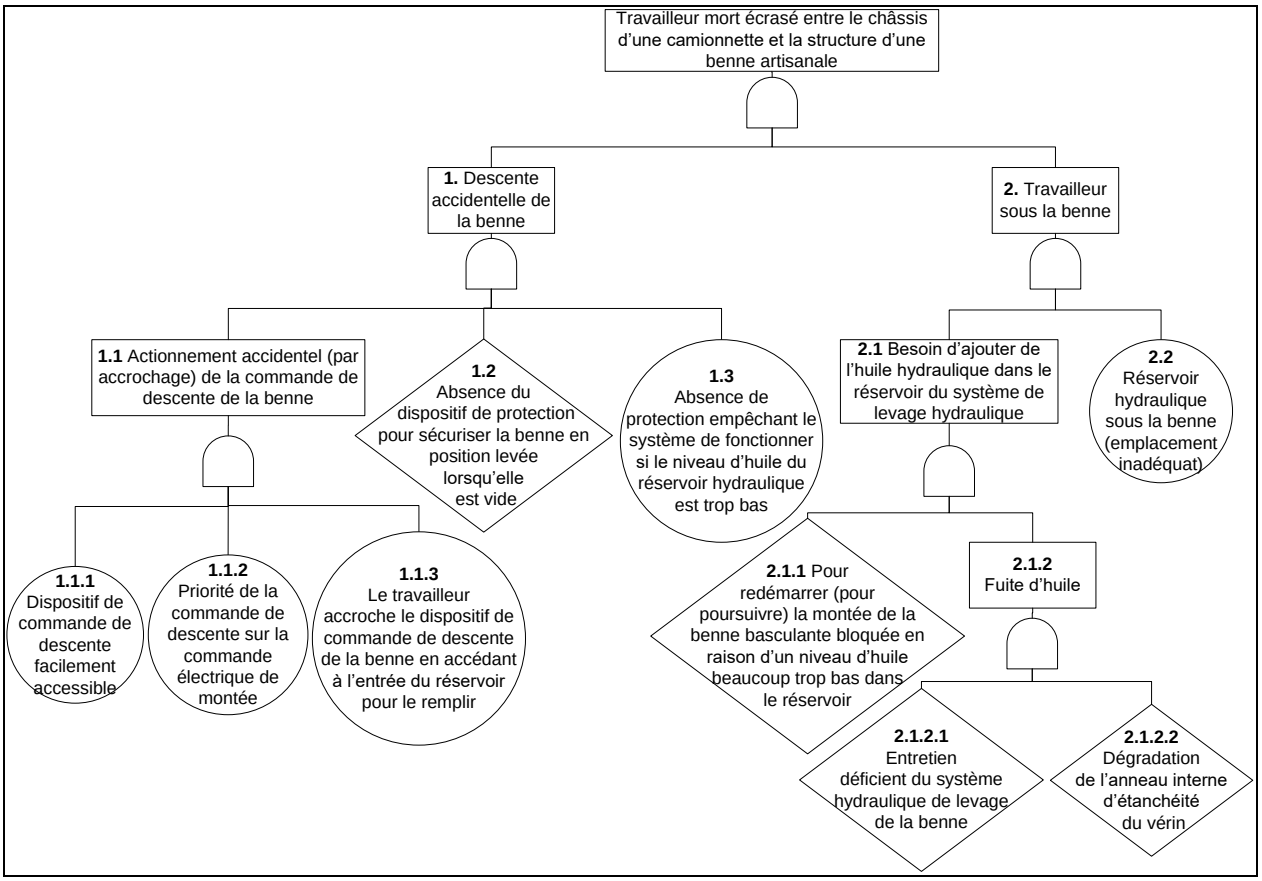


Figure 1.1 : Arbre des causes résumant le contexte accidentel décrit dans le rapport d'enquête EN003876 de la CNESST

La figure 1.1 relate des causes apparentées à celles identifiées dans la littérature :

- l'accès à une pièce en mouvement : descente de la benne sur le travailleur (cf. figure 1.1, cases 1 et 2);
- l'absence de moyens de protection (cf. figure 1.1, cases 1.2 et 1.3).

Cet arbre des causes montre également qu'une dégradation de l'équipement (figure 1.1, case 2.1.2.2) jumelée à un entretien déficient (figure 1.1, case 2.1.2.1) ont contribué à l'accident. Les causes expliquant l'actionnement accidentel de la commande de descente de la benne (figure 1.1, case 1.1) témoignent d'une conception inadéquate de la machine.

1.1.2 La sécurité des machines au Québec

En réponse aux 13500 accidents, liés aux machines, survenant en moyenne chaque année au Québec dont 20 en moyenne occasionnent des décès (Fontaine et al., 2006), la Commission de la santé et de la sécurité du travail (CSST)[1] lance en 2005 son plan d'action « Sécurité des machines ». Celui-ci fait de la prévention des accidents liés aux machines l'une de ses priorités (CSST, s.d.). Ce plan d'action stipule la tolérance zéro, particulièrement envers les pièces en mouvement accessibles sur des machines et susceptibles d'occasionner des lésions graves aux travailleurs (CSST, 2010a). Les efforts consentis par la Commission et ses partenaires ont permis, entre 2005 et 2008, de diminuer de 32% les accidents liés à des machines et de 27 % les accidents liés à l'accès à des pièces en mouvement des machines (CSST, 2010b). Plus récemment, en 2014, les accidents liés aux machines ont diminué de 7,4 % par rapport à 2013. En 2014, quatre décès sont survenus à cause d'une machine, comparativement à une moyenne de 10,4 depuis 2006 (CSST, 2015a). Cependant, aucun décès lié à des pièces en mouvements n'a été enregistré, comparativement à une moyenne de 3,1 depuis 2006 (CSST, 2015a). Il s'agit donc d'une baisse des accidents par rapport à la moyenne d'avant 2005.

Divers facteurs peuvent expliquer cette baisse. Par exemple, les efforts consentis par la CNESST dans le cadre de ce plan d'action : la sensibilisation par des publicités, les visites d'inspecteurs dans les milieux de travail, des moyens de pression comme l'apposition de scellés sur des machines et les amendes en cas d'infraction au Règlement sur la santé et la sécurité du travail (RSST) du Québec (Publications du Québec, s.d.).

1.1.3 Problématiques en gestion du risque lié aux machines

La prévention des accidents du travail passe par la gestion des risques qui menacent l'intégrité physique des travailleurs. La norme ISO 12100:2010 (ISO, 2010) en sécurité des machines définit le risque comme la combinaison de la probabilité du dommage et de la gravité de ce

[1] Depuis 2016, la CSST devient la CNESST : Commission des normes, de l'équité, de la santé et de la sécurité du travail. L'appellation CNESST sera préférée dans le présent document, excepté pour les références antérieures à 2016.

dommage : *Risque* = $f(\mathcal{P}_{dommage}, G_{dommage})$. Cette norme destinée aux concepteurs de machines fournit une démarche de gestion du risque en sécurité des machines. Il s'agit d'un processus itératif en deux étapes principales (cf. figure 1.2) : l'appréciation du risque suivie de la réduction du risque (ISO, 2010).

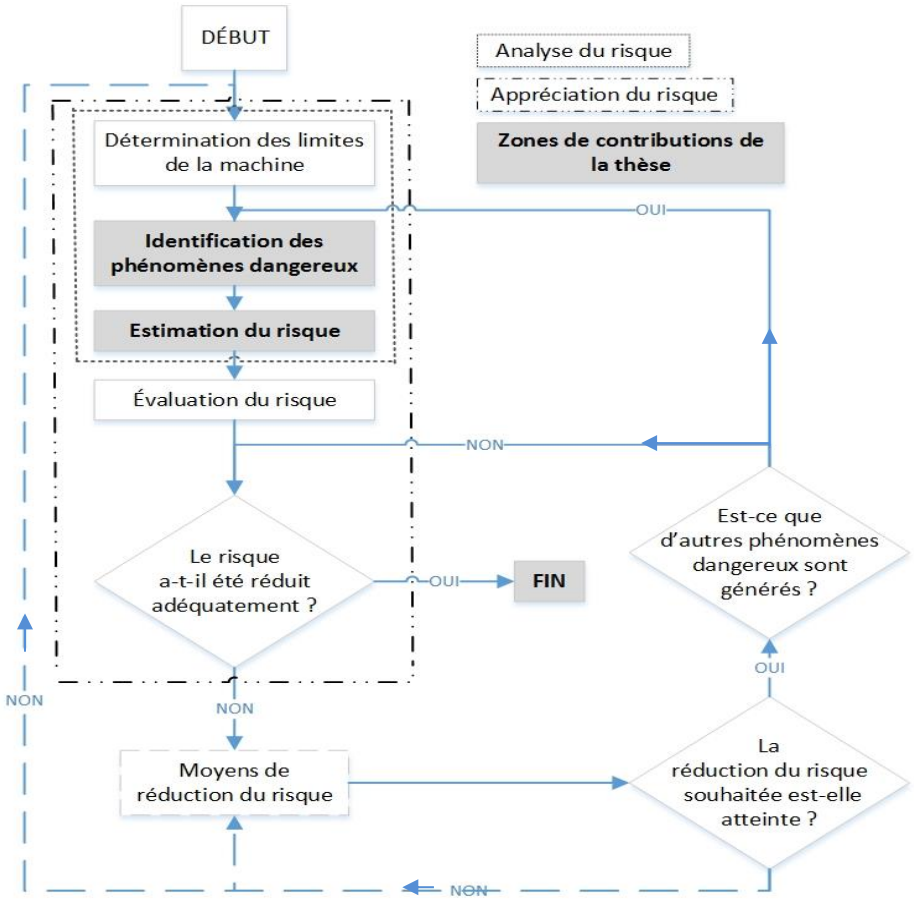


Figure 1.2 : Schéma simplifié du processus de gestion du risque en sécurité des machines (inspiré de l'ISO 12100:2010) et zones de contributions de la thèse

Réduire le risque de manière efficace nécessite une appréciation du risque adéquate et requiert de procéder par ordre de priorité. L'appréciation du risque est composée de l'analyse du risque puis de l'évaluation du risque. L'analyse du risque consiste à :

- déterminer les conditions d'utilisation de la machine ;
- identifier les risques associés menaçant la santé et la sécurité des utilisateurs (phénomènes dangereux, situations dangereuses, événements dangereux) ;

- estimer les risques en déterminant les échelles de cotation des paramètres d'estimation des risques.

L'évaluation du risque consiste à prioriser les actions pour le réduire. L'évaluation consiste à définir l'acceptabilité du risque. Elle est déterminée par consensus entre les différents intervenants concernés (ex. : concepteurs, employeurs, préventionnistes, travailleurs exposés aux risques).

Une fois le risque évalué, des moyens de réduction du risque sont choisis par consensus avec les parties prenantes. À cette étape, on vérifie si l'ajout des moyens de réduction du risque donne l'effet escompté. Ces moyens réduisent le risque en agissant complètement ou partiellement sur l'un des deux principaux paramètres définissant le risque ou les deux.

Initialement, les moyens de réduction du risque sont mis en œuvre par le concepteur de la machine. Ils sont répartis en trois groupes, du plus au moins efficace : 1) prévention intrinsèque (ex : élimination à la source du phénomène dangereux), 2) protection et mesures de prévention complémentaires (ex. : protecteurs, dispositifs de protection) et 3) information pour l'utilisation (ex. : signaux visuels, notice d'instructions). Le groupe 1 est donc à privilégier par rapport aux deux autres. Puis, le groupe 2 est à privilégier vis-à-vis du troisième. Ensuite, la machine est vendue à l'exploitant, mais avec un risque résiduel qui dépendra principalement des conditions réelles d'utilisation de la machine. L'exploitant, le préventionniste en l'occurrence, devra réduire le risque résiduel en adoptant divers moyens, tels que : des mesures organisationnelles (méthodes de travail sécuritaires, surveillance, système de permis de travail), des moyens de protection supplémentaires, le port d'équipements de protection individuelle (EPI), des formations. Dans le cas de l'utilisateur, aucune hiérarchie d'efficacité des moyens de réduction du risque n'est établie par l'ISO 12100:2010.

Les moyens de protection supplémentaires concernent, par exemple, le cas où le préventionniste confie à un intégrateur le soin de modifier la machine, pour qu'elle réponde à des exigences spécifiques de fabrication non prévues par le concepteur d'origine. Ces modifications sont d'autant plus difficiles pour l'intégrateur quand la machine en question n'a pas de norme de type C associée (norme spécifique à un type de machine). En s'aidant de normes plus générales comme l'ISO 12100:2010, d'outils d'identification ou d'estimation du risque, il adapte la machine, au meilleur de ses connaissances, pour satisfaire les critères de fabrication tout en

assurant la sécurité des travailleurs. C'est dans ce genre de contexte que le risque estimé peut changer. En effet, les moyens de protection d'origine risquent de ne plus convenir à la nouvelle utilisation de la machine, d'où la nécessité de mettre à jour le risque estimé et installer des moyens de réduction du risque adaptés. Dans le cas où la démarche d'intégration, les mesures organisationnelles, les EPIs ou les formations sont lacunaires, ces défaillances peuvent conduire à des accidents. Pour prévenir les accidents liés aux machines, plusieurs préventionnistes utilisent la démarche de l'ISO 12100:2010, quoique dédiée à des concepteurs de machines. Ces préventionnistes éprouvent des difficultés notamment à l'étape d'identification des phénomènes dangereux et situations dangereuses, ainsi qu'à celle de l'estimation du risque.

1.2 Organisation de la thèse

Le chapitre 2 présente une recension des écrits au sujet de l'identification des phénomènes dangereux et situations dangereuses, puis de l'estimation du risque. Les avenues de recherche qui pourraient en découler sont identifiées. De celles-ci, des questions de recherche sont soulevées, ainsi que les hypothèses de recherche capables d'y répondre. Enfin, les objectifs de recherche sont également listés. Le chapitre 3 synthétise, aux sections 3.1 et 3.2, l'ensemble du travail effectué. Les sections 3.3 à 3.5 résument chacune le travail associé à l'un des trois articles scientifiques de la thèse (cf. tableau 1.1). Chacun des articles scientifiques fait l'objet d'une annexe (cf. annexes A, B et C). Les deux annexes subséquentes apportent des précisions sur des résultats obtenus au courant de la recherche. Le chapitre 4 discute de l'ensemble du travail réalisé tout en précisant les limites et contraintes de la recherche réalisée. De plus, de nouvelles voies de recherche y sont identifiées. Enfin, le chapitre 5 conclut la thèse en plus d'émettre quelques recommandations.

Tableau 1.1 : Sections du chapitre 3 traitant des trois articles de thèse

Titre de l'article scientifique, son statut et son annexe	**Section du chapitre 3 correspondante**
Article 1 : *Contribution of Dynamic Experience Feedback to the Quantitative Estimation of Risks for Preventing Accidents: A Proposed Methodology for Machinery Safety* (Publié, annexe A)	3.3 Proposition d'une aide à la décision en sécurité des machines
Article 2 : *Application of Logical Analysis of Data to Machinery-Related Accident Prevention Based on Scarce Data* (Publié, annexe B)	3.4 Vérification de l'applicabilité de l'ALD à un échantillon très restreint aux caractéristiques non évidentes
Article 3 : *Estimation of probability of harm in safety of machinery using an investigation systemic approach and Logical Analysis of Data* (Soumis, annexe C)	3.5 Estimation de la probabilité du dommage

CHAPITRE 2 REVUE CRITIQUE DE LA LITTÉRATURE PERTINENTE

Une veille scientifique a été entreprise tout au long de l'étude. À cette fin, le moteur de recherche *Google*, la section « Publications et outils » du site Web de l'IRSST et les bases de données *Compendex* et *Inspec* ont été interrogés. Les mots clés employés se rattachaient à l'identification des phénomènes dangereux et l'estimation du risque. D'abord, la recherche par mots clés portait sur le domaine de la sécurité des machines pour documenter l'état de la recherche sur le sujet. Les mots clés suivants ont été utilisés seuls ou de manière combinée : sécurité des machines, outils d'estimation du risque, *safety of machinery, machinery safety, machine*, equipment, occupational safety, workplace, hazard identification, risk estimation, risk estimation tool, risk assessment, risk management.*

Ensuite, d'autres domaines ont été explorés afin d'en exploiter les moyens de gestion du risque transposables et utiles à la sécurité des machines : la SST en général, l'aérospatial, le biomédical, les finances, le transport, le pétrochimique, le nucléaire. La recherche de sources bibliographiques s'est réalisée en partant des mots clés généraux : *hazard, management, hazard identification, risk management, risk assessment, risk estimation*. Jusqu'ici, les résultats de la recherche bibliographique ont permis de constater qu'une gestion du risque responsable requiert une mise à jour du risque estimé. L'estimation était souvent quantitative et à partir de calculs probabilistes. La mise à jour s'effectuait souvent au moyen d'un retour d'expérience « intelligent », c'est-à-dire inférant des connaissances à partir d'un algorithme de fouille de données (la fouille de données consiste à explorer des données dans le but d'y trouver des caractéristiques non évidentes; c'est le moteur du processus d'extraction de connaissances). Ces constatations ont suggéré l'affinement de la recherche en rentrant les mots clés suivants : probabilité*, mise à jour du risque, actualisation du risque, retour d'expérience, fouille de données, *quantitative risk assessment, probabilit*, risk update, risk monitoring, experience feedback, data mining, machine learning*. Parmi les sources bibliographiques traitant de fouille de données, l'algorithme ALD (analyse logique de données) a retenu l'attention, car il semblait adapté pour l'étude comme expliqué plus loin. Ainsi, les mots clés suivants ont été ajoutés pour la recherche bibliographique : *Logical Analysis of Data, LAD*.

La période fixée pour la recherche par mots clés a couvert les années 1988 à 2017. Au moins 162 sources ont été consultées. De ce nombre, seuls des documents publiés de 1999 à 2017 ont servi à la recension des écrits. Il s'agit de : 51 articles scientifiques, 18 rapports d'études, 7 guides techniques, 7 livres, 2 thèses, 2 règlements, 1 norme et 7 autres sources provenant essentiellement de sites Web. Les sections 2.1 et 2.2 suivantes synthétisent cette recension répartie entre l'intégralité de l'article 1 (annexe A) et les introductions des articles 2 et 3 (annexes B et C).

2.1 Les moyens d'identification du risque et pistes de recherche

À partir d'entrevues dans des entreprises en plasturgie, Chinniah et al. (2014) relatent que la formation, dont celle en santé et en sécurité du travail, rend le personnel moins vulnérable face aux risques menaçant son intégrité physique. En effet, le personnel devient averti face aux risques de dommage qui le guettent. La formation constitue donc un moyen pouvant aider les préventionnistes à identifier les risques associés aux machines. Par exemple, donner aux préventionnistes une formation sur l'analyse du risque passe par l'explication des notions de phénomène dangereux (c.-à-d., source du dommage) et de situation dangereuse (c.-à-d., l'humain dans une zone dangereuse). L'explication de ces notions peut être illustrée par des images fournissant divers types de phénomènes dangereux et de situations dangereuses. Ainsi, les préventionnistes seront entraînés à les identifier.

En outre, pour identifier les risques, les préventionnistes utilisent des tutoriels (ex., Mecaprev https://machines-sures.inrs.fr/mecaprev), des guides (ex., Lupin et Marsot, 2006), des rapports techniques (ex., Worsell et Ioannides, 2000), des normes (ex., ISO, 2010), des règlements (ex., European Machinery Directive 2006/42/EC, 2006) ou apprennent en consultant des rapports d'enquête d'accident. À part ces rapports, les autres sources mentionnées précédemment proposent des exemples de phénomènes dangereux et situations dangereuses illustrées. Toutefois, ils présentent des causes directes d'accident, c'est-à-dire essentiellement techniques. L'interaction de ces causes avec des causes indirectes (c.-à-d. organisationnelles ou humaines) donnerait une représentation plus réaliste du risque à évaluer.

À cette fin, la description des accidents dans les rapports d'enquête que lisent les préventionnistes leur offre des exemples d'interactions entre ces deux types de causes. Cependant, l'apprentissage à partir de rapports d'accidents se base sur un retour d'expérience (REX) statique, car le préventionniste tire de l'information limitée au cas d'accident enquêté dans le rapport. De plus, la longueur des rapports d'enquête d'accident en général et le langage difficilement accessible employé dans ces documents contribuent considérablement à leur échec de diffusion (Lindberg et al., 2010). Comme l'explique Lindberg et al. (2010), divers auteurs (Johnson, 2002; Amoore et Ingram, 2002; Johnson et Holloway, 2003; Lindberg et Hansson, 2006) insistent sur l'importance d'une diffusion efficace des rapports d'enquête d'accident. Kletz (1993) déplore le fait que, souvent, de bons rapports diffusés massivement sont lus, archivés, puis envoyés aux oubliettes. Il est donc nécessaire de rendre plus accessibles les connaissances véhiculées dans les rapports d'accidents et leur diffusion plus efficace. En outre, Lindberg et al. (2010) conseillent que l'essentiel de l'information contenue dans les rapports d'accidents soit acheminé à tout intervenant ayant le pouvoir de l'utiliser pour prévenir les accidents. D'après Beler (2008), un apprentissage plus efficace serait d'inférer des connaissances, sous forme de règles, à partir d'un ensemble d'événements passés. La description de ces événements proviendrait de rapports et serait structurée dans une base de données. L'ensemble des règles pertinentes (générées par un logiciel de fouille de données) constitueraient l'essentiel des connaissances tirées des rapports. La pertinence des règles serait dictée par leur précision de classification. L'alimentation de la base de données par la remontée d'information contenue dans les rapports et l'inférence des règles pertinentes constituent ce qu'on appelle ici, le retour d'expérience dynamique (REX dynamique). Idéalement, l'apprentissage devrait se baser à la fois sur des événements négatifs que positifs. Comme le mentionnent Van Wassenhove et Garbolino (2008), « le retour d'expérience ne se restreint pas à l'analyse des causes de dysfonctionnement, mais peut également servir pour l'analyse des activités pour en déduire les modes de bon fonctionnement (REX « positif ») ». Dans le cadre de la thèse, un événement positif est un événement sans conséquence néfaste pour les personnes comme un presque accident. Ce dernier est un passé-proche ou un incident lors duquel, des conditions rencontrées ont permis d'éviter un dommage corporel. Un événement négatif est un événement redouté, tel qu'un accident. Ce dernier est un événement imprévu et soudain occasionnant un dommage corporel.

La gestion du risque en exploitant un historique de données (ex., d'accidents) est encouragée par la norme en sécurité des machines ISO 12100:2010. Des études (ex., Chinniah, 2015; Malm et al., 2010) proposent ou exploitent des statistiques d'accidents du travail liés aux machines pour aider à identifier les causes possibles d'accident, dans l'optique de gérer le risque de dommage associé. Cependant, il s'agit de statistiques référant à des causes isolées. Connaître leur interaction aurait été bénéfique pour brosser un meilleur portrait du risque. Des études générales en SST (Verma et al., 2014; Cheng et al., 2012; Silva et al., 2012; Rivas et al., 2011) exploitent de telles données, mais tous risques confondus (dommages matériels et physiques). Leurs exploitations empruntent une avenue « intelligente ». Elles infèrent des connaissances sur la source du risque en apprenant d'accidents, grâce à des techniques de fouille de données telles que : les arbres de décision, les règles d'association, les réseaux bayésiens, les machines à vecteurs de support (*Support Vector Machine*). Leur exploitation se base sur des centaines de données. Comme les accidents sont des événements rares, un préventionniste d'entreprises devrait pouvoir exploiter le peu d'accidents de son secteur ou de son lieu de travail pour faire profiter aux travailleurs les connaissances inférées. Malgré la rareté des accidents, l'usage d'un algorithme de fouille de données est pertinent pour inférer des connaissances puisque les caractéristiques d'un ensemble d'accidents sont difficiles à concevoir si un nombre considérable de variables définissent le contexte accidentel. Une méthode d'identification du risque pouvant supporter des données rares, tout en montrant l'interaction entre les principaux facteurs de risque ou causes directes et indirectes d'accidents du travail est donc requise. De plus, une telle méthode destinée aux machines uniquement, contrairement aux dernières sources précitées, serait profitable aux préventionnistes responsables de la sécurité des machines.

Réputé pour le diagnostic et le pronostic de maladies, l'algorithme de fouille de données ALD exploité pour l'inférence de connaissances s'est avéré performant pour identifier et s'informer sur les causes de maladies (Alexe et al., 2006; Alexe et al., 2003; Lauer et al., 2002). Il s'agit d'un algorithme exploitant un processus combinatoire optimisé et basé sur la logique booléenne. L'ALD relève du domaine de l'intelligence artificielle comme on peut le constater dans Ragab et al. (2013). Il sert à la classification d'observations selon deux classes ou plus (multiclasse) (Kim et Choi, 2015a, 2015b). Ces classes peuvent représenter des maladies par exemple. L'ALD génère des connaissances à partir d'une base de données, à condition qu'il y détecte une combinaison de causes distinguant deux types d'états, par exemple : « malade » versus « non

malade » (donc, en santé). Cela étant l'unique condition, il devrait être en mesure de générer des connaissances pour des données rares, contrairement aux autres techniques de fouille de données qui infèrent des règles satisfaisantes à condition d'avoir énormément de données (ex., cas des réseaux de neurones) ou des ensembles fréquents dans les données (ex., cas des arbres de décisions et des règles d'association).

2.2 Les moyens d'estimation du risque et avenues de recherche

Pour estimer le risque, les préventionnistes utilisent couramment des outils matriciels qualitatifs (Paques et al., 2007). Cela s'explique par la facilité d'intégrer les résultats de l'estimation du risque à des politiques de gestion du risque. Toutefois, les matrices d'estimation du risque présentent certaines lacunes. Par exemple, leurs niveaux qualitatifs de risque peuvent surestimer les niveaux quantitatifs de risque correspondants (Cox, 2008). Cela contribue à une prise de décision erronée quant au choix des moyens de réduction du risque et à l'ordre dans lequel ils seront mis à exécution.

En sécurité des machines, Chinniah et al. (2011) se sont penchés sur la subjectivité des outils qualitatifs d'estimation du risque en proposant des règles de construction de matrices ou d'appellation des paramètres d'estimation du risque. Cependant, l'article de Villa et al. (2016) rattaché à l'industrie pétrochimique confirme que l'estimation quantitative du risque est le moyen disponible le plus approprié pour objectiver l'appréciation du risque. Duijm (2015) affirme également que les outils quantitatifs d'estimation du risque sont plus efficaces que ceux qualitatifs. D'après Cox (2008) et Duijm (2015), une gestion efficace du risque passe par une estimation quantitative plutôt que qualitative du risque. L'estimation quantitative consiste essentiellement en l'estimation de la probabilité du dommage. Or, en sécurité des machines, Gauthier et al. (2016) avancent que l'estimation de la probabilité « est un aspect problématique de l'estimation du risque et qu'une attention particulière doit y être accordée ».

Par ailleurs, les outils matriciels qualitatifs offrent une estimation du risque basée sur des paramètres prédéterminés dont le nom et le nombre demeurent immuables. Advenant une évolution du risque nécessitant l'ajout d'un nouveau paramètre pour le décrire, ces outils ne pourront l'intégrer. Prenons par exemple, l'outil d'estimation du risque du tableau 2.1.

Tableau 2.1 : Matrice de risque tirée du rapport technique ANSI B11.TR3 (2000)

Probability of Occurrence of Harm	**Severity of Harm**			
	Catastrophic	**Serious**	**Moderate**	**Minor**
Very Likely	**High**	**High**	**High**	**Medium**
Likely	**High**	**High**	**Medium**	**Low**
Unlikely	**Medium**	**Medium**	**Low**	Negligible
Remote	**Low**	**Low**	Negligible	Negligible

Cet outil décrit globalement le risque en fonction deux paramètres qualitatifs: 1) la gravité du dommage, 2) la probabilité d'occurrence du dommage. Si, dans le milieu de travail, une machine a été modifiée, il est possible que ce changement influence le risque d'accident positivement ou négativement. Dans ce cas, le préventionniste peut souhaiter en évaluer l'impact sur la gravité ou la probabilité d'occurrence du dommage. Il est impossible d'introduire un tel paramètre à même l'outil. Le préventionniste doit donc s'en bâtir un nouveau, selon ses besoins.

Plusieurs raisons peuvent expliquer une mutation du risque, parmi lesquelles : la dégradation de la machine en raison de contraintes environnementales ou de son usage, les modifications apportées à la machine pour répondre à de nouvelles exigences de production. En raison de l'évolution du risque, les moyens de réduction du risque adoptés au temps présent risquent de devenir désuets dans un futur proche ou éloigné. Une méthode favorisant l'itération du processus de gestion du risque afin d'actualiser celui-ci est donc nécessaire.

En sécurité des machines, Aneziris et al. (2013) utilisent l'estimation quantitative du risque pour prévenir les accidents associés à des pièces en mouvement. À partir d'un arbre de défaillances jumelé à un arbre d'événements, ils estiment la probabilité du dommage dû à un contact avec des pièces en mouvement de machines. Bien que leur méthode montre l'interaction entre diverses causes en se basant sur des centaines d'accidents, ces causes et leurs interactions ne peuvent être mises automatiquement à jour une fois les données d'un nouvel accident rentré. Alors, il y a un besoin d'une méthode flexible, c'est-à-dire capable d'intégrer de nouveaux indicateurs d'accidents et inférer automatiquement l'interaction entre les principales causes qui influenceront la probabilité du dommage.

Ainsi, cette thèse contribue à améliorer les étapes d'identification et d'estimation du risque en sécurité des machines. Elle apporte une valeur ajoutée aux outils disponibles dans ce domaine pour identifier et estimer le risque, comme le montreront les résultats rattachés à la thèse. Cette dernière montre que la gestion du risque ne s'arrête pas aussitôt que le risque est adéquatement réduit. Contrairement à la case « FIN » de la figure 1.2, la gestion du risque ne doit pas avoir de fin, elle doit être en mode « veille » constamment. C'est une boucle perpétuelle qui se renouvelle à chaque remontée d'informations pour générer des connaissances actualisées au sujet des risques.

2.3 Questions et hypothèses de recherche

La problématique vécue par les préventionnistes (cf. section 1.1.3) et la recension des écrits amènent les questions de recherche suivantes :

Comment aider les préventionnistes à :

- identifier et estimer efficacement les risques liés aux machines ?
- prioriser les mesures de réduction du risque ?
- suivre l'évolution des risques liés aux machines ?

Basées sur la revue critique de la littérature, les deux hypothèses de recherche suivantes sont jugées appropriées pour répondre à ces questions :

1) Utiliser le retour d'expérience (REX) dynamique basé sur l'analyse logique de données (ALD) permet d'identifier efficacement* les facteurs de risque liés aux machines et d'en suivre l'évolution.
2) Estimer la probabilité du dommage à partir des facteurs de risque identifiés permet d'estimer efficacement** les risques et prioriser les mesures de réduction du risque.

*Par « identifier efficacement les risques » on entend apprendre du passé pour prédire les types d'accidents possibles et noter leurs principales causes et facteurs de risque cibles.

**Par « estimer efficacement les risques » on entend réduire la subjectivité de l'estimation du risque pour rendre la hiérarchisation des risques plus juste et la prise de décision en matière de réduction du risque plus adaptée.

2.4 But de l'étude et objectifs de recherche

Ainsi, le but de l'étude est de proposer une démarche efficace d'identification et d'estimation des risques facilitant leur suivi en milieu de travail. Les risques considérés ici sont ceux liés aux machines et à leurs environnements physique et organisationnel.

Les objectifs de recherche suivants soutiennent ce but :

1) Effectuer une veille scientifique rattachée aux cinq thèmes principaux et secondaires de l'étude :
    - Thèmes principaux : « Identification du risque » et « Estimation du risque »;
    - Thèmes secondaires : « Retour d'expérience », « Fouille de données », notamment « Analyse logique de données (ALD) » et « Actualisation du risque ».
2) Construire le modèle conceptuel d'identification du risque jetant les bases pour son actualisation.
3) Valider le modèle conceptuel.
4) Construire le modèle quantitatif d'estimation du risque.
5) Valider le modèle quantitatif.

CHAPITRE 3 SYNTHÈSE DE L'ENSEMBLE DU TRAVAIL

Une vue **d'ensemble** de l'apport scientifique des trois articles rattachés à la thèse est présentée à la section 3.1. Cette dernière montre la cohésion entre ces articles, également abordée à la figure 3.1 et dans les sections 3.3 à 3.5. La section 3.2 résume la méthodologie encadrant les travaux inhérents aux trois articles scientifiques, donc les travaux de la thèse elle-même. Les sections 3.3 à 3.5 mettent aussi en exergue les résultats qui ont déterminé le cours de l'étude et permis de répondre aux questions de recherche. Elles mentionnent également les ajustements qui ont dû être apportés au cours de la recherche.

La section 3.3 présente une méthode de conception d'un outil qui aidera les préventionnistes dans leur prise de décision en matière de gestion du risque, c'est-à-dire en matière d'identification des situations dangereuses, d'estimation du risque et d'efficacité de la réduction du risque. La section 3.4 traite des travaux ayant abouti à l'article 2. La section 3.5 relate les travaux liés à l'article 3.

3.1 Apport scientifique : l'ALD appliquée à une base de données restreinte pour estimer quantitativement le risque d'accident

Les articles présentés aux annexes A, B et C constituent les contributions scientifiques de la thèse.

L'article 1 fait d'abord une recension des écrits. Il brosse le portrait d'outils et méthodes de gestion du risque en sécurité des machines, mais aussi dans d'autres domaines. Les avantages et inconvénients des méthodes sont soulignés. De là jaillissent les contributions possibles de méthodes d'autres domaines à la sécurité des machines. Les méthodes choisies sont le REX dynamique et l'estimation quantitative du risque (cf. sections 2.1 et 2.2 pour la définition de ces méthodes). Dans cette thèse, le REX dynamique se base sur l'algorithme de fouille de données ALD. Ensuite, la réflexion associée à la contribution des méthodes sélectionnées permet de proposer une méthode proactive d'aide à la décision en gestion du risque lié aux machines. Comme mentionné dans l'article 1 et à l'instar de ce que pensent Villa et al. (2016), l'aspect dynamique de la méthode permet de mettre à jour les risques (c.-à-d., les facteurs de risque et causes possibles d'accidents, ainsi que les probabilités de scénarios accidentels) lors de

changements apportés aux installations ou aux systèmes et ainsi éviter des prises de décision erronées. La méthode présentée est destinée à la réalisation d'un outil informatique. Elle se base sur des accidents et presque accidents du travail liés aux machines. L'utilité d'une telle méthode est appuyée par deux cas d'accidents survenus sur un même convoyeur à courroie.

En raison de l'inaccessibilité à des données de presque accidents, l'ALD a été utilisé pour des données d'accidents uniquement. L'échantillon d'accidents employé pour le REX dynamique était très restreint : 23 accidents du travail liés à des convoyeurs à courroie. Ce nombre a été retenu puisqu'il représente les uniques rapports d'enquête d'accident accessibles depuis le site Web du Centre de documentation de la CNESST et associés à des convoyeurs à courroie. Ce type de convoyeur a été sélectionné pour les raisons suivantes :

- les convoyeurs à courroie ont provoqué le plus d'accidents (16,8%) entre 1990 et 2011, d'après 137 rapports d'enquête d'accident de la CNESST liés à des convoyeurs (le plus récent rapport au moment de créer la base de données datait de 2011);
- ces convoyeurs représentaient la plus grande proportion (8,5 %) d'accidents selon une base de données d'accidents graves et mortels de l'Institut de recherche Robert-Sauvé en santé et en sécurité du travail (IRSST). Cette base de données couvrait 164 accidents allant de 1999 à 2007, rattachés à des machines fixes.

L'ALD a exploité les 23 accidents pour générer des règles dans deux cas d'études (cf. tableau 3.1). Le premier fait l'objet du deuxième article de thèse, tandis que le troisième article de thèse traite du second cas d'étude.

Tableau 3.1 : Cas d'études exploitant l'ALD et l'article de thèse y correspondant

	Répartition des 23 accidents
Cas d'étude n° 1 (article 2)	• 20 accidents de maintenance • 3 accidents de production
Cas d'étude n° 2 (article 3)	• 19 accidents mortels • 4 accidents non mortels

En raison de cet échantillon restreint d'accidents, il fallait tester les aptitudes et performances de l'ALD avant d'aller plus loin dans l'étude. En effet, les études exploitant l'ALD utilisaient,

jusque-là, des bases de données contenant au moins une centaine d'observations (dans le cadre de cette thèse, une observation est un accident du travail). Ainsi, l'article 2 a permis de vérifier l'applicabilité de l'ALD à un échantillon très restreint d'observations aux caractéristiques non évidentes, en raison du nombre important d'indicateurs (c.-à-d. de variables) : 23 indicateurs. Un indicateur est une variable dont la valeur contribue à décrire l'accident du travail. Dans l'article 3, l'algorithme ALD a pu générer des règles (c.-à-d. connaissances) avec une prédiction satisfaisante allant de 72% à 74%. Les règles caractérisent deux classes d'accidents : 1) accidents de maintenance et 2) accidents de production. Afin d'asseoir sa contribution scientifique, l'article 2 présente les avantages et limites des études appliquant la fouille de données à la santé et la sécurité du travail (SST). Aussi, l'article définit et explique l'algorithme ALD à travers un exemple détaillé. Appliquer l'ALD à la sécurité des machines comme à la SST est en soi une nouveauté. L'emploi de l'ALD va au-delà de la génération de règles à partir de causes d'accidents. Il est utilisé pour de l'estimation du risque lié à des convoyeurs à courroie, afin de réduire le risque par ordre de priorité. L'algorithme a analysé des facteurs de risques, tout comme des faits (qu'ils soient sécuritaires ou non), au même titre que les causes d'accidents. L'optique était d'avoir une base de données décrivant le contexte accidentel plutôt que les causes d'accident uniquement.

L'applicabilité de l'ALD étant confirmée, des bases solides existent pour générer des règles à partir d'un échantillon très restreint. L'article 3 emploie l'ALD à un nouveau format de la base de données de l'article 2. Le choix de cette nouvelle configuration visait uniquement à permettre aux préventionnistes, notamment les enquêteurs d'accidents du Québec, de se retrouver dans un univers qui leur est familier. C'est le concept « MELITO » : Moment, Équipement, Lieu, Individu, Tâche, Organisation (CPSST, 2004). Ces six éléments circonscrivent le système accidentel. Ce concept est couramment utilisé lors des enquêtes et analyses d'accidents du travail au Québec, pour saisir la globalité du contexte accidentel. Le formatage des données en fonction du concept MELITO est transposable à tout concept similaire d'enquête d'accident. L'article 3 exploite les règles générées par l'ALD en estimant la probabilité du dommage à travers le calcul de la probabilité de chaque situation accidentelle identifiée par une règle générée. Ces règles caractérisent deux classes d'accidents ou de dommages : 1) accidents mortels et 2) accidents non mortels. La probabilité du dommage sert de référentiel de comparaison pour suivre l'évolution du risque après la survenue d'un nouvel accident, mais aussi pour évaluer l'impact d'une mesure de

réduction du risque. La probabilité du dommage associé à une règle permet également d'en déterminer l'importance par rapport aux autres règles. La hiérarchisation des règles permet d'estimer le risque et d'établir un ordre de priorité dans le processus de réduction du risque.

3.2 Méthodologie globale de la thèse

La figure 3.1 présente la méthodologie de recherche. Après les étapes A) et B) liées par la recension des écrits, la méthodologie illustrée présente les éléments des travaux ayant permis de tester les deux hypothèses de recherche citées en 2.3. Le côté gauche de la figure 3.1 indique les parties des travaux permettant de répondre aux cinq objectifs de recherche annoncés en 2.4. Le côté droit indique le principal apport scientifique de chacun des articles de la thèse. Le premier article porte sur la recension des écrits et propose une méthode de conception d'un outil dynamique d'aide à la décision pour gérer les risques associés aux machines. Le deuxième article contribue principalement à tester la première hypothèse de recherche. Le troisième article contribue essentiellement à tester la deuxième hypothèse de recherche.

Des détails sur la méthodologie et les résultats clés qui ont contribué à l'avancement et l'aboutissement de la recherche sont présentés aux sections 3.3 à 3.5.

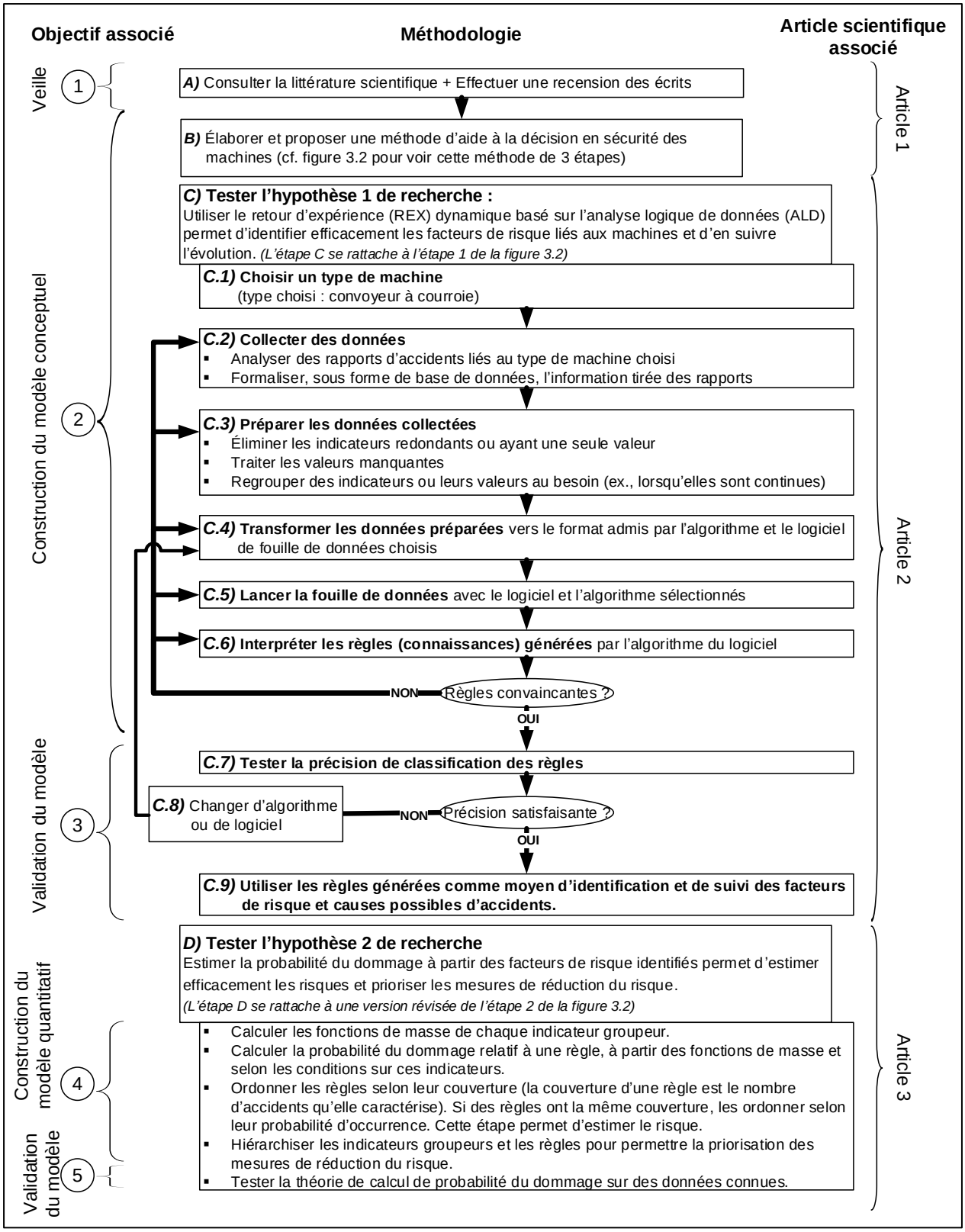


Figure 3.1 : Correspondance entre la méthodologie, les cinq objectifs de recherche et les trois articles scientifiques de la thèse

3.3 Proposition d'une aide à la décision en sécurité des machines

La littérature consultée a permis d'élaborer la méthode de conception d'un outil dynamique d'aide à la décision en sécurité des machines (cf. figure 3.2). La méthode de conception comporte trois étapes composées chacune d'une entrée, d'un traitement et d'une sortie :

- La première étape permet d'identifier les situations dangereuses liées aux machines à partir d'une base de données d'accidents issus de rapports d'enquête. L'algorithme ALD parcourt la base de données pour générer des connaissances sous forme de règles. Chaque règle générée à cette étape représente une situation dangereuse et est une combinaison de causes ou de facteurs de risques. Alors, la règle montre l'interaction entre ceux-ci. Ainsi, à l'étape 1, on connaît les principaux facteurs de risque et causes d'accidents analysés, mais pas leur impact sur le risque.
- C'est pourquoi l'étape 2 cherche à déterminer cet impact. Pour y arriver, il est proposé de collecter les avis *a priori* d'experts au sujet de cet impact. En y jumelant, avec l'inférence bayésienne, les données tirées des rapports, on peut estimer la probabilité des règles et ainsi estimer quantitativement le risque. Les experts visés ici sont des inspecteurs de la CNESST, des préventionnistes d'entreprises ou d'organismes en prévention. En somme, la deuxième étape permet d'estimer quantitativement le risque en calculant la probabilité du dommage associé aux situations dangereuses identifiées.
- La troisième étape permet d'évaluer l'efficacité des moyens de réduction du risque adoptés en mesurant leur effet sur la probabilité du dommage. Cette étape estime également la possibilité qu'un inspecteur de la CNESST appose un scellé sur la machine.

De plus amples renseignements sur la figure 3.2 sont disponibles à la section 5 de l'article 1.

Notez que la figure 3.2 présente un REX idéal, c'est-à-dire qui permet d'apprendre d'événements indésirables (les accidents), mais aussi d'exemples à suivre pour les éviter (les presque accidents).

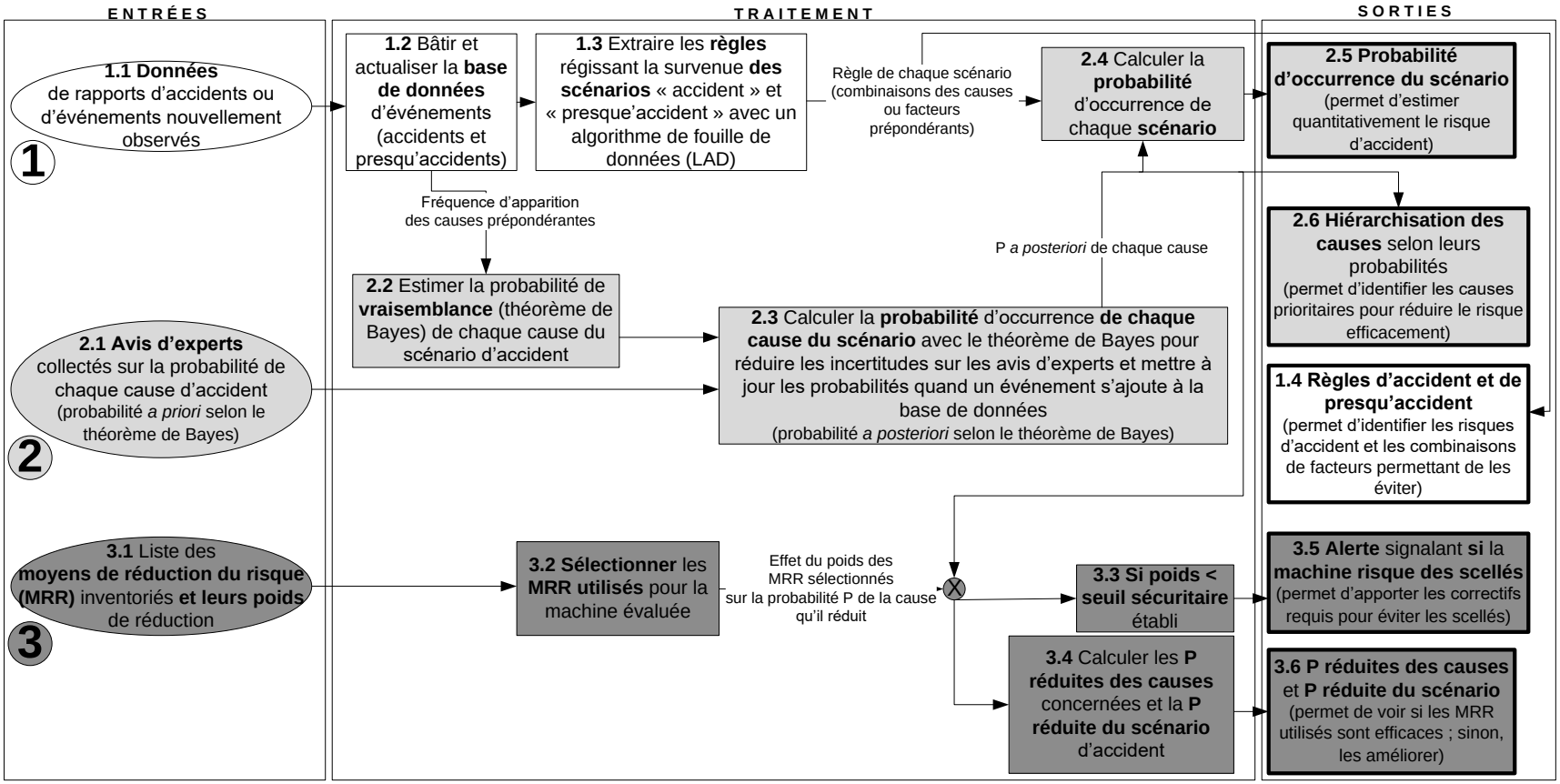


Figure 3.2 : Méthode de conception proposée pour l'outil dynamique d'aide à la décision (version française de la figure 3 de l'article 1)

L'outil visé était informatique. En raison de contraintes temporelles et logistiques, l'informatisation de la méthode a été exclue. Pour les mêmes motifs, les travaux de la thèse se sont limités aux étapes 1 et 2 de la figure 3.2. L'article 2 expérimente l'étape 1 en se basant sur des données d'accidents. L'expérimentation étant concluante, l'article 3 reprend l'étape 1 sous un format différent de la base de données d'accidents. Ce changement de format visait à permettre le calcul de la probabilité du dommage : contrairement à l'article 2 qui caractérisait des risques d'accident selon la tâche en cours (maintenance ou production), l'article 3 caractérise les risques d'accident selon le type de dommage (mortel ou non mortel). Notez que la méthode de calcul de probabilité proposée à l'étape 2 a été ajustée. Ainsi, l'inférence bayésienne annoncée à la figure 3.2 a été remplacée par une autre méthode de calcul présentée à la section 3.5.3. L'inférence bayésienne aurait permis de connaître l'impact des causes et facteurs de risque d'accident sur la probabilité des situations dangereuses identifiées, en collectant les avis *a priori* d'experts (ex., inspecteurs de la CNESST) sur la probabilité d'occurrence de ces causes et facteurs de risque. Contrairement, la démarche de la section 3.5.3 propose une alternative plus rapide pour connaître cet impact, en exploitant les fonctions de masse des causes et facteurs de risque d'accident. La rapidité de cette alternative réside dans le fait qu'aucune intervention humaine externe n'est requise. À l'opposé, l'inférence bayésienne aurait nécessité les interventions humaines externes suivantes : la sollicitation d'experts, des entrevues confidentielles avec eux, un certificat d'éthique autorisant la conduite des entrevues. Enfin, cet ajustement de la méthode de calcul de probabilité proposée à l'étape 2 s'explique par la volonté de suggérer un calcul d'application plus simple, tout en étant efficace, afin de respecter l'échéance de l'étude.

3.4 Vérification de l'applicabilité de l'ALD à un échantillon très restreint aux caractéristiques non évidentes

La fouille de données (étape C.5 de la figure 3.1) sert, normalement, à extraire des connaissances d'un grand nombre d'observations. Pour les raisons évoquées plus tôt, elle a quand même été utilisée pour un échantillon très restreint, mais composé de nombreux indicateurs (23 indicateurs : I_1 à I_{23}). Dans cette thèse, les données utilisées décrivent des accidents du travail

survenus sur des convoyeurs à courroie. Comme ces accidents proviennent du Centre de documentation de la CNESST, ils sont de nature grave (c.-à-d. grave non mortelle) ou mortelle.

Les étapes C.2 à C.8 de la figure 3.1 représentent le processus d'extraction de connaissances requis pour réaliser le REX dynamique. C'est un processus itératif, comme l'illustrent les boucles formées par les flèches reliant ces étapes. Les retours en arrière peuvent survenir à n'importe quelle étape du processus. Ces itérations sont synonymes d'ajustement permettant d'affiner la connaissance tirée des données.

3.4.1 Collecte, préparation et transformation des données

Avant de choisir l'algorithme de fouille de données, il a fallu collecter, analyser et trier l'information tirée des 23 rapports d'accidents liés aux convoyeurs à courroie. L'analyse de l'information textuelle de ces rapports et leur retranscription dans une base de données a constitué un travail colossal. En effet, une lecture complète de chaque rapport s'imposait. L'analyse de chaque rapport permettait d'identifier les facteurs de risques et causes de l'accident, puis de les noter dans un fichier *Excel*. Ces facteurs de risque et causes représentent les valeurs de variables (appelées indicateurs) décrivant les accidents analysés. De rapport en rapport, les causes puis facteurs de risques apparentés étaient regroupés dans une même colonne du fichier *Excel*. La colonne portait le nom de l'indicateur. Les diverses valeurs collectées par indicateur au fil de l'analyse des rapports représentaient ses valeurs possibles dans la base de données.

Le tableau 3.2 et le tableau 3.3 illustrent le processus de retranscription de l'information textuelle des rapports dans la base de données, en considérant deux exemples d'accidents référencés : CSST (1997) et CSST (2004a).

Tableau 3.2 : Exemple de retranscription d'information extraite de rapports d'enquête d'accident au sujet d'un facteur de risque

Référence du rapport d'enquête d'accident	**Extrait du rapport**	**Titre de la colonne *Excel* regroupant les valeurs de l'indicateur**	**Valeurs de l'indicateur inscrites dans le fichier *Excel* (la base de données)**
CSST (1997)	« Un programme de prévention existe depuis 1983. » (p. 6/14 du rapport)	→ Existence d'un programme de prévention N.B. Cet indicateur se nomme I_1 dans l'article 2 (cf. tableau 3.5).	→ Oui
CSST (2004a)	« L'employeur n'a pas élaboré de programme de prévention. » (p. 5/12 du rapport)		→ Non

Tableau 3.3 : Exemple de retranscription d'information extraite de rapports d'enquête d'accident au sujet d'une cause accidentelle

Référence du rapport d'enquête d'accident	**Extrait du rapport**	**Titre de la colonne *Excel* regroupant les valeurs de l'indicateur**	**Valeurs de l'indicateur inscrites dans le fichier *Excel* (la base de données)**
CSST (1997)	« [...] Monsieur "A" [...] a été coincé mortellement (asphyxié par compression) entre la courroie du convoyeur et la structure métallique de l'équipement. » (p. 13/14 du rapport)	→ Zone dangereuse accessible où l'accident s'est produit N.B. Cet indicateur se nomme I_{16} dans l'article 2 (cf. tableau 3.5).	→ Zone de coincement entre la courroie et le châssis
CSST (2004a)	« [...] M. "A" est trituré et coincé mortellement entre la structure du convoyeur et le tambour de queue du convoyeur. » (p. 3/12 du rapport)		→ Tambour de queue

Après la retranscription de l'information issue des rapports d'enquête, il fallait s'assurer que la base de données était dépourvue d'indicateurs redondants. Une fois la vérification réalisée, elle comportait 32 indicateurs (donc 32 colonnes) décrivant 23 accidents (donc 23 lignes). Ensuite, comme expliqué en section 3.3 de l'article 2, les indicateurs comprenant au moins 40 % de valeurs manquantes ont dû être éliminés. Le niveau de détails variant considérablement d'un rapport d'enquête d'accident à l'autre est à l'origine des valeurs manquantes. En effet, les rapports d'enquête d'accident analysés pouvaient aller d'une dizaine à une centaine de pages, annexes incluses. Un dernier indicateur a été supprimé puisqu'il ne comportait qu'une valeur unique dans toute la base de données. Il n'apportait donc pas de connaissances pour distinguer les classes d'accidents. Après ces suppressions, il restait 6 indicateurs comprenant 4 à 26 % de valeurs manquantes. Comme l'explique la section 3.3 de l'article 2, l'impact de ces pourcentages sur la précision de classification des règles est minime. Ainsi, des valeurs approximatives ont été attribuées aux valeurs manquantes restantes, en appliquant la méthode décrite à la section 8.2.2 de Pyle (1999). Celle-ci suggère d'approximer une valeur manquante à la valeur qui perturbe le moins l'écart-type S_o de l'ensemble des valeurs connues de l'indicateur. En d'autres termes, l'objectif est de remplacer la valeur manquante par une valeur qui permet d'obtenir un écart-type final S_f le plus proche de S_o. (cf. tableau 3.4, cas de l'indicateur I_1). Dans le cas où les valeurs possibles de remplacement induisent des S_f également distancés de S_o, la valeur possible de remplacement sera celle qui permet d'obtenir la moyenne finale M_f la plus proche de la moyenne initiale M_o (cf. tableau 3.4, cas de l'indicateur I_8).

Tableau 3.4 : Exemple de remplacement des valeurs manquantes

Indicateur	**Moyenne et écart-type initiaux avec valeurs manquantes**	**Valeurs possibles de l'indicateur**	**Moyenne et écart-type initiaux avec les valeurs approximatives possibles des valeurs manquantes**	**Valeur approximative choisie**
I_1	$M_0 = 0.60$; $S_0 =$ **0.50**	0	$M_f = 0.61$; $S_f =$ **0.50**	→ $\boldsymbol{I_1 = 0}$
		1	$M_f = 0.65$; $S_f = 0.49$	
I_8	$M_0 =$ **0.32**; $S_0 = 0.48$	0	$M_f =$ **0.30**; $S_f = 0.47$	→ $\boldsymbol{I_8 = 0}$
		1	$M_f = 0.35$; $S_f = 0.49$	

Après l'élimination des 9 indicateurs et le remplacement des valeurs manquantes restantes, la base de données comportait 23 accidents et 23 indicateurs (cf. tableau 3.5).

Tableau 3.5 : Définition des 23 indicateurs utilisés dans la base de données

Indicateur	**Définition**	**Valeurs possibles**
I_1	Existence d'un programme de prévention	Non = 0; oui = 1
I_2	Ancienneté au poste	0 à 4 ans = 1; 5 à 10 ans = 2; 20 à 24 ans = 3
I_3	Travailleur de la compagnie ou sous-traitant	Compagnie = 1; sous-traitant = 2
I_4	Activité habituelle du travailleur	Non = 0; oui = 1
I_5	Travailleur formé à l'utilisation de cette machine en particulier	Non = 0; oui = 1
I_6	Accident dû au lieu ou à l'environnement de travail (ex. : encombrement des lieux)	Non = 0; oui = 1
I_7	Agent causal	Angle rentrant = 1; zone de coincement = 2
I_8	Dysfonctionnement de la machine	Non = 0; oui = 1
I_9	Accident survenu lors d'une mise en route intempestive de la machine	Non = 0; oui = 1
I_{10}	Accident survenu lorsque la machine était en marche automatique	Non = 0; oui = 1
I_{11}	Protection présente (c.-à-d., sur place au moment de l'accident)	Non = 0; oui = 1
I_{12}	Avertissements et signalisations lacunaires ou défaillants impliqués dans l'accident	Non = 0; oui = 1
I_{13}	Équipement de protection individuelle (EPI), vêtement de travail ou outil en cause lors de l'accident	Non = 0; oui = 1
I_{14}	Procédure de cadenassage appliquée	Non = 0; oui = 1
I_{15}	Contribution du système de commande à l'accident	Non = 0; oui = 1
I_{16}	Zone dangereuse accessible où l'accident s'est produit	Zone entre la courroie ou le châssis et une autre structure = 1; poulie = 2; rouleau = 3; tambour d'entraînement = 4; tambour tendeur = 5; tambour de queue = 6
I_{17}	Gestion déficiente de la santé et de la sécurité au travail (SST)	Non = 0; oui = 1
I_{18}	Présence d'un comité santé-sécurité (CSS)	Non = 0; oui = 1
I_{19}	Dommage	Avant-bras écrasé = 1; bras amputé = 2; mort = 3
I_{20}	Autre équipement (en plus du convoyeur) impliqué dans l'accident	Non = 0; oui = 1
I_{21}	Absence d'un dispositif d'arrêt d'urgence accessible en cause	Non = 0; oui = 1
I_{22}	Circonstance accidentelle	Écrasement = 1; coincement = 2; frappé par = 3; étranglement = 4; arrachement = 5; Entraînement et asphyxie par ensevelissement = 6
I_{23}	Partie du corps dans la zone dangereuse et siège de la lésion	Corps = 1; tronc = 2; tête = 3; bras gauche = 4; bras droit = 5; les deux bras = 6; cou = 7

Plus de détails quant à la collecte, la préparation et la transformation des données (étape C.2 à C.4 de la figure 3.1) sont disponibles aux sections 3.1 à 3.4 de l'article 2.

3.4.2 Transformation des données selon les algorithmes et logiciels testés

Cette section porte notamment sur les étapes C.4, C.5 et C.8 de la méthodologie illustrée à la figure 3.1.

Avant de découvrir l'algorithme ALD et de le qualifier d'adapté à cette étude, d'autres algorithmes ont été testés pour l'étape de fouille de données. Il s'agit d'algorithmes relatifs à la technique des règles d'association puis celle des arbres de décision. Les autres techniques comme les réseaux de neurones artificiels ont été écartées dès le départ, car cette technique est reconnue gourmande en observations. En effet, elle requiert beaucoup d'observations pour pouvoir générer des règles peu sensibles au bruit dans les données et capables de prédire avec une précision optimale des observations futures. Le bruit désigne un changement subtil apporté à l'une ou plusieurs valeurs d'un indicateur de la base de données.

Selon l'algorithme choisi, il a fallu adapter la transformation des données. Initialement, les valeurs des indicateurs étaient des mots écrits en toutes lettres. Par exemple, la valeur « Oui » pour l'indicateur I_1 (cf. tableau 3.5). Cette représentation convenait aux algorithmes de règles d'association et d'arbres de décisions employés dans le premier logiciel utilisé : *Tanagra* (Rakotomalala, 2010). Toutefois, par la suite, pour les raisons évoquées vers la fin de cette section, l'algorithme ALD a été utilisé. Alors, les valeurs initialement qualitatives ont dû être transformées en valeurs numériques puisqu'il s'agit d'une exigence de l'ALD. Ainsi, à titre d'exemple, la valeur « Oui » de l'indicateur I_1 est devenue 1 comme indiqué plus tôt au tableau 3.5.

Comme les règles d'association et les arbres de décision sont basés sur la détection d'ensembles fréquents ou éléments caractéristiques fréquents dans les données, ces deux techniques ont été utilisées en espérant trouver, dans l'échantillon d'accidents de maintenance et de production, des ensembles fréquents à l'origine de règles convaincantes. Une règle est convaincante si sa conviction vaut au moins 1 (Brin et al., 1997). Une conviction supérieure à 1 signifie que la règle n'est pas le fruit du hasard.

Considérons une règle de la forme : $A \rightarrow B$. La lettre *A* représente l'antécédent et *B* la conséquence. Dans le cadre de cette thèse, l'antécédent *A* serait un indicateur d'accident ou une combinaison d'indicateurs respectant certaines conditions. La conséquence *B* symboliserait la classe (le type d'accident). La conviction se définit ainsi :

$$Conviction = \frac{P(A\&B)}{P(A)\cdot P(B)} = \frac{Support}{P(A)\cdot P(B)} = \frac{Confiance}{P(B)} \qquad \text{(Éq. 1)}$$

où :

P(A&B) : la probabilité d'avoir *A* et *B* dans les données. C'est le support de la règle.

P(A) : la probabilité d'avoir *A* dans les données.

P(B) : la probabilité d'avoir B dans les données.

P(A&B)/P(A) : la probabilité d'avoir l'occurrence *B* dans les données sachant qu'on a déjà *A*. C'est la confiance de la règle (Brin et al., 1997).

Les règles d'association et les arbres de décision ont été lancés avec le logiciel *Tanagra* en testant la base de données d'accidents de maintenance et de production. L'annexe D présente les résultats détaillés de la fouille de données lancée avec ce logiciel pour les règles d'association. Dans ce cas, l'algorithme « *Supervised Association Rule* » basé sur l'approche « *A priori* » a été utilisé. Quant aux arbres de décisions, l'annexe E présente les résultats détaillés liés à l'utilisation des algorithmes « *C4.5* » et « *ID3* » rattachés à cette technique.

Tanagra a généré trois règles d'association attitrées à la classe « Accidents en production » (cf. tableau 3.6). Pour générer des règles avec *Tanagra*, un paramétrage de l'algorithme choisi est requis. Dans le cas des règles d'association, le paramétrage consiste essentiellement à indiquer au logiciel la valeur minimale souhaitée pour le support, la confiance et la conviction. En ce qui concerne les règles d'association générées pour la classe « Accidents en production », les valeurs minimales sont mentionnées au tableau 3.6. Le support minimal a été calculé selon la logique suivante :

- chercher à obtenir des ensembles fréquents caractéristiques pour toutes les observations de la classe « Accidents en production ». Comme la base de données contient 3 accidents en production, le support minimal avoisinait $3/23 \approx 0{,}13$;

- si aucune règle n'avait été obtenue avec ce minimum, la valeur minimale du support aurait été réduite jusqu'à l'obtention d'une règle, tout en maintenant la valeur minimale de la conviction.

Les seuils minimaux relatifs au support et à la confiance sont déterminés par l'utilisateur, selon la nature des données et de l'objectif de la fouille de données. L'idéal est de rechercher une confiance la plus élevée possible puisque c'est un indicateur de « précision » des règles d'association (Rakotomalala, s.d.). Par défaut, *Tanagra* suggère une confiance minimale de 0,75. En effectuant plusieurs tests sous ce logiciel avec un même support minimal de 0,1, mais des confiances différentes, nous constatons que les règles d'association obtenues avec une confiance allant de 0,67 à 0,99 demeurent identiques à celles listées au tableau 3.6. Cependant, 15 règles sont obtenues lors des tests avec une confiance 0,66 et un support de 0,1. En effet, une astuce pour limiter la prolifération de règles est d'augmenter la confiance (Rakotomalala, s.d.). En ce qui concerne la thèse, il est préférable de limiter la prolifération des règles puisque gérer les risques associés aux machines est plus facile avec un nombre contrôlé de règles. Ainsi, à partir d'une observation présentant les mêmes caractéristiques qu'une règle d'association de la classe « Accidents en production » dans le cas du tableau 3.6, il y aura 70% de risque que cette observation présage un accident en production. La conviction minimale choisie au tableau 3.6 assure que la conviction soit supérieure à 1 afin d'éviter que les règles soient le fruit du hasard.

Tableau 3.6 : Règles d'association obtenues pour la classe « Accident en production » en fonction du paramétrage choisi

Paramétrage	**Règles d'association pour la classe « Accident en production »**
$Support_{min} = 0,1$ $Confiance_{min} = 0,7$ $Conviction_{min} = 1,1$	1 (I_5 = Oui) ET (I_{13} = Non) ET (I_{22} = Deux)
	2 (I_2 = Un) ET (I_5 = Oui) ET (I_{13} = Non)
	3 (I_1 = Non) ET (I_{13} = Non) ET (I_{22} = Deux)

La même logique a été suivie pour la classe « Accident en maintenance ». Comme il y avait 20 accidents en maintenance sur les 23 accidents de la base de données, la recherche de règles d'association a commencé avec un support minimal de 0,8 proche de 20/23 ≈ 0,87. N'ayant

obtenu aucune règle d'association, le support minimal a été réduit itérativement jusqu'à 0,04. Ce qui correspond à 1 accident sur 23. Aucune règle n'a été trouvée sous ces différents supports minimaux. Notez que pour cette classe, la confiance et la conviction minimales étaient identiques à celle de la classe « Accident en production ». Seulement pour la dernière tentative, la confiance a été réduite jusqu'à 0,6. Aucune règle n'a pu être générée.

N'arrivant pas à obtenir de règles pour caractériser et distinguer les deux classes, la génération d'un arbre de décision a été testée pour classer les accidents de la base de données. Toutefois, comme le montre l'annexe E, aucun arbre n'a pu être obtenu.

Comme mentionné dans l'article 2, les techniques de règles d'association et d'arbres de décision négligent certaines observations si celles-ci ne respectent pas le seuil du nombre d'ensembles fréquents demandé par l'utilisateur du logiciel. La rareté des ensembles fréquents dans la base de données utilisée pour cette thèse explique l'échec des algorithmes de règles d'association et d'arbre de décision précédemment cités. Alors, il a fallu trouver un autre type d'algorithme, d'approche différente, convenable à la base de données. Une recherche a donc été entreprise sur différentes études de fouille de données, dans le but de trouver un algorithme potentiellement performant, indépendamment de la taille de l'échantillon et du nombre d'ensembles fréquents. Au fil des recherches, nous comprenons petit à petit que l'algorithme ALD a le potentiel de caractériser des classes, malgré la rareté d'ensembles fréquents dans les données. Ce potentiel, suggéré entre autres par Lauer et al. (2002), a été confirmé dans l'article 2 de la thèse. En effet, le but de l'ALD est de trouver les caractéristiques qui distinguent des classes, au lieu de chercher des ensembles fréquents caractérisant des classes. L'ALD ne néglige aucune observation lors de l'apprentissage, à moins que des observations de classes différentes soient représentées par le même vecteur de données, donc soient contradictoires. En raison de ces aptitudes de classification, l'ALD a été adopté de manière définitive en vue des étapes C.5 et C.9 de la figure 3.1.

Le tableau 3.7 donne un aperçu de la base de données transformée pour l'ALD et utilisée dans l'article 2. Le chiffre 1 dans la colonne intitulée « Classe » correspond aux accidents de maintenance. Le chiffre 0 de cette colonne représente les accidents de production.

Tableau 3.7 : Aperçu de la base de données d'accidents des convoyeurs à courroie utilisée dans l'article 2

Accident No.	Classe	Indicateurs																						
		I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}	I_{11}	I_{12}	I_{13}	I_{14}	I_{15}	I_{16}	I_{17}	I_{18}	I_{19}	I_{20}	I_{21}	I_{22}	I_{23}
1	1	1	3	1	1	1	1	2	0	0	1	0	0	0	0	0	1	0	1	3	1	0	1	2
2	1	1	2	1	1	1	1	1	1	0	1	0	0	1	0	0	2	0	1	3	0	0	1	1
. . .																								
23	0	0	1	1	1	1	0	1	0	0	1	0	0	0	0	0	4	1	0	3	1	0	2	2

L'algorithme ALD n'étant pas disponible dans *Tanagra*, un autre logiciel, nécessitant une licence d'utilisation, a dû être utilisé pour l'article 2. Il s'agit de : *cbmLAD* (Yacout, 2010b).

3.4.3 Connaissances générées et précision de classification

Cette section porte notamment sur les étapes C.6 et C.7 de la méthodologie illustrée à la figure 3.1.

Le logiciel *cbmLAD* a généré des connaissances à partir de la base de données de 23 accidents. Les connaissances sont présentées sous forme de règles, avec leurs interprétations, au tableau 3.8 ci-dessous.

Tableau 3.8 : Règles générées à partir de la base de données, ainsi que leurs couvertures et prévalences relatives dans chaque classe

	Règle No.	Règle	Règle interprétée : accident survenu…	Couver-ture	Prévalence relative
Accidents de maintenance	P_1^+	$I_1 > 0{,}5$	...alors qu'un programme de prévention existait	12/20	60%
	P_2^+	$I_6 > 0{,}5$	...en raison d'un environnement lacunaire (ex., encombrement des lieux de travail)	14/20	70%
		$I_{16} > 1{,}5$	...à cause d'une zone dangereuse accessible : un angle rentrant impliquant une poulie *OU* un rouleau *OU* un tambour d'entraînement *OU* un tambour tendeur *OU* un tambour de queue		
		$I_{23} < 6{,}5$	...en présence d'une partie du corps dans la zone dangereuse : le corps *OU* le tronc *OU* la tête *OU* le bras gauche *OU* le bras droit *OU* les deux bras		
	P_3^+	$I_{22} < 1{,}5$	...en raison d'un écrasement	11/20	55%
Accidents de production	P_1^-	$I_1 < 0{,}5$	...en l'absence d'un programme de prévention	3/3	100%
		$I_8 < 0{,}5$	...alors qu'il n'y avait aucun dysfonctionnement du convoyeur		
		$I_{16} < 4{,}5$	...à cause d'une zone dangereuse accessible : un angle rentrant impliquant une poulie *OU* un rouleau *OU* un tambour d'entraînement		
		$I_{22} > 1{,}5$	...en raison d'un coincement *OU* d'un étranglement		

N.B. La couverture est le nombre d'accidents que décrit la règle.

Chaque règle représente un scénario accidentel ou situation dangereuse. On constate que les indicateurs I_1, I_6, I_8, I_{16}, I_{22}, I_{23} au tableau 3.8 permettent de distinguer les accidents de maintenance analysés par rapport à ceux de production. Les conditions établies par *cbmLAD* sur ces indicateurs permettent de caractériser le contexte accidentel pour chacune des classes d'accidents analysés. En effet, cette caractérisation est faite au moyen des règles composées de combinaisons de ces conditions. Par exemple, le contexte accidentel de la classe des accidents de maintenance est caractérisé par la disjonction de trois règles : P_1^+, P_2^+, P_3^+ (Éq.2). Celui de production est caractérisé par une seule règle : P_1^- (Éq. 3).

Contexte d'accident de maintenance = P_1^+ OU P_2^+ OU P_3^+ (Éq. 2)

Contexte d'accident de production = P_1^- (Éq. 3)

avec :

$P_1^+ = (I_1 > 0{,}5)$

$P_2^+ = (I_6 > 0{,}5)$ ET $(I_{16} > 1{,}5)$ ET $(I_{23} < 6{,}5)$

$P_3^+ = (I_{22} < 1{,}5)$

$P_1^- = (I_1 < 0{,}5)$ ET $(I_8 < 0{,}5)$ ET $(I_{16} < 4{,}5)$ ET $(I_{22} > 1{,}5)$

Une discussion relative aux règles est présentée à la section 6 de l'article 2.

Comme dans tout processus d'exploration de données, une fois la connaissance générée, il faut la valider. La section 4 de l'article 2 détaille le processus de validation en passant par deux méthodes : *Leave-One-Out Cross-Validation*, puis *5-Fold Cross-Validation*. La première est choisie en raison du peu de données. Elle consiste, dans notre cas, à effectuer 23 fois l'apprentissage, à partir de bases de données de 22 accidents et garder le 23ème accident restant pour la base de données « Test ». Bien que la première méthode de validation soit adaptée pour les bases de données restreintes, son défaut est le manque de représentativité de toutes les classes dans la base de données « Test ». Pour pallier cet inconvénient, la seconde méthode est utilisée. Celle-ci consiste à répartir les données, à 5 reprises. À chaque reprise, l'apprentissage est effectué à partir de 80 % de la base de données, tandis que la validation est réalisée à partir du 20 % restant de la base de données, tout en prenant soin d'y représenter les classes d'accidents selon des proportions similaires à la base de données initiale. La première méthode aboutit à une précision de classification des règles de 74 %, la seconde, à 72 %. Il s'agit de précisions satisfaisantes pour cette base de données restreintes, sachant que des précisions allant de 71,4 à 74,4 % ont été considérées comme adéquates pour une base de données de 768 observations relative au diabète (Boros et al., 2000). Ainsi, l'article 2 a montré que l'ALD est un algorithme approprié pour la problématique traitée dans cette thèse. Il est donc capable de caractériser et distinguer deux classes d'observations malgré l'échantillon très restreint. Notons toutefois qu'avec les règles obtenues, il est impossible de généraliser les connaissances tirées. Compte tenu de l'aspect très restreint des échantillons d'apprentissage et de test, les combinaisons d'indicateurs formant les règles caractérisent uniquement les cas des entreprises concernées par les 23 accidents analysés. Ainsi, les moyens à adopter pour réduire le risque et la hiérarchisation obtenue pour prioriser les facteurs de risque se limitent aux cas de ces entreprises uniquement.

Ces moyens de réduction du risque et leur hiérarchisation ne peuvent donc être étendus à d'autres entreprises tels quels.

La précision de classification est adéquate et valide donc le modèle conceptuel (c.-à-d., les règles générées par *cbmLAD*). Cette précision valide également que l'ALD est approprié pour l'analyse d'échantillons restreints de données. Ceci étant validé, le calcul de la probabilité du dommage à partir de règles générées par l'algorithme ALD est traité à la section 3.5.

3.5 Estimation de la probabilité du dommage

3.5.1 Nouvelle extraction de connaissances

L'article 3 traite de calculs de probabilité du dommage. Dans l'article 2, les accidents étaient classés selon le type de tâche : maintenance ou production. Dans une même classe, la gravité du dommage variait et les nuances entre les niveaux de gravité de ces dommages étaient discutables. Ainsi, pour évaluer le risque, il aurait fallu discuter avec les survivants des accidents et les collègues des victimes décédées puisque l'appréciation du risque est une affaire de consensus entre les parties prenantes. Le but de l'article 3 étant d'illustrer une méthode, il était plus judicieux de choisir un cas avec des niveaux de gravité indubitable. Ainsi, dans le cas de l'article 3, les classes d'accidents représentent deux types de dommage indubitables : dommage mortel ou dommage non mortel. Ce choix a impliqué d'adapter, pour les besoins de l'article 3, la base de données d'accidents utilisée dans l'article 2. En effet, les mêmes accidents composent la base de données des deux articles, mais sont classés différemment, comme mentionné plus tôt au tableau 3.1. La nouvelle base de données restreinte étant différente de la précédente, il a fallu exécuter à nouveau le processus d'extraction de connaissances pour caractériser les nouvelles classes : « Mortel » et « Non mortel ».

Le choix de disposer d'une classe « Mortel » et une autre « Non mortel » permet d'obtenir un espace échantillon aux événements complémentaires, condition requise afin de calculer des probabilités. Dans l'article 2, la classe d'accidents liés à la maintenance n'est pas complémentaire à celle de production, car des accidents peuvent survenir lors d'autres types de tâches, tel que le réglage. Ainsi, pour calculer la probabilité du dommage, un espace échantillon d'accidents

comprenant uniquement des accidents ayant lieu durant des tâches de maintenance et de production est insuffisant. Il aurait fallu couvrir tous les types de tâches qui existent.

Finalement, l'article 3 vise à estimer quantitativement le risque encouru sur une machine, en prenant pour exemple des convoyeurs à courroie. Rappelons que la norme ISO 12100:2010 en sécurité des machines définit le risque comme étant la combinaison de la probabilité du dommage ($\mathcal{P}_{dommage}$) et de la gravité de ce dommage ($G_{dommage}$). La base de données choisie est telle qu'elle permet la prise en compte de ces deux paramètres principaux. En effet, les deux classes choisies (« Mortel » et « Non mortel ») permettent de considérer la gravité du dommage. Calculer la probabilité de chaque règle caractéristique d'une classe correspond à estimer la probabilité du dommage relatif à la situation dangereuse que représente la règle. Ultimement, calculer la probabilité d'occurrence de chaque classe permettrait d'aboutir à la probabilité globale du dommage.

La liste suivante décrit les transformations apportées à la base de données de l'article 2 au profit de l'article 3 :

- la classe désignant le type de tâche (maintenance ou production) lors de l'accident est devenue un indicateur (voir l'indicateur V_{20} au tableau 3.9);
- les anciens indicateurs I_{19}, I_{22} et I_{23} (cf. tableau 3.9) ont été éliminés pour former les classes « Mortel » et « Non mortel » dans l'article 3;
- les données liées au moment des accidents (V_1 : quart de travail) ont été ajoutées aux indicateurs initiaux, ainsi que le type de convoyeur à courroie (V_2).
- L'ajout de ces deux indicateurs vient du fait que la base de données est formatée en fonction d'un concept d'enquête et analyse d'accidents appelé « MELITO ». Ce choix a été motivé par le fait que les préventionnistes et enquêteurs d'accident du travail du Québec sont familiers avec ce concept. De surcroît, modéliser un accident selon un tel concept permet d'obtenir un modèle de type « système » comme le préfèrent Hollnagel (2004), Dekker (2006) et Leveson (2011). Ils trouvent les modèles « systèmes » plus adaptés que les modèles séquentiels et épidémiologiques puisqu'ils prennent en compte la globalité du contexte entourant un événement. Plus de détails sur la vision de Hollnagel (2004), Dekker (2006) et Leveson (2011) sont disponibles à la section 2 de l'article 3;

- la base de données de l'article 3 a été formatée selon MELITO :
    - d'abord, les 23 indicateurs ont été répartis selon 6 groupes (cf. tableau 3.9 et figure 3.3). Chacun de ces groupes correspond à une lettre du concept : MELITO, d'où l'appellation « indicateur groupeur » attribué à ces six lettres. Le tableau 3.9 illustre la correspondance entre les indicateurs utilisés dans l'article 2 et ceux employés dans l'article 3;
    - ensuite, les valeurs possibles de chaque indicateur groupeur ont été déterminées à partir des valeurs associées aux indicateurs correspondant, grâce à un algorithme d'apprentissage non supervisé (cf. définition à l'encadré ci-dessous). Il s'agit de l'algorithme de classification hiérarchique HAC (*Hierarchical Agglomerative Clustering*) de *Tanagra*.

La figure 3.3 illustre le passage de la base de données de l'article 2 à celle de l'article 3. La section 4 de l'article 3 détaille davantage ce formatage.

Définition : Apprentissage supervisé vs apprentissage non supervisé

En apprentissage supervisé, les classes d'observations sont connues. Par exemple, l'ALD est un algorithme d'apprentissage supervisé. Un tel algorithme vise à caractériser chacune des classes qui lui sont indiquées.

En apprentissage non supervisé, les observations existent sans être classées. Les classes et leur nombre sont donc inconnus. Un algorithme d'apprentissage non supervisé vise à déceler des regroupements dans les données, en identifiant des caractéristiques communes entre des observations.

Tableau 3.9 : Description et correspondance entre les indicateurs utilisés dans les articles 2 et 3

Indicateur dans l'article 2	**Indicateur correspondant ou ajouté dans l'article 3**	**Définition de l'indicateur**	**Indicateur groupeur dans l'article 3**
---	V_1	Quart de travail	***M***
---	V_2	Type de convoyeur à courroie (machine)	***E***
I_7	V_3	Agent causal	***E***
I_8	V_4	Dysfonctionnement de la machine	***E***
I_9	V_5	Accident survenu lors d'une mise en route intempestive de la machine	***E***
I_{10}	V_6	Accident survenu lorsque la machine était en marche automatique	***E***
I_{11}	V_7	Protection présente (c.-à-d., sur place au moment de l'accident)	***E***
I_{12}	V_8	Avertissements et signalisations lacunaires ou défaillants impliqués dans l'accident	***E***
I_{13}	V_9	Équipement de protection individuelle (EPI), vêtement de travail ou outil en cause lors de l'accident	***E***
I_{15}	V_{10}	Contribution du système de commande à l'accident	***E***
I_{16}	V_{11}	Zone dangereuse accessible où l'accident s'est produit	***E***
I_{20}	V_{12}	Autre équipement (en plus du convoyeur) impliqué dans l'accident	***E***
I_{21}	V_{13}	Absence d'un dispositif d'arrêt d'urgence accessible en cause	***E***
I_6	V_{14}	Accident dû au lieu ou à l'environnement de travail (ex. : encombrement des lieux)	***L***
I_2	V_{15}	Ancienneté au poste	***I***
I_3	V_{16}	Travailleur de la compagnie ou sous-traitant	***I***
I_4	V_{17}	Activité habituelle du travailleur	***I***
I_5	V_{18}	Travailleur formé à l'utilisation de cette machine en particulier	***T***
I_{14}	V_{19}	Procédure de cadenassage appliquée	***T***
---	V_{20}	Opération en cours lors de l'accident	***T***
I_1	V_{21}	Existence d'un programme de prévention	***O***
I_{17}	V_{22}	Gestion déficiente de la santé et de la sécurité au travail (SST)	***O***
I_{18}	V_{23}	Présence d'un comité santé-sécurité (CSS)	***O***
I_{19}	---	Dommage	---
I_{22}	---	Circonstance accidentelle	---
I_{23}	---	Partie du corps dans la zone dangereuse et siège de la lésion	---

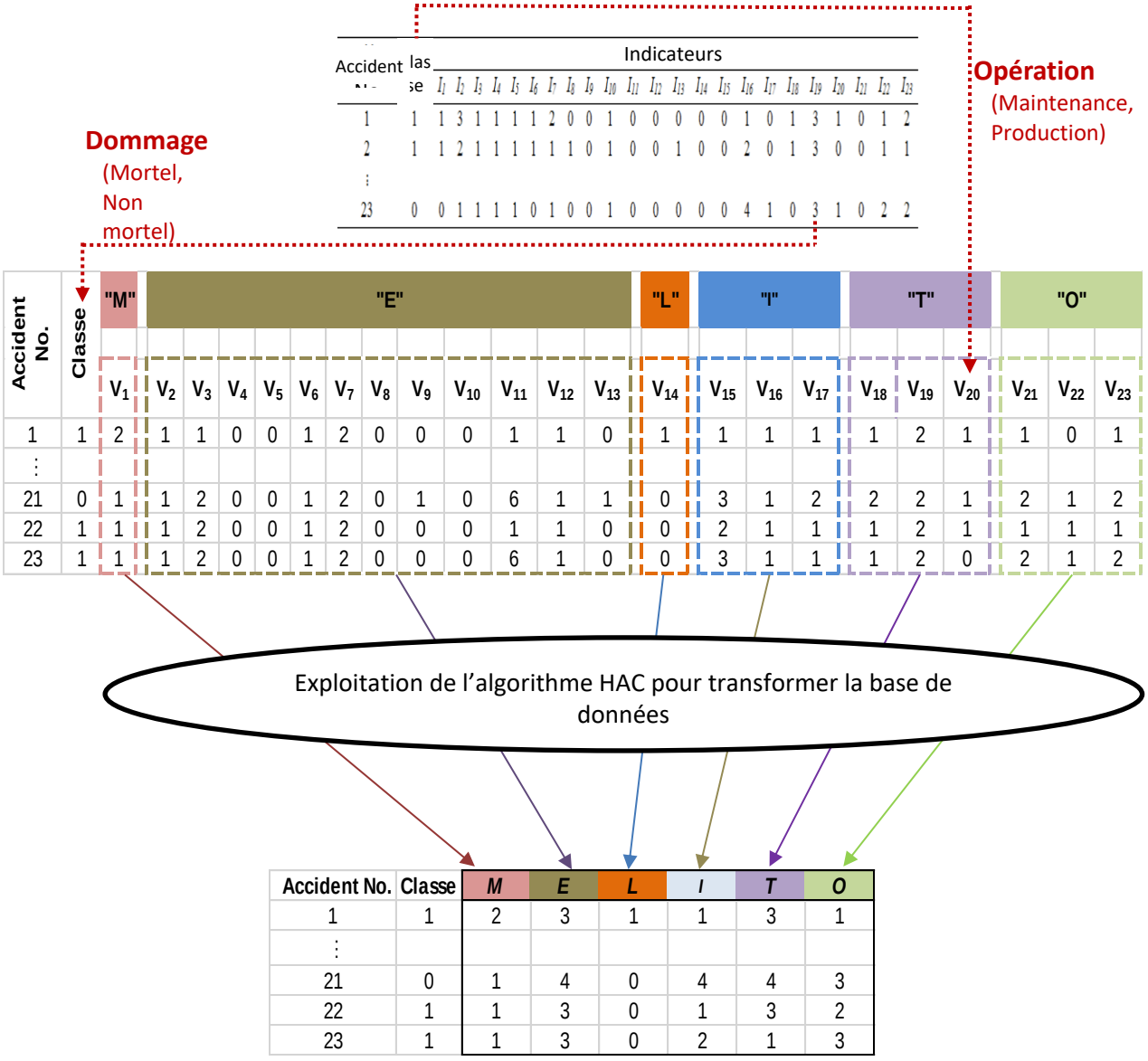


Figure 3.3 : Passage de la base de données de l'article 2 à celle de l'article 3

Comme l'illustre la figure 3.1, la recherche associée à l'article 3 commence à partir de l'étape C.2 : la collecte des données. L'étape C.1 concernant le type de machine est exclue. La base de données utilisée pour l'article 3 étant une réorganisation de celle de l'article 2, le même type de machine est donc utilisé. L'exécution des étapes C.2 à C.5 de la figure 3.1 constituant le processus d'extraction de connaissances de l'article 3 est similaire à celle de l'article 2. En revanche, à l'étape C.5, la fouille de données a été lancée avec le logiciel, sans licence d'utilisation, *LAD-WEKA* (http://www.lia.ufc.br/~tiberius/lad/) qui est une mise en œuvre de

l'ALD dans l'environnement du logiciel *WEKA* (Hall et al., 2009). Le tableau 3.10 et le tableau 3.11 listent les règles générées avec le logiciel. La colonne « Accidents couverts » de ces tableaux liste les numéros d'accidents caractérisés par chaque règle. Chacun des 23 accidents est représenté par un numéro comme le montrent ce tableau et la figure 3.4.

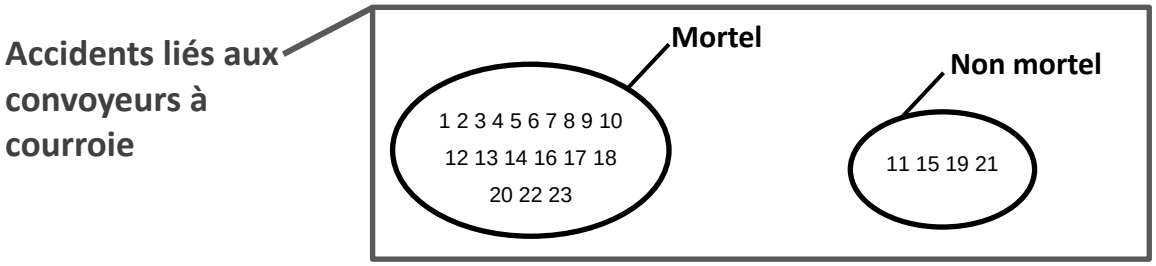


Figure 3.4 : Répartition, par classe, des accidents associés aux convoyeurs à courroie

Tableau 3.10 : Règles générées par *LAD-WEKA* pour la classe « Mortel » et ordonnées en fonction de leur couverture

Règle							**Accidents couverts**	**Couverture**
P_6^+ =		($E \leq 4.5$)	ET			($O \leq 2.5$)	1 2 3 4 5 7 12 13 16 17 20 22	12/19
P_2^+ =		($E \leq 3.5$)	ET			($O \leq 2.5$)	1 2 3 4 5 7 13 16 17 20 22	11/19
P_5^+ =		($E \leq 4.5$)	ET	($I \leq 2.5$)			1 2 5 6 8 9 12 14 18 22 23	11/19
P_4^+ =	($M \leq 1.5$) ET	($E \leq 3.5$)					3 5 7 9 13 16 17 18 22 23	10/19
P_3^+ =		($E \leq 3.5$)	ET	($I \leq 2.5$)			1 2 5 8 9 18 22 23	8/19
P_7^+ =				($I \leq 2.5$)	ET	($O > 2.5$)	6 8 9 14 18 23	6/19
P_1^+ =		($E > 4.5$)					10	1/19

Tableau 3.11 : Règles générées par *LAD-WEKA* pour la classe « Non mortel » et ordonnées en fonction de leur couverture

Règle					**Accidents couverts**	**Couverture**
P_1^- =		(E = 4) ET	(I > 2,5)		15 21	2/4
P_3^- =		(E ≤ 4,5) ET	(I > 2,5) ET	(O > 2,5)	11 21	2/4
P_5^- =	(M ≤ 1,5) ET	(E = 4) ET		(O ≤ 2,5)	15 19	2/4
P_7^- =	(M ≤ 1,5) ET	(E > 3,5) ET		(O ≤ 2,5)	15 19	2/4
P_2^- =	(M ≤ 1.5) ET	(E = 4) ET	(I ≤ 2,5) ET	(O ≤ 2,5)	19	1/4
P_4^- =	(M ≤ 1,5) ET	(E ≤ 4,5) ET	(I > 2,5) ET	(O > 2,5)	21	1/4
P_6^- =		(E > 3,5) ET	(I > 2,5) ET	(O ≤ 2,5)	15	1/4
P_8^- =		(E = 4) ET	(I > 2,5) ET	(O > 2,5)	21	1/4
P_9^- =	(M > 1,5) ET	(E ≤ 3,5) ET	(I > 2,5) ET	(O > 2,5)	11	1/4
P_{10}^- =	(M > 1.5) ET		(I > 2,5) ET	(O > 2,5)	11	1/4
P_{11}^- =		(E ≤ 3,5) ET	(I > 2,5) ET	(O > 2,5)	11	1/4
P_{12}^- =	(M ≤ 1,5) ET	(E > 3,5) ET	(I > 2,5) ET	(O ≤ 2,5)	15	1/4

L'interprétation des règles (étape C.6 de la figure 3.1) se fait comme au tableau 2 de l'article 3, grâce au tableau A.2 de cet article permettant d'interpréter les valeurs des indicateurs groupeurs.

Ces règles générées avec *LAD-WEKA* ont une précision de classification de 78 % (étape C.7 de la figure 3.1). Cette précision de classification étant satisfaisante, le modèle conceptuel inhérent à l'article 3 est donc validé. L'étape C.8 de la figure 3.1 est donc ignorée. On passe directement à l'étape C.9 où les règles générées servent d'outil d'identification et de suivi des facteurs de risques et causes possibles d'accidents (pour plus de détails à ce sujet, voir la section 8 de l'article 3).

3.5.2 Autre logiciel utilisé pour l'ALD

Curieux de distinguer les habiletés de l'algorithme ALD selon le logiciel utilisé, un logiciel libre d'accès a été utilisé pour l'article 3, au lieu du logiciel payant *cbmLAD*. L'emploi de *LAD-WEKA*

a permis de constater qu'il est beaucoup moins optimisé que *cbmLAD* en ce qui a trait à la génération de règles. Il faut ajuster certains paramètres de *LAD-WEKA* afin de réduire le nombre de règles aux caractéristiques redondantes. Malgré ce paramétrage, certaines règles restent imbriquées dans d'autres (elles se chevauchent). Cette différence au niveau des règles générées s'expliquerait par le fait que *cbmLAD* exploite une version optimisée de l'approche MILP (*Mixed 0-1 Integer and Linear Programming*), tandis que *LAD-WEKA* utilise une approche par énumération. Boros et al. (2000) décrivent l'approche par énumération. Ryoo et Jang (2009) expliquent l'approche MILP.

À des fins de tests uniquement, la base de données de l'article 2 a été utilisée pour comparer les performances des deux logiciels. Après l'ajustement des paramètres de *LAD-WEKA*, nous obtenons les règles imbriquées suivantes parmi les 10 générées pour la classe « Accident en maintenance » :

- la règle : ($I_{22} \leq 1{,}5$);
- la règle : ($I_{22} \leq 1{,}5$) ET ($I_6 > 0{,}5$).

La règle $I_{22} \leq 1{,}5$ aurait suffi. En outre, en dépit de l'ajustement du paramétrage, *LAD-WEKA* génère plus de règles que *cbmLAD*. Le rapport de la quantité de règles versus la précision de classification est plus intéressant avec *cbmLAD* qu'avec *LAD-WEKA*. Le tableau 3.12 montre la différence entre ces deux logiciels en matière du nombre de règles générées et leur précision de classification.

Tableau 3.12 : Comparaison des performances de *cbmLAD* et *LAD-WEKA* selon le nombre de règles générées et leur précision de classification

	Logiciel utilisé		
Classe	***cbmLAD*** **(paramétrage automatique)**	***LAD-WEKA*** **(avec paramétrage par défaut)**	***LAD-WEKA*** **(avec ajustement du paramétrage)**
Accident en production	1 règle générée	129 règles générées	5 règles générées
Accident en maintenance	3 règles générées	19 règles générées	10 règles générées
Précision de classification	74%	87%	65%

Notons la précision de classification de 87% avec *LAD-WEKA* quand le nombre de règles générées totalise 148. Étant donné qu'obtenir plus de règles permet de couvrir différentes caractéristiques possibles des observations, il est normal d'arriver à une plus grande précision de classification que dans les cas où l'on en a moins. L'inconvénient est de devoir gérer un nombre considérable de caractéristiques d'accidents, alors que l'étude vise à cibler les plus importantes pour y remédier dans le cadre d'une gestion du risque d'accident.

Pour l'article 3, l'algorithme-glouton *Greedy Set-Covering* (GSC) de *LAD-WEKA* au lieu de l'*IteratedSampling* (IS) est utilisé afin de réduire le nombre de règles obtenues (Bonates et Gomes, 2014). Dans le GSC, l'heuristique s'exécute une seule fois, tandis que dans l'IS, elle s'exécute de manière itérative. Pour les besoins de l'article 3, la performance du logiciel *LAD-WEKA* est suffisante. En effet, l'article 3 se focalise sur les calculs de probabilités du dommage. Il est vrai que le nombre de règles générées par le logiciel rallonge le temps de recherche des règles complémentaires couvrant l'ensemble des observations d'une classe de dommage (une étape qui précède le calcul de la probabilité de la classe). Cependant, la génération d'un nombre plus important de règles ne pose pas de problème méthodologique quant à la procédure de calcul de la probabilité du dommage.

3.5.3 Calcul de la probabilité du dommage associé à une situation dangereuse

En premier lieu, la probabilité du dommage est considérée comme la probabilité du dommage relatif à une situation dangereuse. En d'autres termes, c'est la probabilité d'une règle. Cette probabilité est estimée en passant par le calcul de la fonction de masse de chaque indicateur groupeur. La fonction de masse $p_Y(y)$ est construite à partir de l'occurrence de l'indicateur groupeur dans toute la base de données.

L'indicateur groupeur est représenté par le symbole y. L'expression $p_Y(y)$ est équivalente à $\mathcal{P}[Y = y]$. La probabilité $\mathcal{P}[Y = y]$ est approximée à la fréquence relative de la valeur y dans la base de données. Les fonctions de masse des indicateurs groupeurs de la base de données « MELITO » se trouvent du tableau 3.13 au tableau 3.18.

Tableau 3.13 : Fonction de masse associée à l'indicateur groupeur « *M* »

Valeurs possibles de « *m* »	1	2
$p_M(m)$	$^{16}/_{23}$	$^{7}/_{23}$
	0,696	0,304

Tableau 3.14 : Fonction de masse associée à l'indicateur groupeur « *E* »

Valeurs possibles de « *e* »	1	2	3	4	5
$p_E(e)$	$^{1}/_{23}$	$^{3}/_{23}$	$^{12}/_{23}$	$^{6}/_{23}$	$^{1}/_{23}$
	0,043	0,130	0,522	0,261	0,043

Tableau 3.15 : Fonction de masse associée à l'indicateur groupeur « *L* »

Valeurs possibles de « *l* »	0	1
$p_L(l)$	$^{5}/_{23}$	$^{18}/_{23}$
	0,217	0,783

Tableau 3.16 : Fonction de masse associée à l'indicateur groupeur « *I* »

Valeurs possibles de « *i* »	1	2	3	4
$p_I(i)$	$^{6}/_{23}$	$^{6}/_{23}$	$4/_{23}$	$7/_{23}$
	0,261	0,261	0,174	0,304

Tableau 3.17 : Fonction de masse associée à l'indicateur groupeur « *T* »

Valeurs possibles de « *t* »	1	2	3	4
$p_T(t)$	$^{3}/_{23}$	$^{1}/_{23}$	$^{9}/_{23}$	$^{10}/_{23}$
	0,130	0,043	0,391	0,435

Tableau 3.18 : Fonction de masse associée à l'indicateur groupeur « *O* »

Valeurs possibles de « *o* »	1	2	3
***p*O(*o*)**	$^{4}/_{23}$	$^{10}/_{23}$	$^{9}/_{23}$
	0,174	0,435	0,391

En considérant les fonctions de masse précédentes et en supposant que les indicateurs groupeurs sont indépendants entre eux, la probabilité d'une règle se calcule comme dans l'exemple suivant concernant la règle P_7^+ :

$\mathcal{P}[P_7^+] = \mathcal{P}[I \leq 2{,}5] \times \mathcal{P}[O > 2{,}5]$ puisque $P_7^+ = (I \leq 2{,}5)$ ET $(O > 2.5)$ d'après le tableau 3.10.

$\mathcal{P}[P_7^+] = \mathcal{P}\big[\big((I = 1) \cup (I = 2)\big) \cap (O = 3)\big]$
$\mathcal{P}[P_7^+] = (\mathcal{P}[I = 1] + \mathcal{P}[I = 2]) \times (\mathcal{P}[O = 3])$
$\approx (0{,}261 + 0{,}261) \times (0{,}391) \leftarrow$ valeurs issues des fonctions de masse de *I* et de *O*
$\mathcal{P}[P_7^+] \approx 0{,}204$

En second lieu, pour avoir une vue d'ensemble des probabilités des scénarios accidentels, il est proposé d'estimer globalement la probabilité du dommage. Elle équivaut à la probabilité de se trouver dans une classe d'accidents. Ainsi, la probabilité d'un dommage mortel est celle de la classe d'accidents mortels; la probabilité d'un dommage non mortel est celle de la classe d'accidents non mortels. Toutefois, comme le montrent les différents scénarios de calculs suivants, cette probabilité est difficile à estimer. En effet, un processus par essai-erreur (section 3.5.4) poursuit l'objectif de trouver des probabilités complémentaires pour les classes d'accidents à partir des règles générées. Cependant, on aboutit à une somme différente de 1 pour les probabilités des deux classes d'accidents. D'après les scénarios de calculs suivants, cela s'expliquerait par l'imbrication de certaines règles dans d'autres et l'absence de complémentarité entre les règles des deux classes, même si ces dernières sont de type complémentaire : c'est-à-dire, même si « Mortel » est la négation de « Non mortel ».

Malgré l'impossibilité d'estimer la probabilité globale du dommage, la probabilité du dommage par règle est d'une grande utilité pour les hiérarchiser, donc pour prioriser les situations dangereuses. En effet, plusieurs des règles générées par classe comportaient la même couverture (cf. tableau 3.10 et tableau 3.11). Hiérarchiser l'une par rapport à l'autre demeurait impossible.

Maintenant, avec la probabilité d'une règle, il est possible d'établir un ordre de priorité parmi les règles de même couverture (cf. tableau 3.19 et tableau 3.20).

Tableau 3.19 : Hiérarchie des règles de la classe « Mortel » en fonction de leurs couvertures puis de leurs probabilités

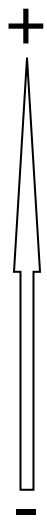
	Règle	**Couverture**	**Probabilité de la règle**	**Importance**
Mortel	$P_6^+ = (E \leq 4{,}5)$ ET $(O \leq 2{,}5)$	12/19	0,582	
	$P_5^+ = (E \leq 4{,}5)$ ET $(I \leq 2{,}5)$	11/19	0,499	
	$P_2^+ = (E \leq 3{,}5)$ ET $(O \leq 2{,}5)$	11/19	0,423	
	$P_4^+ = (M \leq 1{,}5)$ ET $(E \leq 3{,}5)$	10/19	0,484	
	$P_3^+ = (E \leq 3{,}5)$ ET $(I \leq 2{,}5)$	8/19	0,363	
	$P_7^+ = (I \leq 2{,}5)$ ET $(O > 2{,}5)$	6/19	0,204	
	$P_1^+ = (E > 4{,}5)$	1/19	0,043	

Tableau 3.20 : Hiérarchie des règles de la classe « Non mortel » en fonction de leurs couvertures puis de leurs probabilités

	Règle	**Couv.**	**Probabilité**	**Importance**
Non mortel	P_3^- = ($E \leq 4{,}5$) ET ($I > 2{,}5$) ET ($O > 2{,}5$)	2/4	0,179	
	P_7^- = ($M \leq 1{,}5$) ET ($E > 3{,}5$) ET ($O \leq 2{,}5$)	2/4	0,129	
	P_1^- = ($E = 4$) ET ($I > 2{,}5$)	2/4	0,125	
	P_5^- = ($M \leq 1{,}5$) ET ($E = 4$) ET ($O \leq 2{,}5$)	2/4	0,110	
	P_{11}^- = ($E \leq 3{,}5$) ET ($I > 2{,}5$) ET ($O > 2{,}5$)	1/4	0,130	
	P_4^- = ($M \leq 1{,}5$) ET ($E \leq 4{,}5$) ET ($I > 2{,}5$) ET ($O > 2{,}5$)	1/4	0,125	
	P_6^- = ($E > 3{,}5$) ET ($I > 2{,}5$) ET ($O \leq 2{,}5$)	1/4	0,089	
	P_{12}^- = ($M \leq 1{,}5$) ET ($E > 3{,}5$) ET ($I > 2{,}5$) ET ($O \leq 2{,}5$)	1/4	0,062	
	P_2^- = ($M \leq 1{,}5$) ET ($E = 4$) ET ($I \leq 2{,}5$) ET ($O \leq 2{,}5$)	1/4	0,058	
	P_{10}^- = ($M > 1{,}5$) ET ($I > 2{,}5$) ET ($O > 2{,}5$)	1/4	0,057	
	P_8^- = ($E = 4$) ET ($I > 2{,}5$) ET ($O > 2{,}5$)	1/4	0,049	
	P_9^- = ($M > 1{,}5$) ET ($E \leq 3{,}5$) ET ($I > 2{,}5$) ET ($O > 2{,}5$)	1/4	0,040	

Après la hiérarchie établissant l'importance entre les règles, une hiérarchie au niveau des indicateurs groupeurs permet de prioriser l'un par rapport l'autre en vue de la réduction du risque. Ainsi, en se basant sur la fréquence des indicateurs dans l'ensemble des règles d'une classe :

- Classe "Mortel": *E* (*f* = 6) > *I*, *O* (*f* = 3) > *M* (*f* = 1);
- Classe "Non mortel": *E*, *O* (*f* = 11) > *I* (*f* = 10) > *M* (*f* = 7);

on constate par exemple que l'équipement est le plus fréquent pour la classe « Mortel ». Alors, au moment de réduire le risque de dommage mortel inhérent à une situation dangereuse (une règle),

il faudra y parvenir en commençant par remédier aux causes et facteurs de risque (indicateurs) rattachés à l'équipement (indicateur groupeur).

3.5.4 Processus par essai-erreur pour estimer la probabilité globale du dommage

Un processus par essai-erreur en 4 étapes a été testé. La première examine la probabilité globale du dommage en considérant toutes les règles générées par classe. La deuxième cherche à comprendre le comportement de la probabilité en passant par une base de données pour laquelle toutes les observations possibles sont connues. L'exemple de la table logique « ET » est donc examiné à cette fin. Basée sur le comportement des probabilités calculées pour la table logique, la troisième étape propose d'éliminer toute règle dont les accidents couverts le sont déjà par une règle partageant au moins une condition similaire. La quatrième étape vise à réduire la redondance d'information entre les règles en cherchant à considérer seulement le nombre minimal de règles couvrant tous les accidents.

3.5.4.1 Probabilité globale du dommage à partir de toutes les règles générées

Pour ce faire, nous posons que la probabilité du dommage équivaut à la probabilité de l'union des règles générées par classe. Ce choix se base sur le fait que toute classe est caractérisée par l'union des règles qui lui sont associées. La probabilité d'une union est définie par la formule de Poincaré (Éq. 4) (Jourdain, 2013). Elle stipule que pour un espace probabiliste de n éléments A_1, A_2, ... A_n appartenant à A et partageant ou non des intersections, la probabilité de leur union se calcule par :

$$\mathcal{P}[\bigcup_{i=1}^{n} A_i] = \sum_{k=1}^{n}\left((-1)^{k+1} \sum_{1\leq i_1<i_2<\cdots<i_k\leq n} \boldsymbol{\mathcal{P}}(A_{i1} \cap A_{i2} \cap \ldots \cap A_{ik})\right) \quad \text{(Éq. 4)}$$

Dans notre contexte, A représente une classe d'accidents. Chacune des règles d'une classe correspond à un élément A_i. Alors, la probabilité globale du dommage correspondant à chacune des classes sera :

$$\mathcal{P}[Mortel] = \mathcal{P}[\bigcup_{i=1}^{n} P_i^+] \quad \text{(Éq. 5)}$$

où P_i^+ représente une règle de la classe « Mortel ».

$\mathcal{P}[Non\ mortel] = \mathcal{P}[\bigcup_{i=1}^{n} P_i^-]$ (Éq. 6)

où P_i^- représente une règle de la classe « Non mortel ».

Ainsi : $\mathcal{P}[Mortel] = \mathcal{P}[P_1^+ \cup P_2^+ \cup P_3^+ \cup P_4^+ \cup P_5^+ \cup P_6^+ \cup P_7^+]$

$$\mathcal{P}[Non\ mortel] = \mathcal{P}[P_1^- \cup P_2^- \cup P_3^- \cup P_4^- \cup P_5^- \cup P_6^- \cup P_7^- \cup P_8^- \cup P_9^- \cup P_{10}^- \cup P_{11}^- \cup P_{12}^-]$$

Lors du calcul de probabilité des intersections issues de la formule de Poincaré, une attention est portée aux cas particuliers suivants :

- les règles imbriquées dans d'autres. Cela occasionne une intersection avec inclusion. Par exemple, P_2^+ est incluse dans P_6^+. Alors, au moment de calculer la probabilité de l'intersection entre ces deux règles dans la formule de Poincaré, on aura : $\mathcal{P}[P_2^+ \cap P_6^+] = \mathcal{P}[P_2^+]$ et non $\mathcal{P}[P_2^+ \cap P_6^+] = \mathcal{P}[P_2^+] \times \mathcal{P}[P_6^+]$;

- les règles ayant des intersections vides entre elles. Par exemple : P_1^+ et P_2^+ ne peuvent partager d'intersection puisque leurs conditions ($E > 4{,}5$) et ($E \leq 3{,}5$) sont contradictoires. La probabilité de leur intersection sera donc nulle. Il en va de même pour toute intersection d'une combinaison de règles comprenant P_1^+ et P_2^+.

Dans le cas des intersections non vides et dépourvues d'inclusion, la probabilité de l'intersection est le produit des probabilités de chaque règle. Par exemple, $\mathcal{P}[P_1^+ \cap P_7^+] = \mathcal{P}[P_1^+] \times \mathcal{P}[P_7^+]$.

Ainsi, $\mathcal{P}[Mortel] \approx 0{,}979$.

En procédant de même pour la classe "Non mortel", on obtient : $\mathcal{P}[Non\ mortel] \approx 0{,}119$.

Tout compte fait, la somme des probabilités des deux classes en considérant toutes les règles générées vaut 1,099 > 1 (tous les calculs relatifs à la formule de Poincaré ont été effectués avec *Excel* pour tout le processus par essai-erreur). La complémentarité des probabilités n'est donc pas respectée. Les tests suivants avec la table logique « ET » permettront d'élucider la situation.

3.5.4.2 Probabilité globale du dommage : comparaison entre la base de données d'accidents et la table logique « ET »

Prenons trois indicateurs A, B et C décrivant des accidents. Supposons que la population d'accidents est celle d'une table « ET » à trois variables (cf. tableau 3.21).

Tableau 3.21 : Table « ET » à trois variables (indicateurs) décrivant 8 types d'accidents possibles

Accident No.	**Classe**	***A***	***B***	***C***
1	0	0	0	0
2	0	0	0	1
3	0	0	1	0
4	0	0	1	1
5	0	1	0	0
6	0	1	0	1
7	0	1	1	0
8	1	1	1	1

L'équation logique décrivant le comportement d'une table « ET » à trois variables est :

$$Classe = A \cap B \cap C \qquad \text{(Éq. 7)}$$

En d'autres termes, la « *Classe* = 1 » est décrite par la règle : (*A* = 1) ET (*B* = 1) ET (*C* = 1). En revanche, la « *Classe* = 0 » est décrite par les trois règles suivantes : (*A* = 0), ou (*B* = 0), ou (*C* = 0). L'union de ces trois règles forme la négation de l'équation 7. En effet, la loi de De Morgan stipule que : $\overline{A \cap B \cap C} = \bar{A} \cup \bar{B} \cup \bar{C}$. Ainsi, les probabilités des classes « 1 » et « 0 » doivent être complémentaires.

Pour calculer la probabilité de chaque classe, nous utilisons les fonctions de masse $p_Y(y)$ pour *A*, *B*, puis *C* disponibles au tableau 3.22.

Tableau 3.22 : Fonctions de masse associées aux indicateurs *A*, *B* et *C* de la table « ET »

	y* = *a		***y* = *b***		***y* = *c***	
Valeurs possibles de « *y* »	0	1	0	1	0	1
$p_Y(y)$	$^4/_8$	$^4/_8$	$^4/_8$	$^4/_8$	$^4/_8$	$^4/_8$
	0,50	0,50	0,50	0,50	0,50	0,50

En calculant la probabilité de chaque classe à partir des fonctions de masse et en supposant l'indépendance des indicateurs, nous avons :

$\mathcal{P}[Classe = 1] = \mathcal{P}[(A = 1) \cap (B = 1) \cap (C = 1)]$

$= \mathcal{P}[A = 1] \times \mathcal{P}[B = 1] \times \mathcal{P}[C = 1]$

$\mathcal{P}[Classe = 0] = \mathcal{P}[(A = 0) \cup (B = 0) \cup (C = 0)]$

où $\mathcal{P}[A = 1] = \mathcal{P}[B = 1] = \mathcal{P}[C = 1] = \mathcal{P}[A = 0] = \mathcal{P}[B = 0] = \mathcal{P}[C = 0] = 0{,}5$ d'après le tableau 3.23.

Par la formule de Poincaré, on obtient :

$\mathcal{P}[Classe = 0] = \mathcal{P}[A = 0] + \mathcal{P}[B = 0] + \mathcal{P}[C = 0] - \mathcal{P}[A = 0] \times \mathcal{P}[B = 0] - \mathcal{P}[A = 0] \times \mathcal{P}[C = 0] -$
$\mathcal{P}[B = 0] \times \mathcal{P}[C = 0] + \mathcal{P}[A = 0] \times \mathcal{P}[B = 0] \times \mathcal{P}[C = 0]$

Alors :

$\mathcal{P}[Classe = 1] = 0{,}125$

$\mathcal{P}[Classe = 0] = 0{,}875$

⎫ Leur somme vaut 1 → Probabilités **complémentaires**.

Si *LAD-WEKA* génère des règles pour la table « ET » (cf. tableau 3.23), la somme des probabilités des deux classes donne une valeur différente de 1, en considérant toutes les règles générées. En effet, contrairement au cas de l'équation 7, l'union des règles générées par le logiciel pour la « *Classe* 0 » n'est pas le complément de la règle générée pour la classe « 1 ». Alors, la loi de De Morgan n'est pas vérifiable pour ces règles-là. Les calculs ci-après montrent ce propos.

Tableau 3.23 : Règles générées par *LAD-WEKA* à partir de la table « ET »

	Accidents couverts
« *Classe* = 1 »	
$P_1^+ = (A = 1)$ ET $(B = 1)$ ET $(C = 1)$	8
« *Classe* = 0 »	
$P_1^- = (B = 0)$	1 2 5 6
$P_2^- = (C = 0)$	1 3 5 7
$P_3^- = (A = 0)$	1 2 3 4
$P_4^- = (A = 1)$ ET $(C = 0)$	5 7
$P_5^- = (A = 1)$ ET $(B = 0)$ ET $(C = 1)$	6
$P_6^- = (A = 1)$ ET $(B = 0)$	5 6

En calculant la probabilité de chaque classe à partir des fonctions de masse tout en considérant les conditions sur les indicateurs groupeurs et leur indépendance, on obtient :

$$\mathcal{P}[Classe = 1] = \mathcal{P}[P_1^+]$$

$$= \mathcal{P}[(A = 1) \cap (B = 1) \cap (C = 1)]$$

$$= \mathcal{P}[A = 1] \times \mathcal{P}[B = 1] \times \mathcal{P}[C = 1]$$

$$\mathcal{P}[Classe = 0] = \mathcal{P}[P_1^- \cup P_2^- \cup P_3^- \cup P_4^- \cup P_5^- \cup P_6^-]$$

$$= \mathcal{P}[(B = 0) \cup (C = 0) \cup (A = 0) \cup ((A = 1) \cap (C = 0)) \cup ((A = 1) \cap (B = 0) \cap (C = 1)) \cup ((A = 1) \cap (B = 0))]$$

Encore une fois, la formule de Poincaré permet de calculer la probabilité de l'union des règles de la classe « 0 ».

Alors :

$\mathcal{P}[Classe = 1] = 0{,}125$

$\mathcal{P}[Classe = 0] \approx 0{,}438$

} Leur somme vaut 0,563 < 1 → Probabilités **non complémentaires**.

La somme des probabilités des classes « Mortel » et « Non mortel », ainsi que celle des classes « 0 » et « 1 » de la table « ET » sont ≠ 1. Cette constatation montre que considérer toutes les règles générées par *LAD-WEKA* pour la classe d'une base de données ne permet pas nécessairement d'estimer la probabilité de la classe. Pour y arriver, il faudrait sélectionner uniquement l'échantillon de règles qui permettraient d'atteindre la complémentarité entre les probabilités des classes, comme ce fut le cas pour la table « ET » avec son équation logique. D'après les résultats pour la table « ET », cela pourrait être possible en éliminant chaque règle dont tous les accidents couverts (cf. les accidents biffés au tableau 3.24 et au tableau 3.25) sont également caractérisés par une autre règle partageant au moins une même condition avec la règle précédente.

3.5.4.3 Probabilité globale du dommage en considérant le critère du partage de conditions

En appliquant ce critère aux règles générées à partir de la base de données « MELITO » (cf. tableau 3.24 et tableau 3.25), les règles suivantes : P_1^+, P_4^+, P_5^+ et P_6^+ sont retenues pour calculer la probabilité de la classe « Mortel ». Puis, les règles P_1^-, P_3^- et P_7^- sont retenues pour calculer la probabilité de la classe « Non mortel ».

Tableau 3.24 : Choix des règles pour le calcul de probabilité du dommage « Mortel »

Règles obtenues avec *LAD-WEKA* pour la classe « Mortel »		**Accidents couverts**
P_1^+	**(*E* > 4,5)**	**10**
~~P_2^+~~	~~(*E* ≤ 3,5) ET (*O* ≤ 2,5)~~	~~1 2 3 4 5 7 13 16 17 20 22~~
~~P_3^+~~	~~(*E* ≤ 3,5) ET (*I* ≤ 2,5)~~	~~1 2 5 8 9 18 22 23~~
P_4^+	**(*M* ≤ 1,5) ET (*E* ≤ 3,5)**	**3 5 7 9 13 16 17 18 22 23**
P_5^+	**(*E* ≤ 4,5) ET (*I* ≤ 2,5)**	**1 2 5 6 8 9 12 14 18 22 23**
P_6^+	**(*E* ≤ 4,5) ET (*O* ≤ 2,5)**	**1 2 3 4 5 7 12 13 16 17 20 22**
~~P_7^+~~	~~(*I* ≤ 2,5) ET (*O* > 2,5)~~	~~6 8 9 14 18 23~~

Tableau 3.25 : Choix des règles pour le calcul de probabilité du dommage « Non mortel »

Règles obtenues avec *LAD-WEKA* pour la classe « Non mortel »		**Accidents couverts**
P_1^-	**($I > 2,5$) ET ($E = 4$)**	**15 21**
~~P_2^-~~	~~($M \leq 1,5$) ET ($E \neq 4$) ET ($I \leq 2,5$) ET ($O \leq 2,5$)~~	~~19~~
P_3^-	**($E \leq 4,5$) ET ($I > 2,5$) ET ($O > 2,5$)**	**11 21**
~~P_4^-~~	~~($M \leq 1,5$) ET ($E \leq 4,5$) ET ($I > 2,5$) ET ($O > 2,5$)~~	~~21~~
~~P_5^-~~	~~($M \leq 1,5$) ET ($E \neq 4$) ET ($O \leq 2,5$)~~	~~15 19~~
~~P_6^-~~	~~($E > 3,5$) ET ($I > 2,5$) ET ($O \leq 2,5$)~~	~~15~~
P_7^-	**($M \leq 1,5$) ET ($E > 3,5$) ET ($O \leq 2,5$)**	**15 19**
~~P_8^-~~	~~($E \neq 4$) ET ($I > 2,5$) ET ($O > 2,5$)~~	~~21~~
~~P_9^-~~	~~($M > 1,5$) ET ($E \leq 3,5$) ET ($I > 2,5$) ET ($O > 2,5$)~~	~~11~~
~~P_{10}^-~~	~~($M > 1,5$) ET ($I > 2,5$) ET ($O > 2,5$)~~	~~11~~
~~P_{11}^-~~	~~($E \leq 3,5$) ET ($I > 2,5$) ET ($O > 2,5$)~~	~~11~~
~~P_{12}^-~~	~~($M \leq 1,5$) ET ($E > 3,5$) ET ($I > 2,5$) ET ($O \leq 2,5$)~~	~~15~~

Ainsi : $\mathcal{P}[Mortel] = \mathcal{P}[P_1^+ \cup P_4^+ \cup P_5^+ \cup P_6^+]$

$\mathcal{P}[Non\ mortel] = \mathcal{P}[P_1^- \cup P_3^- \cup P_7^-]$

Alors :

$\mathcal{P}[Mortel] \approx 0,935$

$\mathcal{P}[Non\ mortel] \approx 0,292$

} Leur somme vaut 1,227 ≠ 1→ Probabilités **non complémentaires**.

Ainsi, ce n'est pas la méthode de calcul appropriée pour estimer la probabilité d'une classe.

3.5.4.4 Probabilité globale du dommage en considérant le minimum de règles couvrant tous les accidents

Pour éviter des redondances entre les règles considérées pour le calcul, une autre tentative serait de considérer uniquement le nombre minimal de règles permettant de couvrir tous les accidents. Ainsi, pour la classe « Mortel », il s'agit des règles : P_1^+, P_6^+ et P_7^+ (cf. tableau 3.26). Quant à la classe « Non mortel », il s'agit de : P_3^- et P_7^- (cf. tableau 3.27).

Tableau 3.26 : Règles minimales couvrant les accidents de la classe « Mortel »

	Règles minimales couvrant les accidents										
	P_6^+	P_7^+	P_6^+	P_7^+	P_1^+	P_6^+	P_7^+	P_6^+	P_7^+	P_6^+	P_7^+
Les 19 accidents de la classe « Mortel »	1 2 3 4 5	6	7	8 9	10	12 13	14	16 17	18	20 22	23

Tableau 3.27 : Règles minimales couvrant les accidents de la classe «Non mortel »

	Règles minimales couvrant les accidents		
	P_3^-	P_7^-	P_3^-
Les 4 accidents de la classe « Non mortel »	11	15 19	21

Ainsi :

$\mathcal{P}[Mortel] = \mathcal{P}[P_1^+ \cup P_6^+ \cup P_7^+]$

$\mathcal{P}[Non\ mortel] = \mathcal{P}[P_3^- \cup P_7^-]$

Alors :

$\mathcal{P}[Mortel] \approx 0{,}821$

$\mathcal{P}[Non\ mortel] \approx 0{,}308$

Leur somme vaut 1,129 ≠ 1 → Probabilités **non complémentaires**.

3.5.4.5 Issue des calculs de probabilités

Comme mentionné plus tôt, les classes « Mortel » et « Non mortel » ont été choisies pour avoir un espace échantillon d'accidents aux événements complémentaires. Cependant, les règles générées par *LAD-WEKA* pour une classe ne sont pas complémentaires par rapport à celles de l'autre classe (c.-à-d. l'ensemble des règles d'une classe ne correspond pas à la négation des règles de l'autre classe). La loi de De Morgan n'est donc pas vérifiée. Cela expliquerait l'impossibilité d'aboutir à des probabilités complémentaires pour ces deux classes d'accidents. En revanche, la probabilité de chaque règle a pu être obtenue en supposant que les indicateurs groupeurs étaient indépendants entre eux. Chacune de ces probabilités représente la probabilité

du dommage spécifique à une situation dangereuse représentée par une règle. Jumelées à la couverture de chaque règle, ces probabilités permettent de hiérarchiser les situations dangereuses et ainsi d'établir les priorités en matière de réduction du risque.

La base de données « MELITO » ne comprend pas tous les cas possibles d'accidents liés à un convoyeur à courroie. Si elle comprenait tous les cas possibles, comme pour la table logique « ET » amenant à une sortie « 1 » ou « 0 », il aurait été possible de déduire avec l'ALD la fonction booléenne caractérisant chaque classe. Les fonctions booléennes induiraient des probabilités complémentaires comme c'est le cas pour la somme des probabilités des sorties « 1 » et « 0 » de la table « ET ». Étant donné que les règles générées par l'ALD, pour une base de données au nombre d'observations incomplet, forment une fonction booléenne partiellement définie (Crama et al., 1988) plutôt qu'une fonction booléenne et comme le calcul des probabilités dépend des règles, il semble normal d'avoir obtenu une somme des probabilités des deux classes, différentes de 1. Il est donc envisageable que la complémentarité pour les classes caractérisées par ces règles ne soit pas atteinte. L'expression « fonction booléenne partiellement définie » signifie que l'ensemble des règles obtenues ne décrit pas tous les cas possibles d'accidents. Comme on ne peut vérifier une méthode en se basant uniquement sur un cas partiellement connu, il a été proposé de tester le calcul de probabilité globale du dommage à l'aide d'une table « ET » à trois variables ayant tous les cas possibles.

CHAPITRE 4 DISCUSSION GÉNÉRALE

4.1 Récapitulatif

Cette thèse par articles vise à proposer une démarche d'identification et d'estimation efficace des risques liés aux machines et à leurs environnements physique et organisationnel. La démarche doit faciliter le suivi de ces risques en milieu de travail. Bien que la démarche soit appliquée à des convoyeurs à courroie, elle est transposable à d'autres types de machines.

Dans cette optique, l'article 1 propose une méthode dynamique d'aide à la décision en gestion du risque lié aux machines. Cette méthode sert à la conception d'un outil dynamique d'aide à la décision. L'aide à la décision se base principalement sur l'identification du risque et l'estimation quantitative du risque qui peuvent être mises à jour après la survenue d'un accident ou la prise en compte d'un moyen de réduction du risque. Dans l'article 1, l'estimation quantitative du risque proposée est probabiliste et bayésienne. Elle se base sur la combinaison d'avis d'experts à propos des fréquences de causes accidentelles prépondérantes signalées par les règles déduites d'une base de données d'accidents et de presque accidents (cependant, un peu plus tard dans l'étude, par souci de respecter l'échéance du projet, une approche autre que bayésienne présentée en 3.5.3 a été envisagée pour l'estimation quantitative du risque). L'article 1 propose l'algorithme ALD pour générer les règles. La performance de cet algorithme dans le domaine biomédical (ex., Alexe et al., 2006; Brauner et al., 2007; Hammer et Bonates, 2005), les finances (ex., Cavali. et Moriggia, 2002) et la maintenance (ex., Bennane et Yacout, 2012) porte à le suggérer pour générer les règles. La méthode proposée dans l'article 1 se base sur une recension exhaustive des écrits. L'utilité de la méthode y est présentée. Toutefois, l'application de la méthode était prévue pour les articles subséquents de la thèse.

La première étape de la méthode proposée dans l'article 1 porte sur l'extraction de connaissances sous forme de règles générées par l'ALD, depuis une base de données d'accidents et de presque accidents. Les entreprises et organismes qui compilent les accidents et passés-proches peuvent utiliser cette approche. Comme les passés-proches ou incidents n'étaient pas disponibles, l'article 2 (puis l'article 3) n'a considéré que des données d'accidents. L'objectif premier de l'article 2 était de démontrer l'applicabilité de l'algorithme ALD à une base de données très restreinte. Disposer d'un échantillon restreint d'accidents est une situation réaliste dans le contexte

industriel. Les accidents étant des événements rares, les entreprises ont accès à un nombre limité de rapports d'accidents les concernant.

Lors de l'analyse des résultats de *cbmLAD*, les valeurs de couverture des règles générées ont montré le potentiel de l'ALD à servir à l'estimation du risque. Cette constatation est à l'origine du second objectif de l'article 2 : montrer que cet algorithme est utile pour prévenir les accidents liés aux machines. Ainsi, l'idée d'exploiter l'ALD à des fins d'estimation quantitative du risque a émergé. C'est ainsi que dans l'article 2, le risque est estimé en se basant sur la fréquence des caractéristiques accidentelles qui ressortaient lors de la phase de validation de l'ALD. Les caractéristiques accidentelles désignent les valeurs d'un indicateur dans une règle générée. La méthode d'estimation du risque de l'article 2 ne s'appuie pas sur des probabilités d'occurrence des caractéristiques accidentelles. Elle se base uniquement sur la fréquence d'apparition de ces caractéristiques et la couverture associée aux règles. L'application de l'ALD à la sécurité au travail, et plus précisément à la sécurité des machines, est novatrice.

La deuxième étape de la méthode proposée dans l'article 1 portait sur l'estimation de la probabilité 1) du dommage et 2) de chaque scénario accidentel représenté par une règle, en adoptant une approche bayésienne. En raison de contraintes temporelles, les étapes 2 et 3 de la méthode de l'article 1 n'ont pu être réalisées comme prévu. Une alternative à l'estimation quantitative annoncée dans l'article 1 a dû être adoptée pour l'article 3. Cette alternative détaillée à la section 3.5.3 consiste à estimer la probabilité de chaque scénario (c.-à-d., de chaque règle) à partir des probabilités des conditions formant la règle. La probabilité d'une condition dépend de la fonction de masse associée à son indicateur.

L'estimation quantitative du risque, présentée à l'article 2, se base sur les couvertures des règles et les fréquences d'apparition de leurs conditions. Cette estimation du risque est simple et efficace d'application. Toutefois, l'estimation quantitative du risque dans l'article 3 est plus précise, mais plus complexe à mettre en œuvre en raison des connaissances requises en calculs probabilistes. Par exemple, avec la méthode d'estimation du risque de l'article 2, si deux règles avaient la même couverture, il serait difficile d'établir une hiérarchie entre elles, d'où le manque de précision. Ainsi, avec la méthode d'estimation du risque de l'article 3, si des règles avaient la même couverture, leur différence au niveau de leurs probabilités permettrait de les hiérarchiser.

Pour l'article 3, un autre logiciel a été utilisé : *LAD-WEKA*. Utiliser deux logiciels différents pour l'ALD a montré que l'approche algorithmique de *cbmLAD* est bien plus optimisée que celle de *LAD-WEKA*. En effet, le rapport du nombre de règles générées versus leur précision de classification est plus satisfaisant pour le premier que pour le second. En outre, *cbmLAD* réalise un paramétrage automatique selon les données analysées, contrairement à *LAD-WEKA* qui n'offre à l'utilisateur que la possibilité d'un paramétrage manuel. À partir de ce constat, il est recommandé au préventionniste qui souhaite utiliser l'ALD pour de la prévention d'accidents, d'opter pour un logiciel optimisé capable de générer le moins de règles possibles (donc capable de générer l'essentiel de la connaissance), tout en fournissant une précision de classification adéquate. En revanche, si l'utilisation de l'ALD a pour but de présenter une méthode dont la faisabilité est indépendante de la performance de cet algorithme, comme dans l'article 3 de cette thèse, *LAD-WEKA* peut être utilisé. Une fois la faisabilité de la méthode démontrée, il serait souhaitable d'utiliser un logiciel plus performant que *LAD-WEKA* afin de générer des règles optimisées pour prévenir les accidents.

En somme, l'article 1 a proposé une méthode dans l'optique d'atteindre le but de la thèse. L'article 2 visait principalement à montrer la faisabilité de l'étape 1 de la méthode suggérée dans l'article 1. En effet, les aptitudes de l'ALD en fouille de données pour un nombre restreint d'observations étaient inconnues. La précision adéquate de classification (72% - 74%) obtenue avec l'ALD dans l'article 2 permet de confirmer l'aptitude de cet algorithme pour l'identification des facteurs de risque à travers les règles qu'il génère. La nécessité de s'adapter aux contraintes temporelles explique la démarche alternative employée pour cette thèse. Les points C.1 à D de la figure 3.1 constituent cette démarche. L'article 3 traite de l'entièreté de cette démarche pour identifier et estimer efficacement des risques de manière à en faciliter leur suivi en milieu de travail, tout en considérant les aspects « machines », « environnement physique et organisationnel ». La prise en compte de ces aspects constituant le contexte accidentel s'est concrétisée en considérant les indicateurs appartenant au moment (*M*) de l'accident, à l'équipement (*E*) impliqué, au lieu (*L*) de l'accident, à l'individu (*I*) impliqué, la tâche (*T*) en cours et au système organisationnel (*O*). Pour donner une meilleure vue d'ensemble, la base de données a été formatée selon le concept MELITO. Notez que cette étape de formatage n'est pas obligatoire, tout dépendra des besoins du préventionniste.

4.2 Confirmation de la première hypothèse de recherche

Comme mentionné à l'introduction, pour identifier les risques liés aux machines, les préventionnistes ont accès, entre autres, à des normes, des guides et des rapports d'enquêtes d'accidents. Ces rapports peuvent provenir d'une banque de documents comme celle du Centre de documentation de la CNESST. Les préventionnistes eux-mêmes peuvent avoir rédigé des rapports à la suite d'une enquête d'accident ou de passé-proche en interne. Dans le cas des rapports, la connaissance est consultée et interprétée par le préventionniste qui les lit ou les a rédigés. Nombreux sont les rapports envoyés aux oubliettes une fois lus (Kletz, 1993). Or, ces rapports représentent une mine d'information instructive pour la prévention d'accidents. Afin d'exploiter cette mine à bon escient, il serait pertinent d'en extraire l'essentiel de l'information contenue dans les rapports d'accidents (Lindberg et al., 2010). Vu ces constats, il a été supposé qu'utiliser le REX dynamique basé sur l'ALD permette d'identifier efficacement les facteurs de risque liés aux machines et d'en suivre l'évolution (c'est la première hypothèse de recherche). Rappelons que le REX dynamique consiste à apprendre du passé par l'inférence de connaissances à partir d'événements enregistrés dans une base de données, grâce à de la remontée d'information. L'inférence est réalisée par un algorithme de fouille de données.

Les études en sécurité du travail (Cheng et al., 2012; Verma et al., 2014; Silva et al., 2012) ou plus précisément en sécurité des grues sur les chantiers de construction (Raviv et al., 2017) qui utilisent la fouille de données à des fins de prévention d'accidents traitent de cas ayant minimalement des centaines d'accidents. De plus, les trois études en sécurité au travail susmentionnées utilisent des techniques de fouille de données comme les règles d'association ou les arbres de décisions. Or, ces techniques deviennent efficaces s'il existe des ensembles suffisamment fréquents dans les données pour caractériser des classes d'accidents, par exemple. Quand l'échantillon se raréfie, les chances d'obtenir des ensembles fréquents s'amenuisent. Il devient donc impossible d'utiliser ces techniques. C'est ce qui explique le choix de l'algorithme de fouille de données ALD. L'essence même de cet algorithme est de trouver des différences dans les données de classes diverses afin de caractériser ces classes et ce, peu importe le nombre d'accidents différents disponibles. L'ALD ne sert donc pas à trouver des ensembles fréquents de caractéristiques pour définir des classes diverses.

Quant à l'étude de Raviv et al. (2017), elle emploie un apprentissage non supervisé avec l'algorithme *k-means* pour former des regroupements d'accidents survenus sur des grues et connaître les caractéristiques de ces regroupements. En revanche, dans la présente thèse, la gestion du risque proposée se fait par rapport à un type d'accident et comme les types (classes) d'accidents sont connus d'avance, un apprentissage supervisé est requis. Cela explique aussi le choix de l'ALD comme algorithme pour l'apprentissage supervisé (l'encadré à la fin de la section 3.5.1 définit ces deux types d'apprentissage).

Finalement, Aneziris et al. (2013) ont généré manuellement un nœud-papillon. Ce dernier consiste en un schéma illustrant des liens de cause à effet. Au centre du schéma, on retrouve un événement indésirable (ex., contact avec une pièce en mouvement). La partie gauche du nœud est un arbre de défaillances, tandis que sa partie droite est un arbre d'événements. L'événement redouté est engendré par l'arbre de défaillances. Si l'événement redouté n'est pas contrôlé, il peut dégénérer en plusieurs événements qui affecteront la population, l'environnement ou le travailleur, d'où l'arbre d'événements à droite. Aneziris et al. (2013) ont construit ce nœud-papillon à partir de causes génériques tirées de leur propre analyse de rapports d'enquêtes d'accidents liés à des pièces en mouvement sur des machines. Ce nœud-papillon représente la connaissance générique tirée de 3000 accidents. Contrairement à l'apprentissage automatique que propose cette thèse, l'apprentissage d'Aneziris et al. (2013) était manuel et le nœud-papillon proposé était immuable du côté de l'utilisateur. Seul le concepteur pouvait mettre à jour les causes d'accidents. La méthode proposée dans cette thèse permet à l'utilisateur (le préventionniste) de mettre à jour, à sa guise, les causes et facteurs d'accidents.

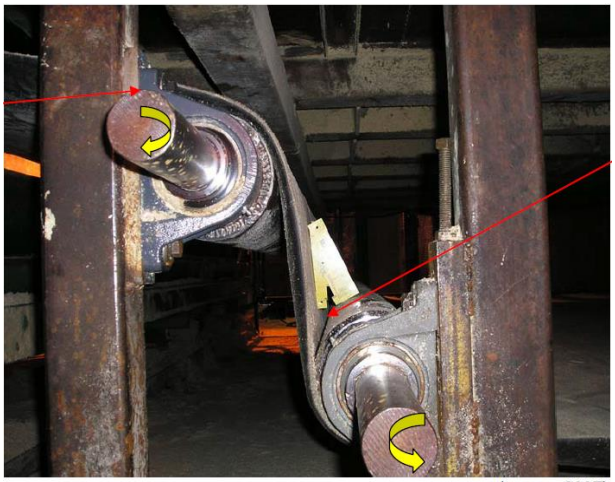
4.2.1 Première hypothèse de recherche : contributions à l'avancement des connaissances

Le REX dynamique basé sur l'ALD contribue à identifier efficacement les facteurs de risque puisque les règles que génère l'algorithme permettent d'apprendre du passé, en transmettant l'essentiel de la connaissance tirée des rapports d'enquêtes d'accidents et ce, peu importe leur nombre. La connaissance inférée et transmise par l'ALD permet de prédire, avec une précision de classification adéquate, les types d'accidents possibles (ex., accident de maintenance ou accident de production, accident grave non mortel ou accident mortel). En d'autres termes, si la

combinaison d'indicateurs d'accident composant une règle est observée en entreprise, l'événement redouté se trame. À titre illustratif, la règle P_2^+ caractérisant 70 % des accidents de maintenance analysés (cf. tableau 3.8) aurait permis, lors d'une inspection des lieux de travail d'une scierie décrite dans Brulotte et Roberge (2006) , de constater que les principales conditions pouvant induire un accident de maintenance étaient réunies sur le convoyeur de l'entreprise et présageaient un tel événement. En effet, cette règle regroupe les principales conditions des accidents de maintenance de 2002 et 2005 décrits dans (Brulotte et Roberge, 2006) :

- facteur de risque : encombrement des lieux de travail. La sciure de bois (cf. photo A de la figure 4.1) bloquait l'accès au convoyeur. Cet empêchement portait le travailleur à emprunter un raccourci en passant, accroupi, sur le brin inférieur de la courroie du convoyeur;
- cause directe : angle rentrant accessible. La photo B de la figure 4.1 illustre deux angles rentrants dépourvus de moyen de protection pour sécuriser le convoyeur. Un rouleau en rotation et le brin inférieur de la courroie créent chaque angle rentrant;
- cause indirecte : partie du corps dans l'angle rentrant.

(source : CSST) (A)



(source : CSST) (B)

Figure 4.1 : Principaux facteur de risque et causes des accidents de 2002 et 2005 survenus dans une même scierie, sur le même convoyeur (photo A : la flèche de gauche montre l'accès normal à la salle du convoyeur, tandis que celle de droite pointe le tas de sciure de bois; photo B : les deux flèches rouges pointent vers des angles rentrants) (Brulotte et Roberge, 2006)

Agir sur chacune des causes d'accident et facteur de risque aurait permis d'éviter l'accident de 2005 survenu pour les mêmes raisons que celui de 2002.

En plus d'aider à prédire les accidents dans une optique de prévention, la connaissance générée par l'ALD permet également :

- aux organismes en prévention de suivre l'évolution des facteurs de risque et causes possibles d'accidents afin de déceler si un accident est imminent et y remédier avant qu'il ne survienne. Ce suivi peut se faire lors d'audits, par exemple;
- aux concepteurs et aux intégrateurs de machines de profiter d'un retour d'expérience. D'ailleurs, la norme générale ISO 12100:2010 en conception de machines prévoit le retour d'expérience comme moyen d'améliorer en continu la sécurité des machines. Les facteurs de risques et causes possibles composant les règles sous forme de conditions, permettent au concepteur de penser au-delà de la machine, en considérant l'environnement (organisationnel et physique), les personnes impliquées (surtout les travailleurs), ainsi que la tâche pour laquelle la machine doit être conçue.

La prépondérance des indicateurs relatifs à l'équipement dans l'explication des accidents analysés a été montrée en fin de section 3.5.3. Ce qui conforte la prescription normative de l'ISO 12100:2010 qui prône, en priorité, les moyens de prévention liés à l'équipement pour réduire efficacement le risque : la prévention intrinsèque suivie des moyens de protection (protecteurs et dispositifs de protection) si le risque résiduel n'est pas suffisamment réduit. En outre, l'article 3 fait ressortir, en sa section 8.1, que les indicateurs de type « Équipement » sont suivis par les indicateurs de types « Individu » puis « Organisation » en matière d'importance dans la survenue des accidents analysés. Ce résultat rappelle celui de Raviv et al. (2017) qui en sont arrivés à la conclusion que les facteurs techniques sont les plus dangereux dans la survenue des accidents analysés impliquant des grues, en plus d'être interreliés aux facteurs humains. Cela montre la nécessité pour le concepteur et l'intégrateur de considérer l'utilisateur de la machine, ses bons et mauvais usages raisonnablement prévisibles de la machine, ainsi que l'environnement dans lequel elle sera utilisée.

Enfin, le REX dynamique basé sur l'ALD est un outil adapté pour de la prévention ciblée puisque l'essence même de cet algorithme est de trouver les différences entre des classes d'événements. Par exemple, selon le type de tâche, les moyens de réduction du risque ne seront pas

nécessairement les mêmes. Ainsi, un protecteur installé sur une machine pour protéger le travailleur lors de la production peut devenir gênant lors d'une intervention de maintenance dans le mécanisme duquel le travailleur était protégé en production. Alors, un autre moyen de réduction du risque doit être utilisé en maintenance, tel que le cadenassage, pour éviter le démarrage intempestif du mécanisme alors que le travailleur a la main dedans par exemple. De plus, gérer les risques en faisant de la prévention ciblée est profitable aux entreprises ayant peu de ressources pour gérer tous les risques « machines » menaçant la sécurité des travailleurs. En effet, par la prévention ciblée, l'ALD communique aux préventionnistes de ces entreprises les causes d'accidents et facteurs de risques essentiels à la survenue des accidents analysés.

La méthode d'identification des risques basée sur le REX dynamique est de nature itérative. Lorsqu'un nouvel accident d'un certain type surviendra, il pourra être ajouté à la base de données. L'inférence de connaissances (règles) pourra être mise à jour automatiquement par l'utilisateur, en lançant à nouveau l'algorithme ALD.

Toutes ces contributions confirment la première hypothèse de recherche, conformément au sens de l'expression « identifier efficacement les facteurs de risque » donné à la section 2.3.

4.3 Confirmation de la seconde hypothèse de recherche

Comme expliqué au chapitre 2, des outils matriciels qualitatifs sont couramment utilisés pour estimer le risque lié aux machines. L'aspect qualitatif rend l'estimation du risque subjective. Par ricochet, les décisions en gestion du risque découlant de cette subjectivité peuvent être inadaptées. Pour gérer efficacement le risque Cox (2008), Duijm (2015), Hubbard et Evans (2010) privilégient une estimation quantitative du risque plutôt que qualitative. L'aspect quantitatif, portant principalement sur l'estimation de probabilités (ex., probabilité d'occurrence du dommage ou de l'accident, probabilité d'occurrence de l'événement dangereux ayant conduit au dommage), apporte de l'objectivité à l'estimation du risque.

De plus, pour améliorer l'adéquation entre les moyens de réduction du risque et le risque lui-même, la mise à jour du risque est nécessaire en actualisant l'estimation du risque en plus de la mise à jour de l'identification du risque. Autrement, les moyens de réduction du risque mis en

œuvre ou prévus deviendront désuets et n'arriveront pas à assurer la sécurité de l'utilisateur de la machine puisque le risque aura évolué entretemps.

Face à de tels constats, il a été supposé qu'estimer la probabilité du dommage à partir des facteurs de risque identifiés via le REX dynamique basé sur l'ALD, permet d'estimer efficacement les risques et prioriser les mesures de réduction du risque (c'est la seconde hypothèse de recherche).

Bien que le nœud-papillon d'Aneziris et al. (2013) empêche l'utilisateur de mettre à jour les causes génériques d'accidents, il rend possible la mise à jour des fréquences de causes d'accidents et de la probabilité du dommage. Toutefois, cette mise à jour est incomplète, car les indicateurs constituant le risque, tout comme la combinaison de ces indicateurs, pourraient avoir changé dans le temps, ce qui aurait pour effet de modifier autrement la probabilité du dommage.

4.3.1 Seconde hypothèse de recherche : contributions à l'avancement des connaissances

Estimer la probabilité du dommage à partir de facteurs de risque identifiés par les règles contribue à estimer efficacement les risques liés aux machines, puisque la probabilité estimée pour chaque scénario d'accident représenté par une règle permet d'ordonner ces scénarios selon leur importance (cf. tableau 3.19 et 3.20). Cela apporte de l'objectivité dans la hiérarchisation des risques, et ainsi de l'objectivité dans la priorisation des mesures de réduction du risque. Plus la probabilité d'une règle est élevée, plus le scénario d'accident qu'elle représente est important. Ainsi, le risque associé à ce scénario d'accident doit être réduit avant celui des autres scénarios. La réduction du risque consiste à mettre en œuvre des moyens pour réduire la dangerosité des indicateurs (facteurs de risque et causes potentielles d'accident) composant chaque règle, en commençant par l'indicateur le plus récurrent dans les règles. Pour évaluer l'efficacité d'un moyen de réduction du risque, il faut vérifier s'il réduit au moins un des paramètres principaux du risque, soit : la gravité ou la probabilité du dommage. La probabilité du dommage estimée sert de référentiel de comparaison pour évaluer l'effet d'un moyen de réduction du risque. C'est aussi un référentiel pour comparer la probabilité après un nouvel accident à la probabilité initiale.

Combiner l'estimation quantitative du risque au REX dynamique permet également de suivre l'évolution du risque par la possibilité de mettre à jour les probabilités après l'actualisation de

l'identification du risque (c.-à-d., actualisation des facteurs de risques et causes d'accidents après l'enregistrement d'un nouvel accident dans la base de données, par exemple). Contrairement au nœud-papillon d'Anezeris et al. (2013), la démarche proposée dans cette thèse (étapes C.1 à D de la figure 3.1) offre la possibilité à l'utilisateur de mettre à jour les scénarios d'accident (règles), mais aussi les probabilités associées à ces scénarios.

La méthode proposée pour l'estimation de la probabilité du dommage associé à une situation dangereuse ne se cantonne pas à des règles générées par l'ALD, ni au calcul de la probabilité du dommage. Premièrement, elle est aussi applicable à des règles générées par des algorithmes d'inférence de connaissance autres que l'ALD. Deuxièmement, elle peut aussi servir à calculer la probabilité de l'événement dangereux. Ce dernier est un « événement susceptible de causer un dommage » (ISO, 2010). L'événement dangereux est considéré par la norme ISO 12100:2010 comme un sous-élément de la probabilité du dommage. Dans ce cas, chaque classe de la base de données dont il faudrait générer les règles représenterait un type d'événement dangereux (ex., démarrage intempestif). Les indicateurs de la base de données deviendraient les causes et facteurs (ex., l'étape de purge omise dans le processus de cadenassage) contribuant au type d'événement dangereux. La valeur de la probabilité de l'événement dangereux pourrait très bien être intégrée à un outil d'estimation du risque où la probabilité de l'événement dangereux mérite d'être connue. Ce serait un moyen de réduire la subjectivité d'outils d'estimation qualitative du risque. Par exemple, en affectant des valeurs calculées et non arbitraires aux niveaux qualitatifs de la probabilité de l'événement dangereux. À titre illustratif, la troisième colonne de l'outil de la figure 4.2 offre trois niveaux qualitatifs de « Probabilité d'occurrence de l'événement dangereux » :

- « ***O1*** *Très faible : technologie stable, éprouvée et reconnue pour les applications de sécurité*
- ***O2*** *Faible : événement relié à une défaillance technique de probabilité supérieure ou égale à* 10^{-5} *bris/heure (1 bris/100 000 heures); ou bien, événement entraîné par l'action d'une personne qualifiée, expérimentée, formée, effectuant une tâche unique, etc.*
- ***O3*** *Élevée : événement relié à une défaillance technique de probabilité supérieure ou égale à* 10^{-3} *bris/heure (1 bris/1 000 heures); ou bien, événement entraîné par l'action d'une personne sans expérience ou formation particulière* » (Paques et al., 2004).

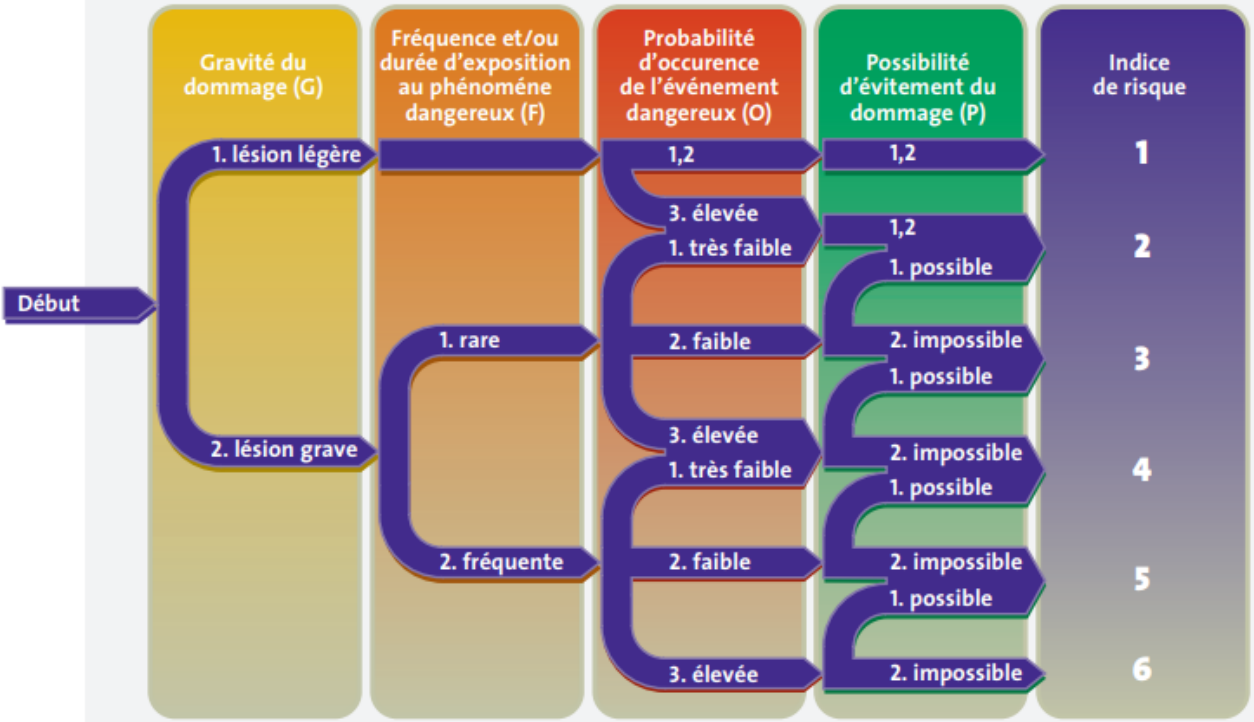


Figure 4.2 : Exemple d'outil d'estimation qualitative du risque (Paques et al., 2004)

Les niveaux O1 à O3 portent sur des événements dangereux d'origine soit technique, soit humaine. Si jamais l'origine de l'événement dangereux n'était pas technique, mais plutôt organisationnelle ou un mélange des trois, il aurait été difficile de sélectionner le niveau approprié pour la situation dangereuse analysée. En ayant des indicateurs techniques, humains et organisationnels décrivant trois types d'événements dangereux dans une base de données, par exemple : 1) démarrage intempestif; 2) démarrage inopiné par un tiers; 3) démarrage inopiné par la victime, il serait possible d'examiner, lors d'une inspection en milieu de travail, le type d'événement dangereux risquant de survenir, grâce aux règles caractérisant chaque type d'événement dangereux. Le type d'événement dangereux identifié pour la situation dangereuse analysée permet de savoir la probabilité d'occurrence lui correspondant. Si plus d'un type d'événement dangereux est attribuable à la situation dangereuse, le préventionniste devra sélectionner le type ayant la plus grande probabilité d'occurrence, pour ensuite faire son choix entre O1, O2 et O3. Cette contribution sert à pallier le fait que « l'évaluation de la probabilité est un aspect problématique de l'estimation du risque » (Gauthier et al., 2016). Aussi, l'exemple précédent montre comment on peut réduire la subjectivité de l'estimation du risque associé à une situation dangereuse. Cependant, l'amélioration, dans ce cas, est au niveau de la formulation d'un

paramètre. La subjectivité au niveau de la description qualitative des autres paramètres demeure, au même titre que la subjectivité rattachée à l'architecture de l'outil.

Avec l'identification du risque liée à la première hypothèse, l'estimation quantitative du risque rattachée à la seconde hypothèse constitue une démarche d'analyse du risque lié aux machines. Les règles générées et leurs probabilités sont facilement interprétables par des préventionnistes non-initiés. Cette démarche peut servir d'une première analyse du risque pour commencer la gestion du risque. Elle sert également à suivre l'évolution du risque, via un changement au niveau des probabilités ou des valeurs des indicateurs qui servent de tableau de bord au préventionniste.

Enfin, à chaque problème lié à la SST, son appréciation du risque. À chaque appréciation du risque, son outil, son modèle d'analyse du risque. Le modèle étant flexible, il permet à chaque entreprise d'adapter la méthode à sa réalité en choisissant les indicateurs décrivant cette réalité, ainsi que leurs valeurs possibles.

Toutes ces contributions confirment la seconde hypothèse de recherche, conformément au sens de l'expression « estimer efficacement les risques » donné à la section 2.3.

4.4 Limites et contraintes

Bien que l'étape non obligatoire du formatage des données traité dans l'article 3 ait pour avantage de donner une vue d'ensemble sur le contexte accidentel, l'apprentissage non supervisé (cf. la définition dans l'encadré de la section 3.5.1) employé pour le processus de fusion des indicateurs apporte de l'incertitude au modèle. En effet, en apprentissage non supervisé, c'est à l'utilisateur de juger du nombre de regroupements admissibles pour le cas à traiter.

Comme le montrent l'article 2 (prévention ciblée selon la tâche) et l'article 3 (prévention ciblée selon le type de dommage), le contenu de la base de données change en fonction du problème à résoudre. Certains indicateurs doivent être enlevés, d'autres doivent être ajoutés, les valeurs des indicateurs changent aussi (ex., dans l'article 2, les valeurs non binaires allaient du plus fréquent au moins fréquent; tandis que dans l'article 3, les valeurs, qu'elles soient binaires ou non, allaient du moins au plus dangereux de manière croissante). Le choix des indicateurs et de leurs valeurs impacte les règles représentant le risque, ainsi que le choix des moyens de réduction du risque. Pour éviter que ces derniers n'entravent la tâche des travailleurs, mieux vaut réaliser, avec tous

les intervenants, l'exercice d'analyse du risque depuis la définition du problème à résoudre, en passant par la conception de la base de données jusqu'à son exploitation.

La démarche d'analyse du risque proposée ne peut remplacer une analyse détaillée du risque au sens de l'ISO 12100:2010. La démarche proposée peut servir de démarrage dans un processus d'analyse du risque ou de « mieux que rien du tout » pour les entreprises qui ignorent comment réaliser une analyse du risque ou les entreprises qui trouvent qu'elles manquent de temps pour analyser le risque. Aussi, la démarche peut être utilisée après l'analyse du risque détaillée selon l'ISO 12100:2010, afin de suivre les facteurs clés du risque ou principales causes d'accidents, ainsi que les probabilités du dommage ou de l'événement dangereux. Quand les moyens de réduction du risque auront été installés ou lorsqu'un nouvel accident sera survenu, réutiliser la démarche pour mettre à jour le risque permet de faire un suivi du risque.

La mise à jour du risque ou le succès du REX dynamique dépend de la remontée d'information par les travailleurs qui sont les plus susceptibles de noter des changements relatifs à la machine ou son environnement. Seules une culture de sécurité et une confiance entre les travailleurs et leurs supérieurs pourront encourager la remontée d'information. Sans remontée d'information, aucun nouvel accident, ni incident, ni modification liée aux moyens de réduction du risque ne sera enregistré. Conséquemment, le portrait du risque n'est pas actualisé. Ainsi, sans remontée d'information, l'entreprise vit dans le passé et prend donc des décisions dépassées.

Les règles générées dans les articles 2 et 3 ne sont pas généralisables vu l'échantillon très restreint d'observations. En revanche, ces règles s'appliquent aux usines concernées par les accidents analysés. En outre, les accidents analysés sont issus de la base de données de la CNESST. Or, cette base ne comporte que des accidents graves et mortels. Les règles générées ne peuvent refléter les indicateurs associés à des dommages moins graves.

L'application de la démarche d'analyse du risque proposée est envisageable pour les grandes entreprises (ex., Hydro-Québec) ou les organismes de prévention afin qu'ils en fassent profiter les entreprises de leurs secteurs ou sous leur gouvernance (notamment les petites entreprises qui manquent de ressource). La démarche proposée exige d'être exécutée par un expert en processus d'extraction des connaissances et en calcul probabiliste. Cependant, une fois que les règles à générer et les probabilités associées deviennent disponibles, le préventionniste est en mesure de prendre les décisions quant aux moyens de réduction du risque à mettre en œuvre et dans quel

ordre. La présence de ce connaisseur est requise pour mettre à jour les règles et les probabilités après l'enregistrement d'un nouvel événement (ex., accident, incident, modification d'un moyen de réduction du risque). Pour pallier cette limite, il faudrait viser l'informatisation de cette méthode.

4.5 Nouvelles voies de recherche

4.5.1 Informatiser la méthode

Actuellement, l'inférence automatique de connaissances après une mise à jour de la base de données par l'utilisateur est lancée manuellement (en lançant le logiciel de fouille de données). De même, en l'état actuel de la démarche proposée, l'actualisation du calcul des probabilités du dommage après mise en œuvre des recommandations pour la sécurité se fera manuellement. Afin de diffuser la méthode proposée par cette thèse, il serait pertinent d'informatiser ces deux processus : génération automatique de connaissances après un nouvel enregistrement dans la base de données et mise à jour des probabilités du dommage associé aux situations dangereuses. L'informatisation de la démarche proposée permettrait de s'affranchir du connaisseur en processus d'extraction de connaissances et en calcul probabiliste. Son apport ne serait requis que pour créer la base de données initiale, notamment sa structure (c.-à-d. le choix des indicateurs d'accidents), tester les premières règles générées et pour mettre en œuvre le calcul de probabilités. Par la suite, l'enregistrement de nouveaux accidents ou d'un nouvel état des moyens de réduction du risque pourrait être entré via une interface du système informatisé, par le préventionniste ou un travailleur autorisé dans l'entreprise. Le système informatisé serait une plateforme créée pour faire le pont entre : 1) le fichier *Excel* contenant la base de données et les calculs probabilistes, 2) puis le logiciel d'inférence de connaissances.

4.5.2 Collecter des avis d'experts

Pour améliorer l'estimation des probabilités du dommage calculées dans l'article 3, il serait pertinent d'y combiner des avis d'experts au sujet de l'occurrence des diverses causes et facteurs de risques des accidents analysés. Comme expliqué dans l'article 1, la collecte d'avis d'experts

dans les calculs probabilistes est utile dans les cas où peu de données sont disponibles. Surtout que, dans notre cas, les accidents moins graves ne sont pas du tout représentés et les accidents graves non mortels sont sous-représentés.

En outre, les avis d'experts peuvent aussi être utiles pour actualiser le risque après la mise en œuvre d'une mesure de réduction du risque, sinon après une amélioration ou une altération d'un moyen de réduction du risque. Les avis seraient récoltés au sujet de l'impact en pourcentage de la mesure de réduction du risque mise en place, de son amélioration et de son altération, sur chaque fonction de masse des indicateurs composant les règles. Les nouvelles fonctions de masse obtenues permettraient de calculer de nouvelles probabilités, d'où l'actualisation du risque.

4.5.3 Exploiter des courbes de fiabilité pour ajouter la notion du temps en prévention des accidents liés aux machines

Avec la démarche d'analyse du risque proposée dans cette thèse, les règles générées informent le préventionniste des indicateurs auxquels il doit prêter attention dans le milieu de travail relatif à la machine en question. Si les conditions constituant une règle sont remplies dans le milieu de travail, cela signifie qu'un accident se trame. Cependant, on ne peut prédire à quel moment. D'où la pertinence d'avoir un historique d'accidents par machines semblables pour savoir le délai qu'il reste pour agir, c'est-à-dire réduire le risque. Par semblable, on entend des machines identiques, installées de la même manière (même inclinaison, usage de la même puissance, convoi de matériaux comparables dans des environnements de travail similaires). Si cette proposition s'avère utopique pour les convoyeurs à courroie, car il est difficile d'avoir la même inclinaison d'un convoyeur à l'autre, elle serait envisageable pour des machines de marques et d'application similaires plus répandues, comme des presses à injection. En effet, lors des visites relatives à l'étude de Chinniah et al. (2014) plusieurs des entreprises visitées pouvaient disposer d'au moins une vingtaine de presses de même marque, produisant des pièces similaires.

Si les accidents analysés dans le cadre de cette thèse étaient survenus dans des conditions similaires (c.-à-d., même type de convoyeur, dans un même environnement de travail et installé selon les mêmes critères d'inclinaison, de puissance, etc.), il aurait été possible de tracer la courbe de fiabilité dans le temps, liée à la survenue des accidents. Plus il y aurait eu d'accidents

similaires documentés, plus la courbe aurait été proche de la réalité générale. Cette courbe serait tracée à partir d'une estimation non paramétrique telle que la méthode de Kaplan-Meier. À l'instar des courbes de fiabilité d'équipements qui permettent d'anticiper les pannes et d'intervenir avant le temps de défaillance, cette courbe de fiabilité liée aux accidents permettrait de prédire le moment où surviendrait le prochain accident. Tracer une telle courbe aurait aussi été possible si, pour une même machine ou un ensemble de machines similaires, l'historique des incidents ou des accidents était disponible. Cela se réaliserait par analogie des historiques de pannes d'équipement similaires à la base de courbes de fiabilité. Étant donné que les machines relatives à cette thèse ne sont pas similaires, une telle courbe de défaillance n'a pu être tracée. Les mesures prises par les préventionnistes avant l'atteinte du temps de l'accident prédit auraient permis de redresser la courbe de fiabilité en diminuant sa pente. Cela aurait pour effet de réduire la probabilité qu'un accident survienne tout en voyant l'effet sur le temps moyen avant défaillance (avant accident). L'avantage d'une telle méthode serait d'avoir deux informations sur une même courbe : en plus d'avoir la probabilité qu'un type d'accident survienne (ce qui permettrait de prioriser un type d'accident par rapport à un autre à des fins de prévention) on dispose du temps auquel ce type d'accident menace d'arriver (ce qui permet de mieux gérer les ressources dans le temps).

4.5.4 Besoin de recherche fondamentale au sujet de la probabilité globale du dommage

Des tentatives ont été menées à la section 3.5.4 pour calculer la probabilité globale du dommage en se basant sur la probabilité de la classe d'accident. Comme résultat, les probabilités des classes « Mortel » et « Non mortel » n'étaient pas complémentaires, contrairement à ce que l'on attend d'un espace échantillon. Deux raisons hypothétiques ont été évoquées :

- les règles obtenues sont des fonctions booléennes partiellement définies plutôt que des fonctions booléennes. Cela empêche de caractériser tous les cas possibles d'accident et réduit donc les chances que l'ensemble des règles d'une classe soit la négation de celle de l'autre classe;

- cette négation n'étant pas rencontrée, la loi de De Morgan n'est pas vérifiée. Alors, la somme des probabilités des classes est ≠ 1.

Ces raisons découlent d'analyse du processus par essai erreur. Il faudrait les approfondir par de la recherche fondamentale, à des fins de preuve mathématique. La recherche fondamentale servirait également à trouver une autre démarche adaptée au calcul de la probabilité globale du dommage.

CHAPITRE 5 CONCLUSION ET RECOMMANDATIONS

Cette thèse propose une démarche transposable d'identification des situations dangereuses et d'estimation quantitative du risque lié aux machines en milieu de travail. L'identification des situations dangereuses est entreprise en exploitant un algorithme d'apprentissage automatique : l'analyse logique de données (ALD). En effet, à partir d'une base de données d'accidents décrits dans des rapports d'enquête de la CNESST, l'ALD génère des règles caractérisant un type d'accident. Chaque règle représente une situation dangereuse et met en évidence l'interaction des principales causes ou facteurs de risque d'accident d'ordre technique, humain ou organisationnel. Ces causes et facteurs de risque sont des indicateurs d'accident. Quant à l'estimation quantitative du risque, elle se base sur l'estimation de la probabilité du dommage en passant par les probabilités des règles générées par l'ALD. Bien que la démarche s'appuie sur des convoyeurs à courroie à des fins démonstratives, elle est applicable à tout type de machine. En effet :

- l'étape C.1 de la démarche illustrée à la figure 3.1 propose de choisir le type de machine sur lequel portera le REX dynamique;
- l'objectif poursuivi par l'analyse du risque déterminera les données à collecter à l'étape C.2 de cette figure. Par exemple, si l'objectif est de connaître la probabilité du dommage, les indicateurs à documenter seront des facteurs de risques ou causes expliquant le contexte accidentel. Si l'objectif est plutôt de connaître la probabilité d'un événement dangereux, les indicateurs seront les éléments contribuant à l'occurrence de l'événement dangereux;
- la méthode de calcul de probabilité présentée en 3.5.3 est applicable au dommage relatif à la situation dangereuse tout comme à l'événement dangereux pouvant déclencher un type d'accident;
- la démarche proposée permet à chaque entreprise de l'adapter à sa réalité en choisissant les indicateurs décrivant cette réalité, ainsi que leurs valeurs possibles.

La démarche proposée comporte plusieurs contributions :

- si appliquée dans un organisme en prévention, elle lui permettra de suivre l'évolution du risque dans son secteur d'activités;

- si appliquée par un préventionniste d'entreprise, elle lui permettra de suivre l'évolution du risque lié à un type de machine de l'entreprise et de le gérer;
- les règles générées pour un type de machine peuvent aider les concepteurs et intégrateurs à identifier les points à améliorer dans leur conception pour augmenter la sécurité de leur équipement;
- la grande capacité discriminante de l'ALD fait de cet algorithme un moyen approprié pour entreprendre de la prévention ciblée. En ciblant les indicateurs qui caractérisent un type d'accident, on peut choisir des moyens de réduction adaptés au risque de cet accident en particulier;
- l'aspect quantitatif de la méthode apporte de l'objectivité dans le processus d'estimation du risque, mais aussi de réduction du risque, grâce à la hiérarchisation des règles générées par l'ALD à partir de leurs probabilités;
- l'approche par retour d'expérience permet la mise à jour des facteurs de risques identifiés et les probabilités calculées;
- l'aspect dynamique de ce retour d'expérience assure l'inférence de connaissances essentielles tirées de rapports d'enquêtes d'accidents;
- le choix de l'ALD assure l'applicabilité du REX dynamique à des bases de données de toutes tailles.

Les règles générées dans cette thèse caractérisent des risques à l'origine d'accidents graves ou mortels. Pour augmenter le pouvoir de généralisation de ces règles à tout type d'accidents liés à un type de machine, des rapports d'enquêtes d'accidents rattachés à des accidents moins graves sont aussi requis. Il est donc recommandé que les accidents moins graves fassent aussi l'objet d'enquêtes et de rapports détaillés d'accidents. Il est aussi recommandé que tous les rapports d'accidents documentent les mêmes indicateurs (standardiser les rapports). Par exemple, éviter que dans un rapport il soit mentionné que l'appréciation du risque avait été réalisée ou non, et omettre cette information dans un autre. Documenter les mêmes indicateurs dans tous les rapports éviterait la présence de données manquantes dans la base de données utilisée dans la démarche proposée. Rappelons que l'absence de données oblige d'éliminer l'indicateur sinon de lui attribuer une valeur qui n'est peut-être pas celle qu'elle était lors de l'accident. Minimiser les

données manquantes optimise donc la qualité des règles générées, car le modèle obtenu approxime mieux la réalité.

Il est aussi recommandé de documenter les presque accidents, c'est-à-dire, les incidents ne causant aucun dommage corporel. Après tout, le retour d'expérience consiste à apprendre d'événements passés négatifs, mais aussi positifs. Encore une fois, avec le grand pouvoir discriminant de l'ALD, il est possible de cibler les indicateurs qui bloquent l'occurrence d'un accident, par simple comparaison des règles caractérisant une classe d'accidents versus une classe de presque accidents.

Enfin, il est recommandé à toute personne qui souhaite utiliser l'ALD pour de la prévention d'accidents, d'opter pour un logiciel optimisé capable de générer le moins de règles possibles tout en fournissant une précision de classification adéquate. Par exemple, cette thèse a permis de constater que *cbmLAD* est plus performant que *LAD-WEKA*.

Dans l'optique d'améliorer la démarche proposée, des recherches sont recommandées pour :

- informatiser la démarche;
- mieux approximer les probabilités calculées en collectant des avis d'experts;
- effectuer de la recherche fondamentale pour trouver une méthode d'estimation de la probabilité globale du dommage;
- ajouter la notion du temps en prévention des accidents liés aux machines, en traçant des courbes de fiabilité illustrant le temps avant qu'un accident arrive, en plus d'informer sur la probabilité de cet accident. Avec la démarche actuelle, on sait qu'un accident est probable et pour quelles raisons, mais on ignore quand il va survenir. En exploitant des courbes de fiabilité, on pourra prédire le moment prévu de l'accident et agir avant cette date pour l'éviter.

Ces recherches contribueront à faciliter la mise à jour et le suivi du risque pour adapter la gestion du risque.

BIBLIOGRAPHIE

Abrahamsson, M. (2002). *Uncertainty in Quantitative Risk Analysis – Characterisation and Methods of Treatment* (Rapport n° 1024). Lund, Sweden: Department of Fire Safety Engineering.

Agard, B. & Kusiak, A. (2005). *Exploration des bases de données industrielles à l'aide du data mining - Perspectives*. Communication présentée au 9ème Colloque National AIP PRIMECA, France (p. 1–9).

Alexe, G. et al. (2006). Breast cancer prognosis by combinatorial analysis of gene expression data. *Breast Cancer Research*, 1–20.

Alexe, S. et al. (2003). Coronary risk prediction by logical analysis of data. *Annals of Operations Research*, *119*, 15–42.

Almuallim, H. & Dietterich, T.G. (1994). Learning Boolean concepts in the presence of many irrelevant features. *Artificial Intelligence*, *69*, 279–306.

Amoore, J. & Ingram, P. (2002). Quality improvement report: learning from adverse incidents involving medical devices. *British Medical Journal*, *325*(7358), 272–275.

Anderson, W.E. (2005). Risk analysis Methodology Applied to Industrial Machine Development. *IEEE Transactions on Industry Applications*, *41*(1), 180–187.

Aneziris, O.M. et al. (2013). Quantification of occupational risk owing to contact with moving parts of machines. *Safety Science*, *51*, 382–896.

Apostolakis, G.E. (2004). How useful is quantitative risk assessment? *Risk Analysis*, *24*(3), 515–520.

Association canadienne de normalisation. (2013). *Control of hazardous energy — Lockout and other methods*. Norme CSA Z460. Mississauga, Ontario, Canada: Association canadienne de normalisation.

American National Standards Institute. (2000). *Risk assessment and risk reduction - A guide to estimate, evaluate and reduce risk associated with machine tools*. Rapport technique ANSI B11.TR3. États-Unis: American National Standards.

American National Standards Institute. (2015). *Safety of machinery*. Norme ANSI B11.0. États-Unis: American National Standards Institute.

Badreddine, A. & Ben Amor, N. (2010). *A new approach to construct optimal bow tie diagrams for risk analysis*. Communication présentée à 23rd International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, Cordoba, Spain (p. 595–604).

Ball, D.J. & Watt, J. (2013). Further thoughts on the utility of risks matrices. *Risk Analysis*, *33*(11), 2068–2078.

Beler, C. (2008). *Modélisation générique d'un retour d'expérience cognitif – Application à la prévention des risques*. (Thèse de doctorat, Université de Toulouse, France).

Bellamy, L.J. et al. (2007). Storybuilder—A tool for the analysis of accident reports. *Reliability Engineering and System Safety*, *92*, 735–744.

Bennane, A. & Yacout, S. (2010). *Processing missing and inaccurate data in a condition based maintenance database*. Communication présentée à IEEE 40th International Conference on Computers and Industrial Engineering (CIE-40), Awaji, Japon (p. 1–5).

Bennane, A. & Yacout, S. (2012). LAD-CBM; new data processing tool for diagnosis and prognosis in condition-based maintenance. *Journal of Intelligent Manufacturing*, *23*, 265–75.

Beriha, G.S., Patnaik, B., Mahapatra, S.S. & Padhee, S. (2012). Assessment of safety performance in Indian industries using fuzzy approach. *Expert Systems with Applications*, *39*, 3311–3323.

Bluff, E. (2014). Safety in machinery design and construction: Performance for substantive safety outcomes. *Safety Science*, *66*, 27–35.

Bonates, T.O. & Gomes, V.S.D. (2014). *LAD-WEKA Tutorial Version 1.0*.

Borg, B. (2002). Predictive safety from near miss and hazard reporting. Tiré de http://signalsafety.ca/files/Predictive-Safety-Near-Miss-Hazard-Reporting.pdf

Boros, E. et al. (2000). An implementation of logical analysis of data. *IEEE Transactions on Knowledge and Data Engineering*, *12*(2), 292–306.

Bounot, J., Mazeau, M. & Jules, D. (1996). La maintenance des bus : analyse des sources d'accidents. *Performances humaines & techniques*, (83), 20–29.

Brauner, M.W. et al. (2007). Logical analysis of computed tomography data to differentiate entities of idiopathic interstitial pneumonias. Dans: Pardalos, P.M., Boginski, V.L., Vazacopoulos, A. (Édit.), *Data Mining in Biomedicine Vol. 7* (p. 193–208). New York: Springer optimization and its applications.

Brin, S., Motwani, R., Ullman, J.D. & Tsur, S. (1997). *Dynamic Itemset Counting and Implication Rules for Market Basket Data*. Communication présentée à SIGMOD/PODS'97, Tucson, AZ, États-Unis (p. 255–264).

Brulotte, L. & Roberge, G. (2006). *Accident mortel survenu à un travailleur le 9 décembre 2005 à l'enrtreprise Henri Radermaker et Fils inc. 1340, route 117 à Rivière-Rouge* (Rapport n° EN-003614). Québec, Canada: CSST.

Buncefield Major Investigation Board (2008). The Buncefield incident 11 December 2005. Bootle, United Kingdom.

Burnham, M. (2015). Targeting zero – Eight questions to ask before using zero as a safety target. *Professional Safety – Safety Management Peer-Reviewed*, 40–45.

Cacciabue, P.C. (2004). Human error risk management for engineering systems: a methodology for design, safety assessment, accident investigation and training. *Reliability Engineering and System Safety*, *83*, 229–240.

Caputo, A.C., Pelagagge, P.M. & Salini, P. (2013). AHP-based methodology for selecting safety devices of industrial machinery. *Safety Science*, *53*, 202–218.

Cavali, E. & Moriggia, V. (2002). Logical data analysis vs. neural networks in the creditworthiness. *Neural Network World*, *4*(2), 371–392.

Centre patronal de santé et sécurité du travail du Québec (CPSST) (2004). Vite, on enquête ! *Convergence*, *20*(2), 6–7.

Cheng, C.-W., Leu, S.-S., Cheng, Y.-M., Wu, T.-C. & Lin, C.-C. (2012). Applying data mining techniques to explore factors contributing to occupational injuries in Taiwan's construction industry. *Accidident Analysis & Prevention*, *48*, 214–222.

Cheng, C.-W., Yao, H.-Q. & Wu, T.-C. (2013). Applying data mining techniques to analyze the causes of major occupational accidents in the petrochemical industry. *Journal of Loss Prevention in the Process Industries*, *26*, 1269–1278.

Childress, S. (2012). How subcontracting affects worker safety. Frontline Enterprise Journalism Group.

Chinniah, Y. (2015). Analysis and prevention of serious and fatal accidents related to moving parts of machinery. *Safety Science*, *75*, 163–173.

Chinniah, Y., Jocelyn, S., Aucourt, B. & Bourbonnière, R. (2014). *Presses à injection de plastique ayant des équipements périphériques – Sécurité lors des interventions de maintenance ou de production* (Rapport n° R-822). Montréal, Québec, Canada: Institut de recherche Robert-Sauvé en santé et en sécurité du travail.

Chinniah, Y., Gauthier, F., Lambert, S. & Moulet, F. (2011). *Experimental analysis of tools used for estimating risk associated with industrial machines* (Rapport n° R-684). Montréal, Québec, Canada: Institut de recherche Robert-Sauvé en santé et en sécurité du travail.

Compiègne, I., Curry, X., Duval, C. & Andéol-Aaussage, B. (2013). *L'analyse de l'accident du travail – La méthode de l'arbre des causes* (Guide nº ED 6163). France: Institut national de recherche et de sécurité (INRS).

Commission de la santé et de la sécurité du travail du Québec (CSST, s.d.). Machines dangereuses. Tiré de http://www.csst.qc.ca/prevention/theme/securite_machines/Pages/accueil.aspx [Consulté le 24 novembre 2013].

Commission de la santé et de la sécurité du travail du Québec (CSST) (1991a). *Rapport d'enquête d'accident* (Rapport n° EN002453). Québec, Canada: CSST.

Commission de la santé et de la sécurité du travail du Québec (CSST) (1991b). *Rapport d'enquête d'accident* (Rapport n° EN002480). Québec, Canada: CSST.

Commission de la santé et de la sécurité du travail du Québec (CSST) (1991c). *Rapport d'enquête d'accident* (Rapport n° EN002496). Québec, Canada: CSST.

Commission de la santé et de la sécurité du travail du Québec (CSST) (1992a). *Rapport d'enquête d'accident* (Rapport n° EN002642). Québec, Canada: CSST.

Commission de la santé et de la sécurité du travail du Québec (CSST) (1992b). *Rapport d'enquête d'accident* (Rapport n° EN002651). Québec, Canada: CSST.

Commission de la santé et de la sécurité du travail du Québec (CSST) (1993). *Rapport d'enquête d'accident* (Rapport n° EN002753). Québec, Canada: CSST.

Commission de la santé et de la sécurité du travail du Québec (CSST) (1995a). *Rapport d'enquête d'accident* (Rapport n° EN002864). Québec, Canada: CSST.

Commission de la santé et de la sécurité du travail du Québec (CSST) (1995b). *Rapport d'enquête d'accident* (Rapport n° EN002891). Québec, Canada: CSST.

Commission de la santé et de la sécurité du travail du Québec (CSST) (1997). *Rapport d'enquête d'accident* (Rapport n° EN003030). Québec, Canada: CSST.

Commission de la santé et de la sécurité du travail du Québec (CSST) (1998). *Rapport d'enquête d'accident* (Rapport n° EN003161). Québec, Canada: CSST.

Commission de la santé et de la sécurité du travail du Québec (CSST) (2000a). *Rapport d'enquête d'accident* (Rapport n° EN003214). Québec, Canada: CSST.

Commission de la santé et de la sécurité du travail du Québec (CSST) (2000b). *Rapport d'enquête d'accident* (Rapport n° EN003249). Québec, Canada: CSST.

Commission de la santé et de la sécurité du travail du Québec (CSST) (2004a). *Rapport d'enquête d'accident* (Rapport n° EN003439). Québec, Canada: CSST.

Commission de la santé et de la sécurité du travail du Québec (CSST) (2004b). *Rapport d'enquête d'accident* (Rapport n° EN003457). Québec, Canada: CSST.

Commission de la santé et de la sécurité du travail du Québec (CSST) (2004c). *Rapport d'enquête d'accident* (Rapport n° EN003478). Québec, Canada: CSST.

Commission de la santé et de la sécurité du travail du Québec (CSST) (2005). *Rapport d'enquête d'accident* (Rapport n° EN003503). Québec, Canada: CSST.

Commission de la santé et de la sécurité du travail du Québec (CSST) (2006a). *Rapport d'enquête d'accident* (Rapport n° EN003569). Québec, Canada: CSST.

Commission de la santé et de la sécurité du travail du Québec (CSST) (2006b). *Rapport d'enquête d'accident* (Rapport n° EN003614). Québec, Canada: CSST.

Commission de la santé et de la sécurité du travail du Québec (CSST) (2007). *Rapport d'enquête d'accident* (Rapport n° EN003657). Québec, Canada: CSST.

Commission de la santé et de la sécurité du travail du Québec (CSST) (2008a). *Rapport d'enquête d'accident* (Rapport n° EN003710). Québec, Canada: CSST.

Commission de la santé et de la sécurité du travail du Québec (CSST) (2008b). *Rapport d'enquête d'accident* (Rapport n° EN003711). Québec, Canada: CSST.

Commission de la santé et de la sécurité du travail du Québec (CSST) (2008c). *Rapport d'enquête d'accident* (Rapport n° EN003733). Québec, Canada: CSST.

Commission de la santé et de la sécurité du travail du Québec (CSST) (2010a). *Plan d'action sécurité des machines* (Document n° DC 900-9123-4). Québec, Canada: CSST.

Commission de la santé et de la sécurité du travail du Québec (CSST) (2010b). *Plan stratégique 2010-2014* (Document n° DC 300-1020). Québec, Canada: CSST.

Commission de la santé et de la sécurité du travail du Québec (CSST) (2012). *Rapport d'enquête d'accident* (Rapport n° EN004024). Québec, Canada: CSST.

Commission de la santé et de la sécurité du travail du Québec (CSST) (2014). *Rapport annuel de gestion 2013* (Rapport n° DC 400-2032-7). Québec, Canada: CSST.

Commission de la santé et de la sécurité du travail du Québec (CSST) (2015a). *Rapport annuel de gestion 2014* (Rapport n° DC 400-2032-8). Québec, Canada: CSST.

Commission de la santé et de la sécurité du travail du Québec (CSST) (2015b). *Données d'exploitation 2014*. Québec, Canada: CSST.

Cooper, D. (2015). Effective safety leadership – Understanding types & styles that improve safety performance. *Professional Safety – Safety Management Peer-Reviewed*, 49–53.

Cox, L.A. (2008). What's wrong with risk matrices? *Risk Analysis*, *28*(2), 497–512.

Crama, Y., Hammer, P.L. & Ibaraki, T. (1988). Cause-effect relationships and partially defined Boolean functions. *Annals of Operations Research*, *16*, 299–326.

de Ruijter, A. & Guldenmund, F. (2016). The bowtie method: a review. Safety Science, 88, 211–218.

Debray, B., Chaumette, S., Descourière, S. & Trommeter, V. (2006). *Méthodes d'analyse des risques générés par une installation industrielle* (Rapport n° INERIS-DRA-2006-P46055-CL47569). Verneuil-en-Halatte, Oise, France: Institut national de l'environnement industriel et des risques (INERIS).

Dei Svaldi, D. & Charpentier, P. (2004). Une étude des accidents en automatisme à partir de la base de données EPICEA. *Hygiène et sécurité du travail*, *196*, 53–73.

Dekker, S. (2006). *The field guide to understanding human error*. Burlington: Ashgate Publishing.

Demichela, M. & Pirani, R. (2013). Human factor analysis embedded in risk assessment of industrial machines: effects on the safety integrity level. *Chemical Engineering Transactions*, *33*, 451–456.

Devooght, J. & Smidts, C. (1992). Probabilistic Reactor Dynamics – I: The Theory of Continuous Event Trees. *Reliability Engineering & System Safety*, *111*, 229–240.

Devooght, J. & Smidts, C. (1996). Probabilistic dynamics as a tool for dynamic PSA. *Reliability Engineering & System Safety*, *52*, 185–196.

Di Gravio, G., Mancini, M., Patriarca, R. & Costantino, F. (2014). *ATM safety management: reactive and proactive indicators – forecasting and monitoring ATM overall safety performance*. Communication présentée à Fourth SESAR Innovation Days, Madrid, Espagne (p. 1–8).

Direction des risques professionnels – Mission Statistiques, Caisse nationale de l'assurance maladie des travailleurs salariés (DRP / CNAMTS) (2015). *Risque AT 2014 : statistiques de sinistralité tous CTN et par CTN* (Étude 2015-149-CTN). France: CNAMTS.

Duijm, N.J. (2015). Recommandations on the use and design of risk matrices. *Safety Science*, *76*, 21–31.

Dulac, N. (2007). *A Framework for Dynamic Safety and Risk Management Modeling in Complex Engineering Systems*. (Thèse de doctorat, MIT, États-Unis).

Dźwiarek, M. (2004). An analysis of accidents caused by improper functioning of machine control systems. *International Journal of Occupational Safety and Ergonomics (JOSE)*, *10*(2), 129 – 136.

Équipes du Programme REX FonCSI (2008). *Le retour d'expérience – Facteurs socio-culturels du REX : sept études de terrain* (Les cahiers de la sécurité industrielle n° 2008-05). Toulouse, France: Fondation pour une culture de sécurité industrielle (FonCSI).

Escobar, R.L. & Lévêque, F. (2014). How Fukushima Dai-ichi core meltdown changed the probability of nuclear accidents? *Safety Science*, *64*, 90–98.

Etherton, J., Main, B., Cloutier, D. & Christensen, W. (2008). Reducing Risk on Machinery: A Field Evaluation Pilot Study of Risk Assessment. *Risk Analysis*, *28*(3), 711–721.

Etherton, J.R. (2007). Industrial machine systems risk assessment: a critical review of concepts and methods. *Risk Analysis*, *27*(1), 71–82.

European Commission (2008). *Risk assessment guidelines for non-food consumer products* (Rapport technique). Draft.

European Machinery Directive 2006/42/EC, 2006.

Feng, S. , Li, Z., Ci, Y. & Shang, G. (2016). Risk factors affecting fatal bus accident severity: Their impact on different types of bus drivers. *Accident Analysis & Prevention*, *86*, 29–39.

Flaspöler, E. et al. (2010). *The human-machine interface as an emerging risk* (Rapport n° TE-80-10-196-EN-N). European Agency for Safety and Health at Work.

Fodor, I.K. (2002). *A survey of dimension reduction techniques*. Livermore, CA, États-Unis: Center for Applied Scientific Computing, Lawrence Livermore National Laboratory.

Fontaine, F. et al. (2006). *La prévention des accidents liés aux pièces en mouvement* (Document n° DC 300-436). Québec, Canada: Commission de la santé et de la sécurité du travail du Québec (CSST).

Gadd, S., Keeley, D. & Balmforth, H. (2003). *Good practice and pitfalls in risk assessment* (Rapport n° 151). Norwich, UK: Health & Safety Laboratory.

Gardner, D., Cross, J.A., Fonteyn, P.N., Carlopio, J. & Shikdar, A. (1999). Mechanical equipment injuries in small manufacturing business. *Safety Science*, *33*, 1–12.

Gauthey, O. (2008). *Le retour d'expérience – État des pratiques industrielles* (Les cahiers de la sécurité industrielle n° 2008-02). Toulouse, France: Institut pour une culture de sécurité industrielle (ICSI).

Gauthier, F., Chinniah, Y., Burlet-Vienney, D., Aucourt, B. & Larouche, S. (2016). *Sécurité des machines - Expérimentation pratique de paramètres et d'outils d'estimation du risque*

(Rapport n^{o} R-940). Montréal, Québec: Institut de recherche Robert-Sauvé en santé et en sécurité du travail (IRSST).

Gauthier, F., Lambert, S. & Chinniah, Y. (2012). Experimental analysis of 31 risk estimation tools applied to safety of machinery. *International Journal of Occupational Safety and Ergonomics (JOSE)*, *18*(2), 245–265.

Giraud, L., Massé, S., Dubé, J., Schreiber, L. & Turcot, A. (2003). *Sécurité des convoyeurs à courroie – Guide de l'utilisateur* (2ème édition). Québec, Canada: Commission de la santé et de la sécurité du travail (CSST).

Gonzalez-Delgado, M. et al. (2015). Factors Associated with Fatal Occupational Accidents among Mexican Workers: A National Analysis. *PLOS ONE*, *10*(3), 1–19.

Hale, A.R. et al. (2007). Modeling accidents for prioritizing prevention. *Reliability Engineering & System Safety*, *92*, 1701–1715.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I.H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, *11*(1), 10–18.

Hammer, P.L. & Bonates, T.O. (2006). Logical analysis of data—An overview: from combinatorial optimization to medical applications. *Annals of Operations Research*, *148*, 203–225.

Hammer, P.L., Kogan, A. & Lejeune, M.A. (2009). Reverse-engineering country risk ratings: a combinatorial non-recursive model. *Annals of Operations Research*, *188*, 185–213.

Hammer, P.L., Kogan, A. & Lejeune, M.A. (2012). A logical analysis of banks' financial strength ratings. *Expert Systems with Applications*, *39*, 7808–7821.

Hammer, P.L., Kogan, A., Simeone, B. & Szedmák, S. (2004). Pareto-optimal patterns in logical analysis of data. *Discrete Applied Mathematics*, *144*, 79–102.

Healey, N. (2006). *Analysis of RIDDOR machinery accidents in the UK printing and publishing industries 2003–2004* (Rapport n° HSL/2006/83). Derbyshire, U.K.: Health & Safety Laboratory (HSL).

Health and Safety Executive (HSE). (2004), Investigating accidents and incidents – A workbook for employers, unions, safety representatives and safety professionals (Workbook nº HSG245). U.K. : Health and Safety Executive (HSE).

Hietikko, M., Malm, T. & Alanen, J. (2011). Risk estimation studies in the context of a machine control function. *Reliability Engineering & System Safety*, *96*, 767–774.

Hollnagel, E. (2004). *Barriers and accident prevention – or how to improve safety by understanding the nature of accidents rather than finding their causes*. Burlington: Ashgate Publishing.

Hounnou, L. & Parrennes, F. (2014). *Anticiper l'évolution des précurseurs de danger par le développement d'une fonction prédictive*. Communication présentée à 19ème Congrès Lambda-Mu Maîtrise des risques et sureté de fonctionnement, Dijon, France (p. 1–8).

Hubbard, D. & Evans, D. (2010). Problems with scoring methods and ordinal scales in risk assessment. *IBM Journal of Research & Development*, *54*(3), 2:1–2:10.

Jabrouni, H. (2012). *Exploitation des connaissances issues des processus de retour d'expérience industriels*. (Thèse de doctorat, Université de Toulouse, Toulouse, France).

Japan International Center for Occupational Safety and Health (JICOSH) (s.d.). Industrial Accidents Statistics in Japan. http://www.jisha.or.jp/english/statistics/accidents2010.html#Fig1

Jocelyn, S., Chinniah, Y. & Ouali, M.-S. (2016). Contribution of dynamic experience feedback to the quantitative estimation of risks for preventing accidents: A proposed methodology for machinery safety. *Safety Science*, *88*, 64–75.

Jocelyn, S., Chinniah, Y., Ouali, M.-S. & Yacout, S. (2017). Application of logical analysis of data to machinery-related accident prevention based on scarce data. *Reliability Engineering & System Safety*, *159*, 223–236.

Johnson, C. (2002). Software tools to support incident reporting in safety-critical systems. *Safety Science*, *40*(9), 765–780.

Johnson, C. & Holloway, C.M. (2003). A survey of logic formalisms to support mishap analysis. *Reliability Engineering and System Safety*, *80*(3), 271–291.

Jourdain B. (2013). *Probabilités et statistiques*. Paris, France : École des Ponts.

Keane, J.M. (2015). Preventing major losses – Changing OSH paradigms & practices. *Professional Safety – Safety Management Peer-Reviewed*, 42–48.

Kim, H.H. & Choi, J.Y. (2015a). Hierarchical multi-class LAD based on OvA-binary tree using genetic algorithm. *Expert System with Applications*, *42*, 8134–8145.

Kim, H.H. & Choi, J.Y. (2015b). Pattern generation for multi-class LAD using iterative genetic algorithm with flexible chromosomes and multiple populations. *Expert System with Applications*, 42, 833–843.

Kjellen, U., Rundmo, T., Sandetory, H. & Sten, T. (1990). Safety analysis of manual tasks in automatic production systems – implications for design. *Accident Analysis & Prevention*, *22*(5), 475–486.

Kletz, T.A. (1993). *Lessons from disaster – How organisations have no memory and accidents recur*. Melksham, UK: Institution of Chemical Engineers.

Kriegel, H.-P. et al. (2007). Future trends in data mining. *Data Mining and Knowledge Discovery*, *15*, 87–97.

Lamy, P. & Charpentier, P. (2009). Estimation des risques machines – Recensement des méthodes et subjectivité des paramètres de l'estimation. *Hygiène et sécurité du travail*, *214*, 37–44.

Lamy, P., Levrat, E. & Paques J.-J. (2006). *Méthodes d'estimation des risques machines : analyse bibliographique*. Communication présentée au 15ème Congrès Lambda-Mu Maîtrise des risques et sûreté de fonctionnement, Lille, France (p. 1–8).

Lannoy, A. (2011). Retour d'expérience technique. Dans: *Management de la sécurité, vol. SE 1 041v2* (p. 1–22). Techniques de l'ingénieur.

Lauer, M.S. et al. (2002). Use of the logical analysis of data method for assessing long-term mortality risk after exercise electrocardiography. *Circulation*, *106*, 685–690.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M.S. & Kruschwitz, N. (2011). Big Data, analytics and the path from insights to value. *MIT Sloan Management Review*, *52*(2), 21–32.

Lebeau, M. & Duguay, P. (2011). *Le coût des lésions professionnelles – Une revue de littérature* (Rapport n° R-676). Montréal, Québec: Institut de recherche Robert-Sauvé en santé et en sécurité du travail (IRSST).

Leveson, N.G. (2011). *Engineering a safer world: Systems thinking applied to safety*. Cambridge: The MIT Press.

Lindberg, A.-K. & Hansson, S.O. (2006). Evaluating the effectiveness of an investigation board for workplace accidents. *Policy and Practice in Health and Safety*, *4*(1), 63–79.

Lindberg, A.-K., Hansson, S.O. & Rollenhagen, C. (2010). Learning from accidents – What more do we need to know?, *Safety Science*, *48*, 714–721.

Lindquist, R. (2011). The secret to sustainment. *Quality Progress*, *44*(8), 40-45.

López-Soto, D., Yacout, S. & Angel-Bello Fr. (2016). Root cause analysis of familiarity biases in classification of inventory items based on logical patterns recognition, *Computers and Industrial Engineering*, *93*(C), 121–130.

Lupin, H. & Marsot, J. (2006). *Sécurité des machines et des équipements de travail – Moyens de protection contre les risques mécaniques* (Guide n° ED 807). Vandœuvre-lès-Nancy, Lorraine, France : Institut national de recherche et de sécurité (INRS).

Malm, T. et al. (2010). Safety of Interactive Robotics – Learning from Accidents. *International Journal of Social Robotics*, *2*(3), 221–227.

Massé, S., Giraud, L., Dubé, J., Vernoux, G., Schreiber, L. & Desrochers, Y. (2003). *Sécurité des convoyeurs à courroie – Principes de conception pour améliorer la sécurité – Guide du concepteur* (2ème édition). Québec : CSST (Commission de la santé et de la sécurité du travail).

Mbaye, S., Kouabenan, R. & Sarnin, P. (2009). *Le retour d'expérience – Processus socio-cognitifs dans l'explication des dysfonctionnements* (Les cahiers de la sécurité industrielle n° 2009-08). Toulouse, France: Fondation pour une culture de sécurité industrielle (FonCSI).

McManus, N. (2014). Incident records – Understanding the past to prevent future hazardous energy incidents. *Professional Safety – Safety Management Peer-Reviewed*, 34–43.

Mili, A., Bassetto, S., Siadat, A. & Tollenaere, M. (2009). Dynamic risk management unveil productivity improvements. *Journal of Loss Prevention in the Process Industries*, *22*, 25–34.

Moatari-Kazerouni, A., Chinniah, Y. & Agard, B. (2015). A proposed health and safety risk estimation tool for manufacturing systems. *International Journal of Production Research*, *53*(15), 4459–4475.

Mortada, M. & Yacout, S. (2011). *cbmLAD – Using logical analysis of data in condition based maintenance*. Communication présentée au 3rd International Conference on Computer Research and Development (ICCRD), Shanghai, Chine (p. 30–34).

Mortada, M.A., Yacout, S. & Lakis, A. (2011). Diagnosis of rotor bearings using logical analysis of data. *Journal of Quality in Maintenance Engineering*, *17*, 371–397.

Musharraf, M. et al. (2013). Human reliability assessment during offshore emergency conditions. *Safety Science*, *59*, 19–27.

Nadeau, S. (2015). La gestion intégrée des risques de santé et de sécurité du travail – Une mode ou une solution ? *Travail et santé*, *30*(5), 4.

Naderpour, M., Lu, J. & Zhang, G. (2014). A situation risk awareness aproach for process systems safety. *Safety Science*, *64*, 173–189.

Nanda Tchiehe, D. & Gauthier, F. (2017). Classification of risk acceptability and risk tolerability factors in occupational health and safety. *Safety Science*, *92*, 138–147.

Nix, D. (2012). The probability paradox: Risk assessment and machine safety. Manufacturing Automation. From: http://www.automationmag.com/factory/2828-the-probability-paradox-risk-assessment-and-machine-safety

Organisation internationale de normalisation. (2007). *Sécurité des machines — Appréciation du risque — Partie 2 : Lignes directrices pratiques et exemples de méthodes*. Norme ISO 14121-2. Genève, Suisse: Organisation internationale de normalisation.

Organisation internationale de normalisation. (2010). *Sécurité des machines — Principes généraux de conception — Appréciation du risque et réduction du risque*. Norme ISO 12100. Genève, Suisse: Organisation internationale de normalisation.

Organisation internationale du travail (OIT) (s.d.). Sécurité et santé au travail. http://www.ilo.org/global/topics/safety-and-health-at-work/lang--fr/index.htm

OSHA (Occupational Safety and Health Administration). (2007). Safety and Health at Work is Everyone's Concern : Risk Assessment Tool. From: http://osha.europa.eu/en/campaigns/hwi/about/material/rat2007 (accessed on July 17th, 2012). Offenhuber, D. (2010). Visual anecdote. *Leonardo, 43*(4), 367–374.

Papazoglou, I.A., Aneziris, O., Bellamy, L., Ale, B.J.M. & Oh, J.I.H. (2015). Uncertainty assessment in the quantification of risk rates of occupational accidents. *Risk Analysis, 35*(8), 1–26.

Paques, J.-J. et al. (2004). *Sécurité des machines : phénomènes dangereux, situations dangereuses, événements dangereux, dommage* (Pochette n^{o} DC 900-337-1PDF (06-11). Montréal, Québec, Canada: Commission de la santé et de la sécurité du travail du Québec (CSST).

Paques, J.-J., Gauthier, F. & Perez, A. (2007). Analysis and Classification of the Tools for Assessing the Risk Associated With Industrial Machines. *International Journal of Occupational Safety and Ergonomics, 13*(2), 173-187.

Paques, J.-J., Gauthier, F., Pérez, A., Charpentier, P., Lamy, P. & David, R. (2006). *Bilan raisonné des outils d'appréciation des risques associés aux machines industrielles* (Rapport n° R-459). Montréal, Québec: Institut de recherche Robert-Sauvé en santé et en sécurité du travail (IRSST).

Parkes, K.R. (2012). Shift schedules on North Sea oil/gas installations: a systematic review of their impact on performance, safety and health. Safety Science, 50(7), 1636–1651.

Poisson, P. & Chinniah, Y. (2016). Managing risks linked to machinery in sawmills by controlling hazardous energies: Theory and practice in eight sawmills. *Safety Science*, *84*, 117–130.

Programme on Safety and Health at Work and the Environment (SafeWork). (2013). *Training Package on Workplace Risk Assessment and Management for Small and Medium-Sized Enterprises*. Genève, Suisse: International Labour Organization.

Publications du Québec (s.d.), "Règlement sur la santé et la sécurité du travail," Publications du Québec. [En ligne]. Disponible : http://legisquebec.gouv.qc.ca/fr/ShowDoc/cr/S-2.1,%20r.%2013. [Consulté le 14 décembre 2016].

Pyle, D. (1999). *Data preparation for data mining*. San Francisco, CA, États-Unis: Morgan Kaufmann Publishers, Inc.

Ragab, A., Yacout, S. & Ouali, M.S. (2015). *Interpretable pattern-based machine learning for condition based maintenance*. Communication présentée à RAMS, 2015 The 61st Annual Reliability & Maintainability Symposium, Palm Harbor, FL, États-Unis (p. 1–17).

Ragab, A., Yacout. S. & Ouali, M.S. (2013). Intelligent data mining for automatic face recognition. *The Online Journal of Science and Technology*, *3*, 97–101.

Rakotomalala, R. (2010). Tanagra (Version 1.4) [Logiciel]. Lyon, France: Université de Lyon 2.

Rakotomalala, R. (s.d.). *Les règles d'association – Market Data Analysis ou L'analyse du panier de la ménagère* (Tutoriels Tanagra). Lyon, France : Université de Lyon 2. Tiré de eric.univ-lyon2.fr/~ricco/cours/slides/regles_association.pdf

Rathnayaka, S., Khan, F. & Amyotte, P. (2011a). SHIPP methodology: Predictive accident modeling approach. Part I: Methodology and model description. *Process Safety and Environmental Protection*, *89*, 151–164.

Rathnayaka, S., Khan, F. & Amyotte, P. (2011b). SHIPP methodology: Predictive accident modeling approach. Part II: Validation with case study. *Process Safety and Environmental Protection*, *89*, 75–88.

Raviv, G., Shapira, A. & Fishbain, B. (2017). AHP-based analysis of the risk potential of safety incidents: Case study of cranes in the construction industry. *Safety Science*, *91*, 298–309.

Rivas, T. et al. (2011). Explaining and predicting workplace accidents using data-mining techniques. *Reliability Engineering & System Safety*, *96*, 739–747.

Ruff, T., Coleman, P. & Martini, L. (2011). Machine-related injuries in the US mining industry and priorities for safety research. *International Journal of Injury Control and Safety Promotion*, *18*(1), 11–20.

Ryoo, H.S. & Jang, I.Y. (2009). MILP approach to pattern generation in logical analysis of data. *Discrete Applied Mathematics*, *157*, 749–761.

Saad, S. & Arakaki, R. (2014). *An Event Processing Architecture for Operational Risk Management in an Industrial Environment*. Communication présentée à Proceedings of the 8th ACM International Conference on Distributed Event-Based Systems, New-York, États-Unis (p. 213–224).

Safe Work Australia (2015). *Work-related traumatic injury fatalities, Australia 2014* (Rapport n^{o} 978-1-76028-882-2). Canberra, Australie: Safe Work Australia.

Santosa, F. (1997). Data mining: an industrial research perspective. *IEEE Computational Science & Engineering*, 6–9.

Saric, S., Bab-Hadiashar, A., Hoseinnezhad, R. & Hocking, I. (2013). Analysis of forklift accident trends within Victorian industry (Australia). *Safety Science*, *60*, 176–184.

Shah, S., Horne, A. & Capellá, J. (2012). Good data won't guarantee good decisions. *Havard Business Review*, 3–5.

Silva, J.F. & Jacinto, C. (2012). Finding occupational accident patterns in the extractive industry using a systematic data mining approach. *Reliability Engineering System Safety*, *108*, 108–122.

Stauffer, D. (2002). How good data leads to bad decisions. *Harvard Management Update*, 25–27.

Swaminathan, S. & Smidts, C. (1999). The event sequence diagram framework for dynamic probabilistic risk assessment. *Reliability Engineering & System Safety*, *63*, 73–90.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

United States Department of Labor. (2002). *Job hazard Analysis* (Guide n° OSHA 3071). États-Unis: Occupational Safety and Health Administration (OSHA).

United States Department of Labor. (2015). *Incident [Accident] investigations: a guide for employers – A systems approach to help prevent injuries and illnesses*. États-Unis: Occupational Safety and Health Administration (OSHA).

van den Honert, A.F. & Vlok, P.J. (2016). Estimating the continuous risk of accidents occurring in the mining industry in South Africa. *South African Journal of Industrial Engineering*, *26*(3), 71–85.

van Loggerenberg, N.J.F. (2014). *Achieving total safety culture through behavior based safety, establishing and maintaining an injury free culture*. Communication présentée à Probabilistic Safety Assessment and Management (PSAM), Honolulu, Hawaii (p. 1–9).

Van Wassenhove, W. & Garbolino, E. (2008). *Retour d'expérience et prévention des risques – Principes et méthodes*. Editions Lavoisier TEC & DOC.

Verma, A., Khan, S.D., Maiti, J. & Krishna O.B. (2014). Identifying patterns of safety related incidents in a steel plant using association rule mining of incident investigation reports. *Safety Science*, *70*, 89–98.

Villa, V., Paltrinieri, N., Khan, F. & Cozzani, V. (2016). Towards dynamic risk analysis: A review of the risk assessment approach and its limitations in the chemical process industry. *Safety Science*, *89*, 77–93.

Wachter, J.K. & Ferguson, L.H. (2013). Fatality Prevention Findings from the 2012 forum. *Professional Safety – Safety Management Peer-Reviewed*, 41–49.

Walter, J. (2017). Safety management at the frontier: cooperation with contractors in oil and gas companies. *Safety Science, 91*, 394–404.

Witten, I.H., Frank, E. & Hall, M.A. (2011). Chapter 5 – Credibility: Evaluating what's been learned. Dans: *Data Mining Practical Machine Learning Tools and Techniques* (3ème édition, p. 147–190). Burlington, MA, États-Unis: Morgan Kaufmann Publishers.

Worsell, N. & Ioannides, A. (2000). *Machinery risk assessment validation literature review* (Rapport n° HSL/2000/18). Broad Lane, Sheffield, Angleterre : Health and Safety Laboratory (HSL).

Yacout, S. (2010a). *Fault detection and diagnosis for condition based maintenance using the logical analysis of data*. Communication présentée à IEEE 40th International Conference on Computers and Industrial Engineering (CIE-40), Awaji, Japon (p. 1–6).

Yacout, S. (2010b). Tool and method for fault detection of devices for condition based maintenance. *Brevet canadien provisoire n° PCT/CA2011/000876*.

Zhang, A., Boardman, A.E., Gillen, D. & Waters II, W.G. (2004). *Towards Estimating the Social and Environmental Costs of Transportation in Canada – A Report for Transport Canada*. Vancouver, Colombie-Britannique: The University of British Columbia, Centre for Transportation Studies.

Zighed, D.A. & Rakotomalala, R. (2011). Extraction de connaissances à partir de données (ECD). Dans: *Technologies logicielles Architectures des systèmes, vol. H 3 744* (p. 1–22). Techniques de l'ingénieur.

ANNEXE A – ARTICLE 1

"CONTRIBUTION OF DYNAMIC EXPERIENCE FEEDBACK TO THE QUANTITATIVE ESTIMATION OF RISKS FOR PREVENTING ACCIDENTS: A PROPOSED METHODOLOGY FOR MACHINERY SAFETY"

Source : http://dx.doi.org/10.1016/j.ssci.2016.04.024

Safety Science 88 (2016) 64–75

Contents lists available at ScienceDirect

Safety Science

journal homepage: www.elsevier.com/locate/ssci

Contribution of dynamic experience feedback to the quantitative estimation of risks for preventing accidents: A proposed methodology for machinery safety

Sabrina Jocelyn [a,b,*], Yuvin Chinniah [b], Mohamed-Salah Ouali [b]

[a] Institut de recherche Robert-Sauvé en santé et en sécurité du travail (IRSST), 505 de Maisonneuve Blvd. West, Montreal, Quebec H3A 3C2, Canada
[b] Department of Mathematical and Industrial Engineering, Polytechnique Montréal, 2500 chemin de Polytechnique, Montreal, Quebec H3T 1J4, Canada

ARTICLE INFO

Article history:
Received 25 July 2015
Received in revised form 11 March 2016
Accepted 25 April 2016
Available online 6 May 2016

Keywords:
Machinery safety
Accident
Quantitative risk estimation
Experience feedback
Updating risk
Logical Analysis of Data (LAD)

ABSTRACT

This paper proposes a methodological approach for designing a dynamic risk identification and estimation support tool for machinery safety. Based on a comprehensive literature review and by updating the risks through dynamic experience feedback integrated into quantitative risk estimation, the methodology makes it possible to better equip machinery safety practitioners to intervene effectively. The methodology combines dynamic risk identification and Logical Analysis of Data (LAD) as two potential methods applied in machinery safety. LAD is an artificial intelligence technique introduced to extract information from accident reports in order to analyze machinery-related accidents in the workplace, which has not been covered in previous studies of machinery safety. The practical relevance and feasibility of the proposed methodology are explained using an example involving two accidents that occurred on the same machine in the same sawmill.

1. Introduction

Machinery-related industrial accidents cause significant bodily injuries, in particular death, and reversible and permanently disabling injuries. In the United Kingdom in 2003 and 2004, half of the accidents related to operating machinery were attributable to printing presses and conveyors (Healey, 2006). In Australia between 2003 and 2013, 84 deaths involving moving machinery occurred (Safe Work Australia, 2014). In the United States between 1980 and 1995, machinery was the second main cause of death at work and, from 1992 to 2001, an annual average of 148 fatal and 318 488 non-fatal accidents occurred on operating machinery (Flaspöler et al., 2010). In

Canada, 70% of the 1975 deaths occurring between 1990 and 2008 were attributable to machinery and farm equipment (McManus, 2014). In Quebec in 2013, machinery-related injuries totalled 3503, including seven deaths and 800 accidents involving access to moving parts (CSST, 2014).

Chinniah (2015) analyzed 106 serious and fatal accidents linked to the moving parts of fixed machinery that occurred in Quebec between 1990 and 2011. The study identifies the following main causes of accident: easy access to moving parts, lack of safeguarding, bypassing safeguards, absence of lockout procedures during maintenance, lack of training, unexperienced workers, modifications to the machinery and their safety control systems, and lack of risk assessment. Other studies that deal with analysis of machinery-related accidents (Lindquist, 2011; Caputo et al., 2013; Gardner et al., 1999) point to some or all of the same causes.

In the field of safety of machinery, there are hundreds of standards. One important standard is ISO 12100 which describes design principles, risk assessment and risk reduction for machinery. It is intended primarily for machine designers and manufacturers, but is widely used when existing machines are modified or residual risk need to be reduced by end users (e.g. in factories). Although ISO 12100:2010 stresses the importance of experience feedback about machinery-related accidents to be used as inputs for safer machine designs, the use of such feedback in reality is quite limited. One example is the lack of consideration for maintenance activities on machinery by machine builders. This translates in machines which are poorly designed and then exposing maintenance personnel to high risks, as supported by the large number of accidents during maintenance activities. Safety practitioners in the workplace learn from accident causes described in accident reports as well as from near misses (i.e. incidents). However, such learning is based on experience feedback described as static, as the knowledge available in the accident and incident reports is limited to case studies. To improve the feedback, a dynamic method which allows new knowledge to be inferred from the information contained in case studies and other circumstantial events seems more adequate. Dynamic experience feedback makes it possible to tap the full potential of accident and incident reports that, at present, are not being used to their fullest.

Moreover, safety practitioners want to choose and apply an optimal risk estimation method that provides useful results with minimal effort (Chinniah et al., 2011). The choice is even more difficult for small and medium-sized enterprises (SMEs) that have few or no risk analysis resources at their disposal. Unfortunately, occupational health and safety (OHS) is often a part time activity for a single resource and optimal usage of the time allocated to risk identification and reduction is crucial. SMEs are considered to be particularly vulnerable to the impact of occupational accidents morally and economically (Programme on Safety and Health, 2013). According to Chinniah (2015), enterprises with limited occupational health and safety resources need, among other things, to prioritize risk assessment or at least hazard identification.

In addition, to reduce the direct costs related to the victim’s absence from work or to the production shutdown that occurs when the machine is sealed by a labor inspector for serious non-compliance, safety practitioners should identify and analyze the risk associated with their machines. To do so, they typically use a qualitative risk matrix tool for risk estimation due to its simplicity and the ease of integrating the results into risk management policies. However, risk matrices have limitations in the area of risk ranking (Cox, 2008). It should be noted that

probability-based quantitative risk estimation tools are rarely used for machinery safety (Gadd et al., 2003), even if some studies (Cox, 2008; Duijm, 2015) find them more effective than qualitative tools. Quantitative risk estimation is mostly used for process safety involving complex systems and reliability considerations (e.g. nuclear or chemical process installations). Risk quantification for occupational injuries, including injuries caused by moving machine parts, has been studied (Papazoglou et al., 2015; Aneziris et al., 2013; Demichela and Pirani, 2013). However, in qualitative and quantitative methods, the risk estimation is frozen in time. The evolution of the machinery is not integrated in the process. New information about usage is neglected.

Hence, three questions arise from these considerations:

1) How to help safety practitioners efficiently identify and estimate machinery-related risks?
2) How to help safety practitioners prioritize risk reduction measures and, at the same time, keep their machines from being sealed by an OHS inspector who identifies a non-compliance with current OHS legislation following a routine inspection or an occupational accident?
3) How to help safety practitioners monitor risk progression?

To answer these questions, it would be relevant for safety practitioners to have access to an efficient (i.e. that enables targeted prevention), easy-to-use tool that provides information about the risks present in a machine and would enable them both to prevent accidents in order of priority and to avoid having seals placed on their equipment. With that in mind, this paper proposes and describes a methodological approach for designing a dynamic tool to support machinery-safety decisions. This paper aims to introduce two potential methods in the field of machinery safety. These concepts concern dynamic experience feedback integrated into quantitative risk estimation and extraction of relevant information from accident reports using Logical Analysis of Data (LAD) as an artificial intelligence technique. Accordingly, the paper focuses especially on: the comprehensive literature review that leads to the proposed methodology, the description and justification of each step of the methodology as well as relationships between them, and the practical relevance and feasibility of the proposed methodology using an example of two accidents that occurred on the same sawmill machine.

The methodology involves updating the risks by integrating dynamic experience feedback into risk estimation, and integrating LAD into dynamic experience feedback. LAD is a data mining technique and optimisation combinatorial algorithm based on Boolean logic. This algorithm is known for its robust performance (even when data is scarce) in medicine, for disease diagnosis and prognosis (e.g. Alexe et al., 2003), finance (Hammer et al., 2009). It is also known for its capacity to characterize and distinguish classes of events. Thus, it is useful for machinery risk identification respecting different contexts of use. Accordingly, it is useful for targeted prevention. LAD has been used neither to occupational safety nor to machinery safety. Thus, proposing LAD for machinery safety is one novelty that this paper brings. What is also original about this approach is that the updating of the probability of a hazardous scenario or event is based on accident reports as well as on new events detected over time through inspections. The combination of these techniques constitutes the added value of the tool over the static qualitative tools generally used in the machinery safety field.

The proposed combination aims first to improve the risk identification and estimation steps of the risk management process in machinery safety (Figure A1). It then aims to perpetuate the risk

estimation process (i.e. update the risk) even if the risk has been adequately reduced. Machine conditions of use can reduce the effectiveness of the risk reduction measures in place, which can in turn affect risk estimation. The risk reduction measures then become insufficient to tackle the new risks.

This technique thus facilitates risk identification since the tool itself will identify the main direct and indirect causes of accidents. Risk estimation will become robust and leave less room for interpretation. Dynamic risk estimation tools are more realistic and thus more effective than static tools, as they reflect the progression of the risk.

The remainder of this paper is divided into five sections. Section 2 presents a comprehensive review of the literature on risk identification and risk estimation tools in machinery safety and on risk management methods used in other engineering fields, as well as the medical and the financial fields. Section 3 introduces LAD technique and explains its main steps to extract information. Section 4 presents the possible contribution of risk management methods from other fields when applied to machinery safety. It also discusses the three questions asked above in light of these results. Section 5 describes in details the proposed methodology for the dynamic tool for supporting decision-making and explains its usefulness with an example. Section 6 presents the conclusion of the paper and further related research areas.

2. Risk Analysis – Available Tools and Techniques

In machinery safety, ISO 12100:2010 defines risk as the combination of the probability of occurrence of harm to humans and the severity of this harm (ISO, 2010). The probability of occurrence of harm is a function of: (1) exposure of the person or persons to a hazard; (2) the probability of occurrence of the hazard; and (3) the possibility of avoiding or limiting the harm. As is true for preventing the consequences related to any hazard, preventing machine-related accidents is achieved through an iterative risk management process (Figure A1) that:

- determines the machine's conditions of use;
- identifies the associated risks that threaten the users' health and safety (hazardous phenomena, situations, and events);
- estimates the risks by determining rankings for the risk estimation parameters;
- evaluates the risks for the purpose of prioritizing the actions to take to reduce them. The evaluation consists of establishing the acceptability of the risks and is determined by agreement among the various stakeholders (e.g. employers, practitioners, workers exposed to the risks);
- checks whether adding risk reduction measures has the intended effect. These means reduce the risk by acting entirely or partially on elements defining the risk.

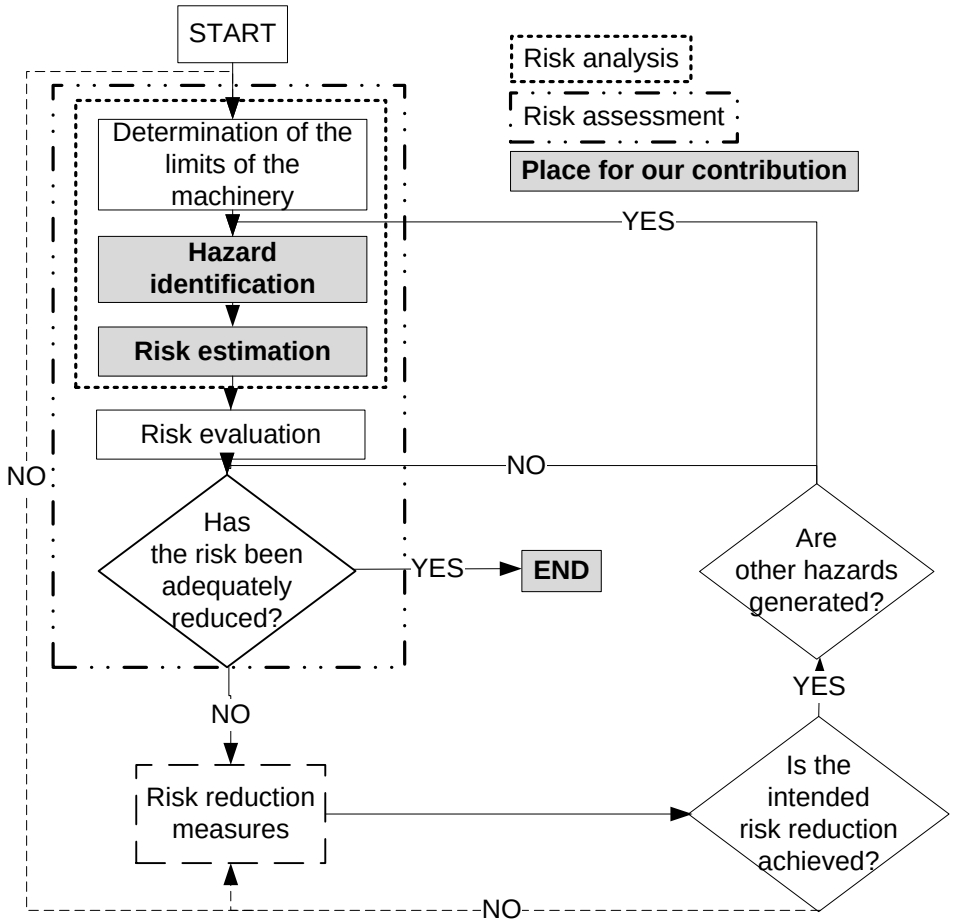


Figure A1: Simplified risk management diagram for machinery safety (inspired by ISO 12100:2010) and the contribution of this research

The risk reduction measures are implemented by the machine designer. They are divided into three groups, from the most effective to the least: (1) inherently safe design (e.g. eliminating the hazard at the source); (2) safeguarding and complementary protective measures (e.g. guards, protective devices); and (3) information for use (e.g. visual signals, instructions). The machine is then sold to the user but with a residual risk that is largely dependent on the actual conditions of use. One problem is that residual risk is not properly documented and explained. The user (employer, safety practitioner, worker) must reduce the risk by adopting various means, such as organizational measures (safe work practices, surveillance, work permit system, etc.), additional means of protection, the wearing of personal protective equipment (PPE), and training. For users, no effectiveness ranking for risk reduction measures is established by ISO 12100:2010 in contrast to the effectiveness ranking mentioned earlier for the machine designer.

Additional means of protection are used, for example, in situations where the machine user has an integrator modify a machine to meet specific manufacturing requirements not considered by the original designer. These modifications are all the more problematic for the integrator when there is no type-C standard (standard specific to a machine type) associated with the machine

concerned. By relying on more general standards such as ISO 12100:2010 and risk identification or estimation tools, the integrator adapts the machine as best he or she can to meet the manufacturing criteria while ensuring the workers' safety. Regulatory compliance prevails and is the minimum that should be achieved. It is in such situations that the risk estimated by the designer can change, as the original means of protection may no longer be suited to the machine's new use. This is why it is necessary to update the estimated risk and install adapted risk reduction measures. In cases where the integration process, organizational measures, PPE, or training are flawed, these shortcomings may lead to accidents, as noted in section 1. The intent in ISO 12100 is that this feedback from users are provided through participation in standardization works, complaints to manufacturers from their users, complaints from regulatory bodies and so on. In reality, this feedback is not systematic.

Sections 2.1 à 2.5 present the tools available for identifying and estimating machinery risks and various risk management techniques used in other fields. The capabilities and limitations of these tools and techniques are discussed, making it possible, in section 4, to highlight the contribution of the techniques from these other fields to risk management in machinery safety.

2.1 Risk identification tools in machinery safety

The literature on machinery safety and on preventing machinery-related accidents includes regulations (European Machinery Directive 2006/42/EC, 2006), standards (ISO, 2010), accident and research guides and reports (U.S. Department of Labor, 2002; Lupin and Marsot, 2006; Worsell and Ioannides, 2000), and electronic and print tools such as (INRS, n.d.; Paques et al., 2004), the Online interactive Risk Assessment (OiRA) of the European Agency for Safety and Health at Work, and the Machine Guarding eTool from OSHA. These resources make it possible to identify the risks encountered or to learn how to do so. For example, sections 1.3 and 1.5 of Annex I of the European Machinery Directive explain the context in which mechanical, electrical, thermal hazards and others can appear on machinery. For instance, it raises concerns about material that can be ejected from a machine at a very low or a high temperature. The hazards section of the type-C standards and Annex B of ISO 12100:2010, which categorizes machinery-related risks, suggest the possible hazards associated with a machine. Similarly, Paques et al (2004) reminds different kinds of hazards to be considered when performing risk analysis: mechanical, electrical, thermal, chemical, ergonomics, etc. It also suggests hazardous situations (e.g. entrapment), events (e.g. unintended start-up) and damage (e.g. amputation) that could be related to a hazard (e.g. mobile part of a machine). That tool derives from ISO's general principles regarding machinery safety. Computer-based tools such as the OSHA Machine Guarding eTool suggest or teach how to identify hazards associated with machines. However, for the most part, these risk identification tools rely on elements associated with the direct causes of the accidents (e.g. hazards). Lamy et al. (2006) note that indirect causes, such as training and work organization, are not often taken into account by risk identification methods and tools: 42 tools out of 108 according to the study. Some risk analysis methods allow this type of cause to be considered. Then again, one must be able to identify them beforehand. The most frequently used are: preliminary risk analysis (PRA), failure mode, effects and criticality analysis (FMECA), HAZOP-type risk diagram analysis, fault tree analysis (FTA), event tree analysis, and bow tie analysis (Debray et al., 2006).

The method proposed in section 5 will facilitate risk identification by suggesting indirect causes of machinery-related accidents in addition to the direct causes.

2.2 Risk estimation tools in machinery safety

Risk estimation is the step that leads to risk evaluation before making decisions, that is to say, before deciding whether the analyzed risks are acceptable. Various risk estimation tools in machinery safety exist and propose rankings for their parameters; they may be qualitative, semi-qualitative or quantitative. Most tools are risk matrices and use qualitative parameters to estimate the risk (Paques et al., 2007). When the parameters qualitatively translate originally quantitative data, risk matrices can mistakenly assign higher qualitative ratings to quantitatively smaller risks (Cox, 2008). To limit this type of bias, design rules (matrix and other) for risk estimation tools are proposed in (Chinniah et al., 2011) and matrix architecture criteria are suggested by Cox (2008). Despite the presence of qualitative risk levels that impeccably translate quantitative risks, Cox (2008) calls for prudence. He shows that, despite their popularity, matrix tools do not always suggest the right decisions for risk management purposes. The tools' popularity is explained by the apparent simplicity of risk matrices (Ball and Watt, 2013). According to Cox, effective risk management requires quantitative data that go beyond the information provided by a matrix tool. Duijm (2015) concurs: quantified risk assessment is preferred over the application of risk matrices, because it leads to fewer methodological uncertainties and less ambiguity in the results. However, those studies refer to risk estimation in general. Risk estimation in the field of machinery, especially by end users, is still not a widely practiced approach. Accidents related to machinery have shown that absence of risk assessment is mostly to be blamed and not a biased or inaccurate risk estimation (Chinniah, 2015). Accidents have occurred in spite of risk assessment for 2 reasons: incomplete hazard identification and no implementation of risk reduction measures.

Moreover, the range of tools available means that risk estimation results vary depending on the tool used, much as they do depending on who performs the estimation (Hietikko et al., 2011; Ball and Watt, 2013). In the frame of this paper, if several safety practitioners in a company evaluate the risk related to a given hazardous situation, they may come up with different results when using the same tool. This is because they do not perceive the risk in the same way, for various reasons such as each individual's culture and knowledge of the hazardous situation. This variability can be reduced by providing a non-matrix tool that intrinsically quantifies the risks. According to Etherton (2007), although a quantified risk estimation may be more effective, it nonetheless generates a degree of subjectivity (e.g. when the conceptual model on which the estimation is based omits a key scenario leading to an undesirable event), hence the need to take the related uncertainties into account (Abrahamsson, 2002) in order to optimize the risk priority order and the choice of risk reduction measures.

2.3 Quantitative risk assessment

Quantitative risk assessment aims to quantify the consequences associated with a risk and the probability that these consequences will occur. However, more often than not the process consists of quantifying the probability that a loss will occur (Dulac, 2007). In machinery safety, Aneziris et al. (2013) have quantified the risks to which workers are exposed when performing various tasks that bring them into contact with the moving parts of a machine. The contact with moving parts is the undesired event. Specifically, the authors modeled the accident sequence by analyzing 3000 accidents related to the undesired event. The model is based on the bow tie principle. The

undesired event is the center of the bow tie. The undesired event is induced by the failure of proactive safety barriers forming a fault tree on the left side of the bow tie. If the undesired event happens and is not controlled, it will generate further events affecting the population, the environment or the worker. These further events are represented by the failure of reactive safety barriers forming an event tree. The probabilities of the failure of the bow-tie barriers are calculated based on the frequency of their possible states in the accidents analyzed. The calculation is framed by the logical arrangement of the barriers in the bow tie. From the analysis of the quantitative model, Aneziris et al. (2013) suggest a list of protection measures against contact with moving parts of machines. In order to support decision making, a risk reduction weight is assigned to each protection measure. The weight varies according to their efficiency and the task involved on the machine. The uncertainties associated with the risk quantification process described in (Dulac, 2007) are dealt with in another article (Papazoglou et al., 2015).

Demichela and Pirani (2013) undertook a quantitative risk analysis. Their model integrates human and organizational factors into the risk analysis to determine the safety integrity level (SIL, as per IEC 62061) actually required for a safety function of a control system for a hydraulic press. The study results show that the required SIL determined semi-quantitatively based on the standard may differ from the required SIL estimated quantitatively based on human and organizational factors. This quantitative estimation was done by integrating into their Integrated Dynamic Decision Analysis software the task analysis carried out on the hydraulic press and the failure mode probabilities associated with human error, for example, in addition to those associated with the technical components. It is therefore important to include human and organizational factors in the risk assessment in order to minimize this distortion of reality.

It should be noted that quantitative risk assessment is used mainly to quantify the probability that the consequence of an undesirable event will occur. The quantitative aspect of this assessment gives it robustness. However, over time and in order always to have available figures that, to the fullest possible extent, are based in reality, the probabilities should be updated on a regular basis or following the occurrence of an event, i.e. updating risk.

2.4 Updating risk

Updating risk entails the updating of the risk parameters for a system. It is used to update the data describing the risk so as to take into account how the system changes in its environment over time. Updating risk allows the risk dynamic to be monitored. It informs decision makers of the impact of their choices: each decision either moves or does not move the system toward a risk level different from the current one. Depending on the risk level fed back to them, the decision makers adjust their aim, if appropriate, by opting for a safer decision (e.g. Dulac, 2007; Swaminathan and Smidts, 1999). In other cases, by using systemic calculations of the probabilities of causes responsible for the consequences, updating risk alerts the decision makers to hazardous scenarios that develop as the system evolves. In some cases, the calculations use system-related data that are entered real time (Swaminathan and Smidts, 1999; Rathnayaka et al., 2011). In others, data entry is done manually (Mili et al., 2009) and is therefore dependant on the good will of the organizations concerned.

The surveyed studies that deal with risk updating essentially do so in the context of a quantitative risk estimation. The updating is done on the occurrence probabilities for the causes of an

undesirable event, as the following examples show. The industries concerned are mainly aerospace, chemical, petroleum and, nuclear. No studies involving machine safety were found.

In aviation, Di Gravio et al. (2014) used a safety monitoring system based on historic events for air traffic management. Swaminathan and Smidts (1999) vaunted the superiority of dynamic methods over static methods for calculating reliability when in the presence of a system whose condition changes over time depending on the disruptions that it experiences. The authors update the probability and time of occurrence of each of the following aviation-related events: (1) successfully completed flight, (2) aircraft crashing in an inhabited area, and (3) aircraft crashing in a deserted area. This updating helps pilots make decisions during the flight, depending on the time that remains for them to take action to avoid the worst outcomes.

According to natural gas and oil industry experts, in order to effectively manage operational risks and reduce the occurrence of accidents, just assessing risks and establishing preventive and mitigating barriers to control them is not enough, and continuous monitoring of the states of these barriers is required, but no concrete solution is proposed (Saad and Arakaki, 2014). To address this shortcoming, Saad and Arakaki (2014) propose an approach that consists of automatically updating the conditions of a plant's bow tie barriers by running a continuous barrier monitoring system. The goal is to minimize human involvement, which is viewed as less reliable than an automated system. However, barrier conditions that can only be passed on verbally (e.g. organizational deficiencies) will continue to be updated manually.

For their part, Badreddine and Ben Amor (2010) proposed a method based on a Bayesian approach to make bow ties dynamic. They present a case study of a petroleum company. According to the authors, forming recommendations based on a static tool to reduce a risk related to a facility will result in lower quality recommendations because the recommendations will be associated with a facility that changes over time and will thus become outdated, whence the necessity of making the bow tie dynamic by using the actual data that characterize the facility.

Naderpour et al. (2014) proposed a dynamic, human-centred approach to assist process system operators in their decision-making in abnormal situations in order to maintain the safety of a facility as it changes. The approach is based on Dynamic Bayesian Networks and fuzzy logic, with the probabilities related to abnormal situations being updated with expert opinions and new data as they become available.

Drawing on several approaches, Escobar and Lévêque (2014) calculated the probability of a nuclear accident using past observations. The study shows the impact of the Fukushima Dai-ichi accident on the probability of this type of event. One of the approaches consisted of using Bayes' theorem for updating the accident occurrence rate based on expert opinions and observations recorded over 40 years. The study shows the importance of taking a system evolution into account: the effectiveness of risk reduction measures varies with time and this changing effectiveness affects the probability of an accident occurring.

Those studies reveal that risk updating begins by drawing on data that document the past. It continues by recording new observations, which could be called "data reporting." Data reporting is a key step in the experience feedback process.

2.5 Dynamic experience feedback to update risk

Van Wassenhove and Garbolino (2008) define experience feedback (ExF) as a trust process between all system actors: executives, operators, designers, maintenance personnel, and ExF facilitators. ExF consists of rigorously structuring data on past events in order to extract knowledge applicable to new projects (Beler, 2008). These data are essential for sound management of the life cycle, maintenance optimization, and probabilistic safety assessment (Lannoy, 2011).

In machinery safety, ISO 12100:2010 mentions reliability data, accident history, and damage to health as being among the factors to consider when estimating the probability of occurrence of a hazardous event. Accident databases are sources of information that can provide experience feedback. Regulations, laws, and standards can be used to establish sound practices. These sources are deemed sufficient for information capitalization in machinery safety (Beler, 2008). Exploitation (dissemination) of information from data is currently being performed in some workplaces and training centers using either their own database or accident databases. However, the exploitation is not done in a way that takes full advantage of ExF, which, in such cases, is static, limited to the factual circumstances at the time of the accident, and used as a diagnostic tool. In the static approach, knowledge is collected and distributed without being processed "intelligently" (Beler, 2008). The knowledge base is consulted mainly for comparison and diagnostic purposes. One way to derive maximum advantage from it is to make the exploitation intelligent, that is, to allow the ExF to infer new knowledge from the capitalized information. This is the dynamic ExF approach.

Capitalizing information of incidents and accidents is important for preventing risk. Bird's pyramid shows this to be true: the probability that a fatal accident will occur increases with the number of incidents and accidents (Borg, 2002). This points to the importance of focusing on the base of Bird's pyramid in order to minimize the more significant undesirable events. Exploiting ExF will shed light on the causes of accidents and incidents on which action needs to be taken to reduce the risk. However, ExF is not a systemic process. It is dependent on the company's occupational health and safety (OHS) culture, on how dedicated its managers and workers are to reporting incident information instead of hiding it out of fear of reprisal, which has the effect of altering the quality of the causal diagnosis of the incident or accident.

Dynamic ExF is widely used to assist decision-making. Cheng et al. (2013) analyzed 349 major accidents that occurred in the petrochemical industry between 2000 and 2010. They then deduced a decision tree with the aim of obtaining the rules governing the occurrence of this type of accident in order to propose means for finding a solution to it. In public transportation, Hounnou and Parrennes (2014) used probabilistic treatment of experience feedback data (a database covering more than 10 years) to define a model for anticipating events precursory to serious incidents or accidents. The model performs a predictive analysis of these types of event and recommends measures to be implemented in order to block their development and control the risks. In OHS, Verma et al. (2014) extracted frequent patterns from a database of 843 events that occurred in a plant between March 2010 and July 2013. The events were accidents, near misses, and incidents involving material and environmental damage. Expressed as association rules, the patterns represent the knowledge acquired and extracted from the database. The rules are risk management tools: the indicators (which are variables) that make them up show the risk factors

requiring risk reduction measures. Cheng et al. (2012) deduce a decision tree from a database of 1542 construction accidents covering the period from 2000 to 2009. The serious work-related accidents are focused on by the researchers to explain the cause and effect relationships using rules revealed by the decision tree. This will allow safety practices and training programs to be improved in the industry. Beriha et al. (2012) propose a model based on fuzzy rules generated from a statistical database. The data regard accidents and costs invested for prevention in different industries. The rules show the relationship between the degree of investment for prevention (the model input) and the type of potential accident (the model output). The type of accidents the model predicts goes from reversible injury without work stoppage to permanently injury with disabling or fatality.

In the finance world, Hammer et al. (2012) carried out a study to evaluate the credit worthiness of banks. They applied different data mining techniques (ordered logistic regression, support vector machine, and LAD) to a database containing 24 explanatory indicators for the year 2001 describing 800 banks' credit risk ratings. They wanted to generate patterns describing the interaction between the indicators explaining the credit risk ratings. Thanks to these patterns, they could predict the banks' credit risk ratings for the subsequent year based on the indicators' value at that year. Compared to the other techniques tested, they found that the LAD ratings are shown to be the most accurate and most successfully cross-validated. Their approach is generalizable to other types of borrowers by the modification of the set of explanatory indicators and the selection of appropriate data sets for model inference.

In medicine, a research team (Alexe et al., 2003) studied coronary risk prediction applying LAD to a 9454-patient database. They showed the effectiveness of that algorithm in identifying and distinguishing between high and low risk patients. The high risk patients were really few (rare data) compared to the low risk ones. LAD has been used in other applications too, in order to develop diagnostic and prognostic systems in cancer research and pulmonology, risk assessment among cardiac patients (Hammer and Bonates, 2006). Therefore, LAD could be the key to suggesting a method to characterize and distinguish machinery contexts.

Behind each of these dynamic ExF-based decision support models is a database of undesirable events. However, it should be noted that, to learn from ExF, it is important not to rely solely on negative events (accidents). Positive events (sound practices) should also be capitalized to support learning. By updating information reporting, it is possible to update the rules (knowledge) that explain the occurrence or non-occurrence of undesirable events. This makes it possible to update the risks of the evaluated system as well as the measures for reducing them.

3. LAD method

LAD is a data mining pattern recognition and classification technique (Boros et al., 2000). It is an optimization combinatorial process based on Boolean logic. LAD consists of browsing any numerical database in order to extract information in the form of patterns that is hidden naturally in the database. The extracted patterns are sets of hidden conditions that characterize and distinguish specific classes of the phenomenon under study. The database is made of a list of n observations described by m factors or indicators. In the context of machinery safety, each observation characterises an accident and the indicators represent risk factors or accident causes. Each indicator has a safety or danger-related value describing the context in which each accident

occurred. Accordingly, each observation is a vector which components are the indicators' values. In order to extract the patterns, data are binarized by transforming the non-binary indicators' values into binary attributes. Then, the patterns are generated by using one or a combination of the following techniques: simple enumeration, combinatorial optimization, or heuristic techniques. As any machine learning technique, LAD is applied in two phases: the learning phase in which the patterns are generated, and the validation or testing phase in which the quality of the generated patterns and their ability of capturing the characteristics of the phenomena under study is tested and validated. Figure A2 presents the main steps of LAD method. It, clearly shows the link between LAD and dynamic ExF: the process learns from the information capitalized in a database and ends with the knowledge inferred from these information. The classification rule formation, also known as the theory formation, is the adequately chosen collection of patterns based on the patterns classification accuracy (Kim and Choi, 2015; Boros et al., 2000).

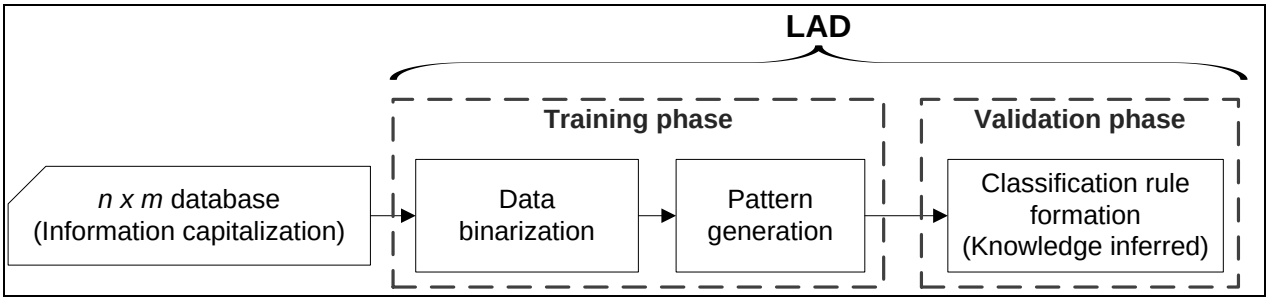


Figure A2: Flow chart illustrating the steps of the LAD method

4. Contribution of: Dynamic Experience Feedback and Quantitative Risk Estimation

The review of tools and techniques presented in section 2 sheds light on the capabilities and limitations of the tools used in machinery safety to identify or estimate risk. It also made it possible to identify the risk management techniques used in other fields: aviation, petroleum, nuclear and process industries, OHS in general, finance, medicine. Integrating dynamic ExF into quantitative risk estimation can only but improve these capabilities. Integrating LAD into ExF brings more efficiency to the machinery risk management process. Tables A1 and A2 present a summary of the tools reviewed and the contribution of dynamic ExF and of quantitative risk estimation. The tables are followed by the explanation of the significance of LAD applied to machinery safety.

Table A1: Risk identification in machinery safety – Contribution of dynamic ExF

Tool	**Risk identification capabilities/limitations of the tool**	**Contribution of dynamic ExF**
Type-C standards and ISO 12100:2010 standard, electronic and paper tools	• Direct causes are suggested • No proposal of indirect causes • Static tools → decisions outmoded in the future	• Preponderant direct and indirect causes are determined • Dynamic method: causes determined by inference from capitalized information and identification of new risks during data reporting
Guides, accident and research reports	• Direct and indirect causes are suggested • Logical combination of causes not always specified • Static tools	• Logical combination of causes determined by inference from capitalized information • Dynamic method
PRA, FMECA, HAZOP, FTA, "Bow tie"	• Risk identification difficult to initiate • Direct and indirect causes can be taken into account but are not suggested • Static tools	• Preponderant direct and indirect causes deduced from a database • Dynamic method

Table A2: Risk estimation in machinery safety – Contribution of quantitative risk estimation and dynamic ExF

Tool	**Risk estimation capabilities/limitations of the tool**	**Contribution of quantitative risk estimation**	**Contribution of dynamic ExF**
Risk estimation tools (essentially matrices)	• Difficult allocation of probability values for the occurrence of damage or the hazardous event • Ranking of risks possible • Bias: under- or over-estimated risks due to the tool's qualitative or semi-qualitative aspect • Biased ranking → decision not always appropriate for managing risks • Static tools	• Allocation of probability values facilitated by the availability of facts related to undesirable events • Less variability: fact-based risk estimation • Robustness conferred by the calculated probability of the undesirable event and its causes • Risk ranking optimized by robustness → more effective management • More objective decision-making assistance due to lower variability	• Dynamic method: ○ Risk updating: probability calculation updated by data reporting ○ Possible monitoring of the frequency of the causes of the undesirable event to correct them in time and prevent the accident ○ Decision making and adjustment support as the system changes for more reality-based decision-making

As shown in sections 2.5 and 3, data mining techniques enable the inference of knowledge, therefore participate to dynamic ExF. The added value of integrating LAD into dynamic ExF is explained as:

- LAD has better prediction rates than other data mining techniques (e.g. decision trees, neural networks) (Yacout, S., 2010);
- LAD can perform on rare data. Contrary to the usual data mining techniques that need a lot of data in order to extract their special features, LAD is able to generate rules not requiring a large amount of data. Indeed, its performance is shown in various studies, such as the ones explained at the end of section 2.5;
- LAD is appropriate for targeted prevention (targeted risk identification): contrary to the other data mining techniques that only characterises classes, the patterns LAD generates also distinguish the classes. Therefore, LAD is suitable for developing a machinery risk management method under different contexts of use of machinery. For instance, the safety practitioners will be able to distinguish various classes of events like: accidents vs. near-misses (non-accidents), maintenance accident vs. production accidents, machine "A"-related accident vs. machine "B"-related accidents, industrial sector "X" accidents vs. industrial sector "Y" accidents. Then, the safety practitioners or any other user will be able to deduce the risk reduction measures specific to each class of event.
- LAD allows generic risk management methods (just like in Hammer et al., 2012): the safety practitioners can choose themselves the observations and indicators regarding the classes of events they want to characterize and distinguish. The indicators and the observations chosen will shape the context of use of the machine no matter what its type or industrial sector is;
- LAD enables risk monitoring: by updating the database, LAD will generate new patterns reflecting the actual state of use of the machinery.

Based on the observations summarized in Tables A1 and A2, as well as the significance of LAD, the three challenges mentioned in section 1 will be discussed and the proposed method for designing a dynamic risk identification and quantitative estimation tool in machinery safety will be explained. An example in section 5.2 illustrates the relevance of such a tool.

- *How to help safety practitioners efficiently identify and estimate machinery-related risks?*

This question can be broken into two parts: (1) how to help safety practitioners efficiently identify machinery-related risks, and (2) how to help them quantitatively estimate these risks?

Table A1 shows that some tools used in machine safety provide safety practitioners with the basic information necessary to begin their analysis of the risk but by suggesting possible direct causes of accidents. Other tools suggest no possible causes of accidents or risk. However, their structure does suggest a procedure to follow in order to identify and analyze the risks; this is the case for the PRA, FMECA, FTA, HAZOP, and bow ties. Embarking on analysis of a machine's risk means not only determining its operating conditions but also identifying its risks or those of the associated work situation. Furthermore, identifying risks is not an easy task or, at least, is not intuitive. It would thus be relevant for safety practitioners to have a tool that suggests the main sources of machinery-related risks, in particular the indirect causes (weak signals) of accidents, in order to take prior preventive action. As Table A1 show that inferring knowledge from databases can reveal the preponderant direct and indirect causes of accidents and incidents, using a database of machinery-related accidents as well as good practices would be relevant for assisting with risk identification. Moreover, efficient risk identification will be possible using LAD, thanks to:

- its ability for targeted prevention;

- the patterns it generates: the indicators combination tells what potential causes or risk factors to verify during an audit in order to detect if an accident or near-miss is can occur.

A LAD-based dynamic ExF is therefore desirable for helping safety practitioners to efficiently identify the risks inherent in machinery.

As for risk estimation (Table A2), the difficulty of attributing from scratch a probability to an event justifies the need for the risk to be estimated quantitatively. This consists of allocating, in a sustained manner and based on real-life data, the probability of harm or an undesirable event occurring. A person's experience and knowledge—or feelings—influence the value or descriptor that he or she attributes to the probability. To minimize such subjectivity, it is preferable to use or combine event histories (i.e. real-life data) with these opinions. As Nix (2012) notes, allocating probabilities to an event is a challenge. To overcome it, he suggests using reliable numeric data. The robustness of quantitative risk estimation makes it the preferred method for efficiently estimating machinery-related risks. Once again, to reduce the epistemic uncertainties relative to the probabilities assigned by expert opinions, use of an event database identical to the one described earlier is recommended. Bayesian inference is a tool of choice for dealing with such uncertainties. Accordingly, quantitative risk estimation supported by an event database built with ExF is the recommended means for helping safety practitioners efficiently estimate the risks of their machinery.

- *How to help safety practitioners prioritize risk reduction measures and, at the same time, keep their machines from being sealed by an OHS inspector?*

When speaking of prioritising, it is understood that the support is for decision-making. In this case, the aim is to help safety practitioners prioritize risk reduction measures and avoid inspectors' seals. To accomplish this, the paper suggest that the causes identified using patterns extracted from the accident and good practices database be ranked in decreasing order of probability of occurrence using a decision support method. The risk reduction measures will be implemented by the safety practitioner based on the priority ranking of the identified causes, from the most probable to the least. Drawing on the work of Aneziris et al. (2013), the tool could propose a list of risk reduction measures with risk reduction weights assigned to each. Risk reduction measures considered indispensable by inspectors would be given the highest weight. For example, knowing that the top reason that leads inspectors to place a seal on a machine is accessibility to danger zones, the highest weight will be assigned to guards. If the machine does not have guards, the tool will notify the safety practitioner that an inspector risks placing a seal on it. Lastly, with a decision support tool that ranks the possible causes of accidents according to their probability of occurrence (quantitative estimation of risk), safety practitioners will know how to prioritize the risk reduction measures for the evaluated machine and avoid an inspector's seals.

- *How to help safety practitioners monitor risk progression?*

Tables A1 and A2 show that, in machinery safety, risk estimation tools and methods are more static than dynamic. They do not consider how a machine evolves in its physical or organizational environment. To ensure the changes in a machine's risks are monitored requires a dynamic decision support tool. Dynamic ExF allows the probabilities of accidents occurring and their causes to be updated. This risk updating is made possible by data reporting, meaning the record of accidents and good practices is updated with the arrival of each new machine-related

event. An inference algorithm based on a method like Bayesian inference can be used to update the calculated probabilities each time new information or data is added to the accident and good practices database. In the frame of this research, the data reporting is manual. It therefore presupposes that a safety culture exists in the organization. Such a culture will ensure that the information (new abnormalities and good practices) is reported for learning purposes in order to make informed decisions. And here one sees what experience feedback is and what its limitations are. Although useful for learning about the past, the process is dependent on the trust established between the workers and their employers as well as the willingness of organizations to maintain the process. Thus, risk updating, which consists of dynamically estimating risk (updating probabilities), is related to experience feedback by the data reporting that it requires. This is how safety practitioners will be helped in monitoring the progression of the risks related to their machines.

5. The Proposed Methodology

5.1 Design process

Figure A3 diagrams the proposed general process for designing a tool based on integrating dynamic ExF (using LAD) into quantitative risk estimation. The process has been developed with the aim of producing a computer-based tool. The process is broken down into three steps. Each step includes the following sections: input, processing and output. Arrows represent the data flow from one block to the next.

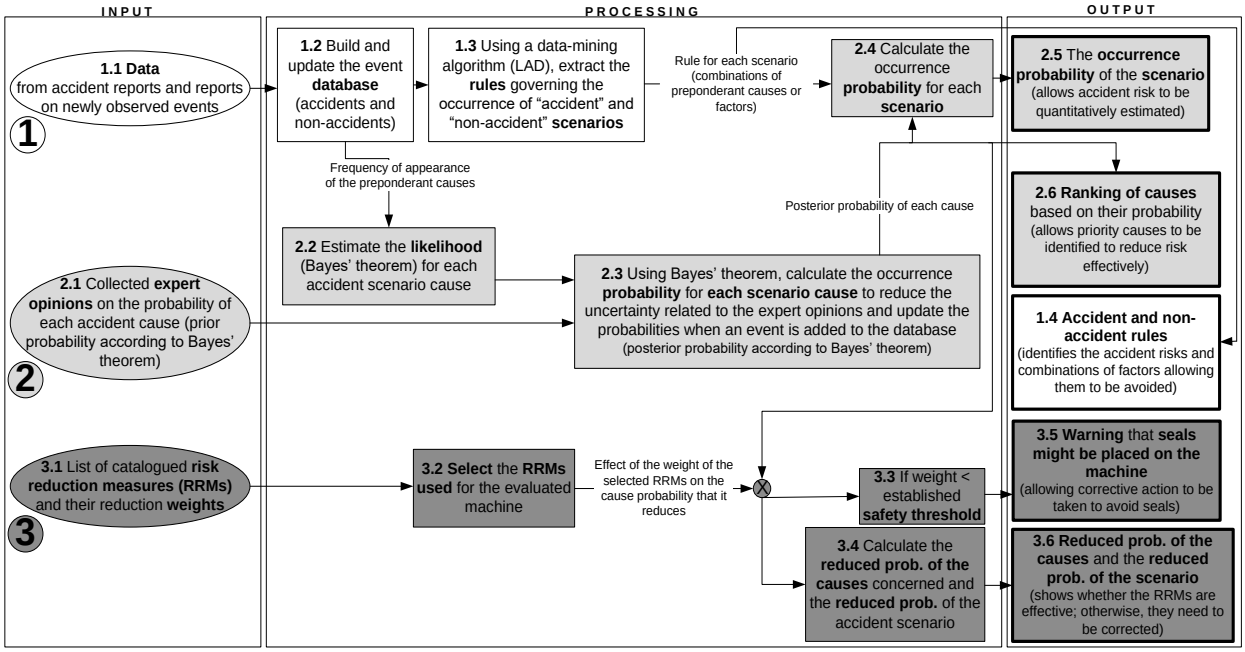


Figure A3: Flow chart of the proposed design method for the dynamic decision support methodology

Step 1 consists of blocks 1.1 to 1.4. It identifies the preponderant direct and indirect causes of accidents and their interaction in the form of logical rules ("accident scenarios"). It also makes it

possible to understand the combinations of causes that, if they fail to materialize, allow the accident to be avoided ("non-accident scenarios"). At this stage, ExF consists of collecting historical data about events (accidents and non-accidents) from which to draw knowledge, i.e. the rules or sets of patterns. It is in block 1.1 that new events are added over time to update the database and infer new rules. Regular audits on safety of machinery can be used to verify that, for example, (i) hazardous zones on machinery are safeguarded, (ii) safety devices are not defeated, (iii) guards are placed back after maintenance, (iv) access to and around machines are safe, (v) workers have been trained properly by asking them basic questions, (vi) lockout procedures are applied, (vii) personnel protective equipment are worn, (viii) preventive maintenance is carried out and (ix) sub-contractors are following safe working procedures and to register any abnormalities or deviations. Investigations following circumstantial events such as near misses or accidents can be made. Results from audits and investigation will update the database. Thus, LAD will infer new knowledge, i.e. will generate updated patterns with a new occurrence frequency. This will update the risk identification (step 1) as well as its estimation (step 2), followed by risk ranking. Thus, contrary to the END block of Figure A1 that suggests ending the risk management process when the risk is adequately reduced, the support tool from Figure A3 encourages to monitor the machinery-related risks continuously. Consequently, this tool should be used in a continuous improvement environment since the possible causes to accident may change through time and use.

Step 2 consists of blocks 2.1 to 2.6. At this stage, ExF consists of collecting expert opinions formulated during the audit of the machine in operating mode. These opinions characterise prior information about the probability of occurrence for each accident cause. Then by applying Bayes' theorem, the likelihood of appearance of the accident causes identified in step 1 are updated with the prior expert's opinions to quantitatively estimate the posterior risk. This makes it possible to predict dynamically the future behaviour of the system based on the probability of occurrence of the causes. When the database is updated in step 1, the causes are too, which automatically updates the probability of occurrence of both the causes and the scenarios. The ranking of the causes is subsequently updated.

Step 3 is comprised of blocks 3.1 to 3.6. At this stage, ExF involves learning the effect of the installed risk reduction measures (RRMs) on the probabilities initially calculated in step 2. The calculation of the effect on the probability is possible due to the tool user selecting the reduction weight of each RRM applied to the machine. This allows the effectiveness of the RRMs used over time or added to be evaluated and monitored (via checks and inspections). In contrast to step 2, where the changes in the accident probabilities can be monitored by adding events, step 3 makes it possible to monitor any changes in the effect that the RRMs have on the estimated risk and to take corrective action as required. This helps with controlling the residual risks of the machine and preventing accidents and avoiding the costs related to seals. Indeed, if the weight of one RRM is less than the minimum weight required by the OHS inspector, the tool will warn the user about the possibility of the inspector sealing the machine.

The patterns generated by LAD have a direct impact on part 1 of the design process suggested. Consequently, LAD has an impact on part 2 and 3, since both depend on the first part. In part 1, no matter how few the data may be, LAD will be able to find at least 1 pattern that both characterizes and distinguishes the database classes, as long as the classes have no contradiction within their observations (contradiction means: if one class of events has an observation

described with the same values of indicators as an observation from another class, it is called a contradiction). LAD facilitates risk identification in a targeted way and determination of risk occurrence frequency.

5.2 Relevance and feasibility demonstration

In this section, the proposed three-step methodology is explained to demonstrate its relevance and feasibility. To do so, a real example from the Quebec sawmill industry involving two accidents on the same belt conveyor and under similar circumstances is discussed. Figure A4 places the problem in context: nip points were involved in both accidents (Brulotte and Roberge, 2006). If the company had learned from accidents on belt conveyors as well as its first accident (2002), the second (2005) could have been avoided.

- *Step 1*

As mentioned in section 2.5, dynamic ExF involves inferring knowledge from capitalized information. In this example, the capitalized information consists of a database of belt-conveyor-related accidents analyzed from investigation reports (steps 1.1 and 1.2 of Figure A3). For example, the accidents can be divided into two classes: maintenance-related accidents and production-related accidents. Then, the knowledge is inferred in the form of set of patterns (output of step 1.3 of Figure A3) by LAD. Since the two accidents involved in this example happened during maintenance activities, one possible pattern for maintenance-related accidents is given as an example. That pattern could involve the following combination of indicators: the working environment surrounding a belt conveyor is poor, and a hazard zone having a running drum is accessible, and a part of the worker's body is inside the hazard zone. Based on the characteristics of the pattern, at the time of the first accident, the safety practitioner could have anticipated that another accident could happen due to the fact that the pattern characteristics are similar to several causes involved in the 2002 accidents. For example, during both accidents, the area was cluttered, which encouraged the worker to take a shortcut, namely to cross the lower strand of the belt conveyor. Moreover, the nip points were accessible because a guard or safety device was never installed on the conveyor. Note that the access to the nip points was not necessary for the task involved in both accidents. The poor working environment cited in the pattern is an indirect cause of both accidents, whereas the two other causes cited are direct causes. As early as 2002, if the sawmill had applied such LAD method, these indicators should have alerted the company's safety practitioners. The pattern would have warned them that accessible nip points are important hazards to be eliminated or significantly reduced. This warning would have helped them identify the related hazard zones: the drums of the conveyor (Figure A4). Eliminating the direct and indirect causes of the first accident would have helped prevent the second from occurring.

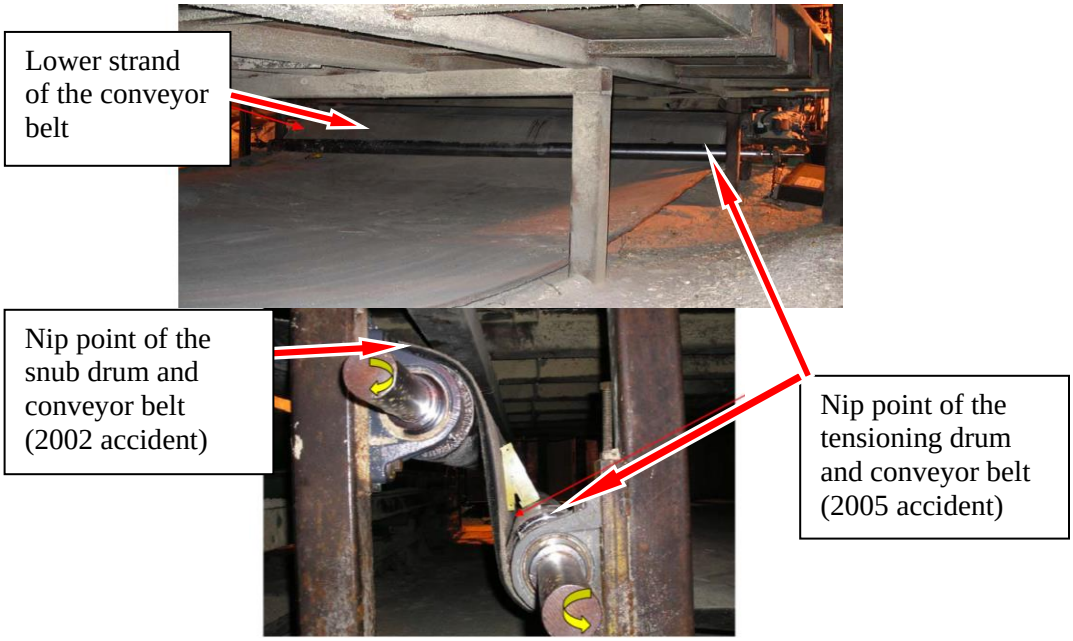


Figure A4: Front and side views of the lower strand of the conveyor belt involved in the accidents (taken from the appendix of report (Brulotte and Roberge, 2006))

- *Step 2*

Each company that has the tool will keep the database up-to-date with its own accidents and good practices. This means that the updated probabilities will concern only that company. In this example, if the safety practitioner had recorded the 2002 accident in the database, the occurrence probabilities for the accident causes would have been updated thanks to the Bayesian inference mentioned in step 2 of Figure A3. The knowledge inferred at step 2 is the update of the probability considering the new occurrence of a circumstantial event. Then, by monitoring the changes in the occurrence probabilities, the safety practitioner would have noticed that the probabilities had increased. In other words, updating the risk might have convinced the safety practitioner to take action to reduce the probabilities based on the methodology's ranking of the causes. This demonstrates the importance of monitoring the frequency or probability of accident causes and finding a solution for them in a timely manner (Table A2) instead of monitoring the changes in the accident frequency and finding a solution once harm has been done.

- *Step 3*

After the 2005 accident, an inspector barred use of the conveyor due to the accessibility of the nip points involved in the two accidents. If the company had been using the tool since 2002, it would have warned of the possibility that an inspector would place seals on the conveyor due to the lack of nip point safeguards. After all, an accessible hazard is the main reason why inspectors put seals on a machine. This would have prevented not only the accident but also the costs associated with a production stoppage.

After the 2002 accident, the employer made incidental corrective improvements to the conveyor. These did not address the three main causes of the accident. By selecting the risk reduction

measures used for the machine in the tool, the company's safety practitioner would have seen that they had no effect on the three main causes of the accident, that the probability associated with the main causes remained the same. This would have alerted the safety practitioner that the risk had not been reduced.

Nonetheless, the proposed process remains dependant on the willingness of the organization involved. The investigation report (Brulotte and Roberge, 2006) shows that, without this willingness, the database will not be updated and consequently neither will the risks. The report also notes that a data reporting program existed at the company. However, in practice, reporting was done only occasionally, contrary to the literature, which recommends regular reporting. In addition, the company has a prevention program but does not maintain it. This example demonstrates the importance of instilling a culture of safety in employers and workers alike. In companies, this is a very important value not only for combating resistance to change and enabling data reporting but also for maintaining ExF. Learning from observed errors and good practices is fundamental to every improvement process and therefore must be used to avoid the occurrence and repetition of undesirable events. The layered audit principle explained by Lindquist (2011) is one solution for instilling a culture of safety.

6. Conclusion

A literature review was undertaken to learn what is already available to help machinery safety practitioners. An examination of risk identification tools in machinery safety led the researcher to focus on risk estimation and what exists on the subject. Research made it clear that, despite the related uncertainties, quantitative risk estimation is the most accurate approach. To obtain a more realistic picture and avoid recommending outdated risk reduction measures, it is necessary to track any changes in the risk. This is where risk updating—or risk dynamic—can help. The literature review showed that using dynamic experience feedback is an interesting way to update risk. The results from the literature review made it possible to suggest a generic method for designing a decision support tool that integrates a LAD-based dynamic experience feedback into quantitative risk estimation in order to identify and update the risks. A novelty of the suggested methodology is this integration of LAD for machinery-related accident prevention, which has not been covered in previous studies. LAD will be useful by guiding safety practitioners with the key indicators to pay attention to, in order to evaluate if the current state of their machine's context (organizational system, environment, machine and workers using the machine) is likely to lead to an accident or not. Also, it will allow the safety practitioners to update the database with newly declared near-misses or accidents or with audit declaration of the machine abnormalities or of its poor environment or organizational system.

The proposed methodology helps improve the risk identification and estimation steps of the ISO 12100:2010 risk management process for machinery safety (Figure A1). As the study showed, it adds value to the tools available in the field for identifying and estimating risk. The methodology demonstrates that risk management does not stop when the risk is adequately controlled. Contrary to what the END box in Figure A1 implies, risk management should never end but must constantly remain in "monitor" mode. It is an endless loop that is renewed with each input of reported data in order to generate updated knowledge about the risks involved.

Basing themselves on specific sources, Caputo et al. (2013) say the context of manufacturing plants and job shops often makes sophisticated probabilistic analysis and detailed risk modeling not applicable, so that simplified methods based on check list (OSHA, 2007; Kjellen et al., 1990) and risk matrix are most often used (European Commission, 2008). However, in section 2.2, the pitfalls of matrices and how quantitative methods are more efficient were shown. In the proposed tool, step 2, which calculates the probabilities, will be a black box for safety practitioners. Only the results of the calculations will be accessible to them: the ranking of the accident causes with their associated probability. This will make the tool easy to use notwithstanding the fact that it is based on a probability calculation. Unlike what would have been the case had a matrix been used to perform a risk appreciation, the safety practitioner will automatically and quickly know which potential accident causes should be reduced and in which order.

The ability of the tool to warn of the possibility that an inspector will affix seals is a factor that encourages its use, as is the fact that it allows the evolution of accident causes to be monitored and corrected in a timely manner. This avoids having to face the human and financial costs associated with seals and accidents. It should be noted that the only points where the tool requires user input are steps 1 and 3: in step 1 to update the database and in step 3 to verify the risk reduction measures used for the machine by checking them off a simple, computerized checklist. Steps 2 and 3 of the tool do not complicate the situation for safety practitioners; on the contrary, they make it easier. The same is true for step 1. Although updating the database may appear onerous, it is actually easier and faster to do than performing a risk analysis in order to begin by identifying from scratch, by oneself, risks that are not always obvious. The decision support tool will already have done much of the work with the combinations of causes that it proposes and the calculated probabilities that show the effect of the risk reduction measures. A way of making the safety practitioners' task easier, irrespective of a company's size, is to motivate workers and make them aware of their responsibilities so that they detect and report on irregularities.

The proposed methodology appears to be a reasonable means for managing the risks related to a machine. The relevance and feasibility of the three-step methodology have been discussed using a real example involving two accidents that occurred under similar circumstances on the same machine in a sawmill. However, comprehensive results at each step of the methodology should be determined, analyzed and validated. To do so, mathematical models, testing and validation of each step should be developed and explained, which cannot be presented simultaneously in this paper. At least two other scientific papers should be focused on these main steps.

Acknowledgment

The contribution of the IRSST and Polytechnique Montréal to this study is gratefully acknowledged.

References

Abrahamsson, M., 2002. Uncertainty in quantitative risk analysis – characterisation and methods of treatment (Report n° 1024). Department of Fire Safety Engineering, Lund, Sweden.

Alexe, S. et al., 2003. Coronary Risk Prediction by Logical Analysis of Data. Ann. Oper. Res., 119, 15-42.

Aneziris, O.M. et al., 2013. Quantification of occupational risk owing to contact with moving parts of machines. Saf. Sci., 51, 382-896.

Badreddine A., Ben Amor N., 2010. A New Approach to Construct Optimal Bow Tie Diagrams for Risk Analysis. In: 23rd International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, Cordoba, Spain, 595-604.

Ball, D.J. and Watt, J., 2013. Further Thoughts on the Utility of Risks Matrices. Risk Anal., 33(11), 2068-2078.

Beler, C., 2008. Modélisation générique d'un retour d'expérience cognitif – Application à la prévention des risques. (Thesis, Université de Toulouse, France).

Beriha, G.S., Patnaik, B., Mahapatra, S.S., Padhee, S., 2012. Assessment of safety performance in Indian industries using fuzzy approach. Expert Syst. Appl., 39, 3311-3323.

Blaise, J.-C., Daille-Lefèvre, B., Lupin, H., Marsot, J., Wélitz, G., 2012. Sécurité des équipements de travail – Prévention des risques mécaniques (Guide n° ED 6122). Institut national de recherche et de sécurité (INRS), Vandœuvre-lès-Nancy, Lorraine, France.

Borg, B., 2002. Predictive Safety from Near Miss and Hazard Reporting.

Boros, E., Hammer, P. L., Ibaraki, T., Kogan, A., Mayoraz, E., Muchnik, I., 2000. An Implementation of Logical Analysis of Data. IEEE Trans. Knowl. Data Eng., 12 (2), 292-306.

Brulotte L., Roberge G., 2006. Accident mortel survenu à un travailleur le 9 décembre 2005 à l'enrtreprise Henri Radermaker et Fils inc. 1340, route 117 à Rivière-Rouge (Report n° EN-003614), Quebec, Canada.

Caputo, A.C., Pelagagge, P.M., Salini, P., 2013. AHP-based methodology for selecting safety devices of industrial machinery. Saf. Sci., 53, 202-218.

Cheng C.-W., Leu S.-S., Cheng Y.-M., Wu T.-C., Lin C.-C, 2012. Applying data mining techniques to explore factors contributing to occupational injuries in Taiwan's construction industry. Accid. Anal. Prev., 48, 214-222.

Cheng C.-W., Yao H.-Q., Wu T.-C., 2013. Applying data mining techniques to analyze the causes of major occupational accidents in the petrochemical industry. J. Loss Prev. Process Ind., 26, 1269-1278.

Chinniah, Y., 2015. Analysis and prevention of serious and fatal accidents related to moving parts of machinery. Saf. Sci., 75, 163-173.

Chinniah, Y., Gauthier, F., Lambert, S., Moulet, F., 2011. Experimental analysis of tools used for estimating risk associated with industrial machines (Report n° R-697). Institut de recherche Robert-Sauvé, Montreal, Quebec, Canada.

Cox, L.A., 2008. What's Wrong with Risk Matrices? Risk Anal., 28(2), 497-512.

CSST (Commission de la santé et de la sécurité du travail du Québec), 2014. Rapport annuel de gestion 2013 (Report n° DC 400-2032-7). CSST, Montreal, Quebec, Canada.

Debray, B., Chaumette, S., Descourière, S., Trommeter, V., 2006. Méthodes d'analyse des risques générés par une installation industrielle (Report n° INERIS-DRA-2006-P46055-CL47569). Institut national de l'environnement industriel et des risques (INERIS), Verneuil-en-Halatte, Oise, France.

Demichela, M., Pirani, R., 2013. Human Factor Analysis Embedded in risk Assessment of Industrial Machines: Effects on the Safety Integrity Level. Chem. Eng. Trans., 33, 451-456.

Di Gravio, G., Mancini, M., Patriarca, R., Costantino, F., 2014. ATM Safety Management: Reactive and Proactive Indicators - Forecasting and monitoring ATM overall safety performance. In: Fourth SESAR Innovation Days, 1-8.

Duijm, N.J., 2015. Recommandations on the use and design of risk matrices. Saf. Sci., 76, 21-31.

Dulac, N., 2007. A framework for Dynamic Safety and Risk Management Modeling in Complex Engineering Systems. (Thesis, MIT, USA).

Escobar R., L., Lévêque F., 2014. How Fukushima Dai-ichi core meltdown changed the probability of nuclear accidents? Saf. Sci., 64, 90-98.

Etherton, J.R., 2007. Industrial machine systems risk assessment: a critical review of concepts and methods. Risk Anal., 27(1), 71-82.

European Commission, 2008. Risk Assessment Guidelines for Non-food Consumerproducts (Technical Report). Draft.

European Machinery Directive 2006/42/EC, 2006.

Flaspöler, E. et al, 2010. The human-machine interface as an emerging risk (Report n° TE-80-10-196-EN-N). European Agency for Safety and Health at Work.

Gadd, S., Keeley, D., Balmforth, H., 2003. Good practice and pitfalls in risk assessment (Report n° 151). Health & Safety Laboratory, Norwich, UK.

Gardner, D., Cross, J.A., Fonteyn, P.N., Carlopio, J., Shikdar, A., 1999. Mechanical equipment injuries in small manufacturing business. Saf. Sci., 33, 1-12.

Hammer, P.L., Bonates, T.O., 2006. Logical Analysis of Data – An overview: from combinatorial optimization to medical applications. Ann. Oper. Res. 148, 203-225.

Hammer, P.L., Kogan, A., Lejeune, M.A., 2009. Reverse-engineering country risk ratings: a combinatorial non-recursive model. Ann. Oper. Res., 188, 185-213.

Hammer, P.L., Kogan, A., Lejeune, M.A., 2012. A logical analysis of banks' financial strength ratings. Expert Syst. Appl., 39, 7808-7821.

Healey, N., 2006. Analysis of RIDDOR Machinery Accidents in the UK Printing and Publishing Industries 2003-2004 (Report n° HSL/2006/83). Health & Safety Laboratory, Derbyshire, UK.

Hietikko, M., Malm, T., Alanen J., 2011. Risk estimation studies in the context of a machine control function. Reliab. Eng. Syst. Saf., 96, 767–774.

Hounnou L., Parrennes F., 2014. Anticiper l'évolution des précurseurs de danger par le développement d'une fonction prédictive. In: 19ème Congrès Lambda-Mu Maîtrise des risques et sureté de fonctionnement, Dijon, France, 1-8.

INRS (Institut national de recherche et de sécurité). Mecaprev - Bibliothèque de solutions de prévention des risques. From: https://machines-sures.inrs.fr/mecaprev/pages/avantpropos.seam?cid=2744

ISO (International Organization for Standardization), 2010. Safety of machinery - General principles for design - Risk assessment and risk reduction. ISO 12100. International Organization for Standardization, Geneva, Switzerland.

Kim, H.H., Choi, J.Y., 2015. Hierarchical multi-class LAD based on OvA-binary tree using genetic algorithm, Expert Syst. Appl., 42, 8134-8145.

Kjellen, U., Rundmo, T., Sandetory, H., Sten, T., 1990. Safety analysis of manual tasks in automatic production systems – implications for design. Accid. Anal. Prev. 22(5), 475-486.

Lamy P., Levrat E., Paques J.-J., 2006. Méthodes d'estimation des risques machines : analyse bibliographique. In: 15ème Congrès Lambda-Mu Maîtrise des risques et sûreté de fonctionnement, Lille, France, pp. 1-8.

Lannoy, A., 2011. Retour d'expérience technique. In: Management de la sécurité (SE 1 041v2). Techniques de l'ingénieur, p. 1-22.

Lindquist, R., 2011. The secret to sustainment. Qual. Prog., 44(8), 40-45.

Lupin, H., Marsot, J., 2006. Sécurité des machines et des équipements de travail – Moyens de protection contre les risques mécaniques (Guide n° ED 807). Institut national de recherche et de sécurité (INRS), Vandœuvre-lès-Nancy, Lorraine, France.

Mc.Manus, N., 2014. Incident Records Understanding the Past to Prevent Future Hazardous Energy Incidents. Prof. Saf., 34-43.

Mili, A., Bassetto, S., Siadat, A., Tollenaere, M., 2009. Dynamic risk management unveil productivity improvements. J. Loss Prev. Process Ind. 22, 25-34.

Naderpour, M., Lu, J., Zhang, G., 2014. A situation risk awareness approach for process systems safety. Saf. Sci., 64, 173-189.

Nix, D., 2012. The probability paradox: Risk assessment and machine safety. Manufacturing Automation. From: http://www.automationmag.com/factory/2828-the-probability-paradox-risk-assessment-and-machine-safety

OSHA (Occupational Safety and Health Administration), 2007. Safety and Health at Work is Everyone's Concern : Risk Assessment Tool. From: http://osha.europa.eu/en/campaigns/hwi/about/material/rat2007 (accessed on July 17th, 2012).

Papazoglou, I.A., Aneziris, O., Bellamy, L., Ale, B.J.M., Oh, J.I.H., 2015. Uncertainty Assessment in the Quantification of Risk Rates of Occupational Accidents. Risk Anal., 1-26.

Paques, J.J. et al., 2004. Sécurité des machines : phénomènes dangereux, situations dangereuses, événements dangereux, dommage (Guide n° DC 900-337-1PDF (06-11)). Commission de la santé et de la sécurité du travail du Québec (CSST) et Institut de recherche Robert-Sauvé, Montreal, Quebec, Canada.

Paques, J.-J., Gauthier, F., Perez, A., 2007. Analysis and Classification of the Tools for Assessing the Risk Associated With Industrial Machines. Int. J. Occup. Saf. Ergon., 13(2), 173-187.

Programme on Safety and Health at Work and the Environment (SafeWork), 2013. Training Package on Workplace Risk Assessment and Management for Small and Medium-Sized Enterprises. International Labour Organization, Geneva, Switzerland.

Rathnayaka, S., Khan, F., Amyotte, P., 2011. SHIPP methodology: Predictive accident modeling approach. Part II: Validation with case study. Process Saf. Environ. Prot., 89, 75-88.

Saad S., Arakaki R., 2014. An Event Processing Architecture for Operational Risk Management in an Industrial Environment. In: Proceedings of the 8th ACM International Conference on Distributed Event-Based Systems, New-York, USA, 213-224.

Safe Work Australia, 2014. Work-related traumatic injury fatalities, Australia 2013. Safe Work Australia.

Swaminathan, S., Smidts, C., 1999. The Event Sequence Diagram framework for dynamic Probabilistic Risk Assessment. Reliab. Eng. Syst. Saf. 63, 73-90.

U.S. Department of Labor, 2002. Job hazard Analysis (Guide n° OSHA 3071). Occupational Safety and Health Administration (OSHA), U.S.A.

Van Wassenhove, W., Garbolino, E., 2008. Retour d'expérience et prévention des risques – Principes et méthodes. Lavoisier.

Verma A., Khan S.D., Maiti J., Krishna O.B., 2014. Identifying patterns of safety related incidents in a steel plant using association rule mining of incident investigation reports. Saf. Sci., 70, 89-98.

Worsell, N., Ioannides, A., 2000. Machinery risk assessment validation literature review (Report n° HSL/2000/18). Broad Lane, Sheffield, England.

Yacout, S., 2010. Fault Detection and Diagnosis for Condition Based Maintenance Using the Logical Analysis of Data. In: 2010 40th International Conference on Computers and Industrial Engineering (CIE), IEEE, pp. 1-6.

ANNEXE B – ARTICLE 2

"APPLICATION OF LOGICAL ANALYSIS OF DATA TO MACHINERY-RELATED ACCIDENT PREVENTION BASED ON SCARCE DATA"

Source : http://dx.doi.org/10.1016/j.ress.2016.11.015

Reliability Engineering and System Safety 159 (2017) 223–236

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

journal homepage: www.elsevier.com/locate/ress

Application of logical analysis of data to machinery-related accident prevention based on scarce data

Sabrina Jocelyn[a,b,*], Yuvin Chinniah[b], Mohamed-Salah Ouali[b], Soumaya Yacout[b]

[a] Institut de recherche Robert-Sauvé en santé et en sécurité du travail (IRSST), 505 de Maisonneuve Blvd. West, Montreal, Quebec, Canada H3A 3C2
[b] Department of Mathematical and Industrial Engineering, Polytechnique Montréal, 2500 chemin de Polytechnique, Montreal, Quebec, Canada H3T 1J4

ARTICLE INFO

Keywords:
Machinery safety
Accident
Prevention
Logical Analysis of Data (LAD)
Pattern recognition
Risk management

ABSTRACT

This paper deals with the application of Logical Analysis of Data (LAD) to machinery-related occupational accidents, using belt-conveyor-related accidents as an example. LAD is a pattern recognition and classification approach. It exploits the advancement in information technology and computational power in order to characterize the phenomenon under study. The application of LAD to machinery-related accident prevention is innovative. Ideally, accidents do not occur regularly, and as a result, companies have little data about them. The first objective of this paper is to demonstrate the feasibility of using LAD as an algorithm to characterize a small sample of machinery-related accidents with an adequate average classification accuracy. The second is to show that LAD can be used for prevention of machinery-related accidents. The results indicate that LAD is able to characterize different types of accidents with an average classification accuracy of 72–74%, which is satisfactory when compared with other studies dealing with large amounts of data where such a level of accuracy is considered adequate. The paper shows that the quantitative information provided by LAD about the patterns generated can be used as a logical way to prioritize risk factors. This prioritization helps safety practitioners make decisions regarding safety measures for machines.

1. Introduction

Data mining is the process of extracting hidden knowledge from data. The knowledge is extracted by means of a specific algorithm, such as support vector machine, neural networks, decision trees, association rules, or logical analysis of data (LAD). For LAD, the knowledge extracted is a set of rules or patterns describing classes of observations. In this paper, observations are accidents. The classes of observations are types of accidents, such as: a maintenance-related accident or a production-related accident. Every observation is a vector of indicators values that are recorded at the time when the accident takes place. The indicators are variables whose values describe the accident. For instance, "Presence of safeguarding" can be an indicator. It may have the value "yes" or "no" at the time of the accident. As another example, the indicator "Worker's

time in current position" may have the value: 0-4 years or 5-10 years, and so on, at the time of the accident.

Data mining techniques can be preferred over traditional methods involving tests of statistical hypothesis where a minimum number of observations is an important requirement. Data mining techniques have been used for risk management in a variety of fields, including finance [1], medicine [2], transportation [3], and occupational health and safety (OHS) [4-7]. The OHS studies deal with accidents or incidents related to workplace hazards or risk factors in general. All kinds of hazards (e.g., violence, emissions, machine-related hazards) are treated simultaneously in these studies. However, so far no study has focused on machine safety.

In OHS, Verma et al. [4] used association rules to extract frequent patterns from a database of 843 events that occurred in a plant between March 2010 and July 2013. The events were accidents, near-misses, and incidents involving material and environmental damage. Expressed as association rules, the patterns represented the acquired knowledge extracted from the database. For instance, some rules showed that behavior-related problems, such as unsafe acts performed by others, resulted in a number of injuries. Moreover, non-compliance with standard operating procedures was involved in property damage cases. The rules arrived at served as risk management tools: the variables that made them up represented the risk factors requiring risk reduction measures. For example, the factors pinpointed by the rules were useful in accident investigation. Discussions were then undertaken with safety experts that led to the identification of the root causes underlying these behavior-related problems: work stress, production pressure, overconfidence, lack of concentration, lack of training for new workers, and lack of supervision for new workers. As a result, it was concluded that some measures, such as training, needed to be provided, mainly to new employees, but also to temporary workers who lacked experience.

Cheng et al. [5] made a decision tree from a database of 1 542 construction accidents covering the period from 2000 to 2009. The researchers used the rules revealed by the decision tree to explain the cause-and-effect relationships. For example, one of the rules generated indicated that accidents related to the collapse of objects were more common under three concurrent conditions: 1) the source of injury was the structure and the construction facilities (e.g., scaffolding), 2) the work was performed under unsafe conditions: use of hazardous methods or procedures, and 3) the worker failed to use safeguards or ignored hazard warning signs. The rules guided preventive actions.

Silva & Jacinto [6] studied the cause-and-effect relationship regarding occupational accidents in the extraction industry. A total of 6 089 accidents from the period 2005–2007 were analyzed. Three patterns were identified. Each of them characterized a specific type of accident: 1) being struck by an object, 2) physical or mental stress, 3) horizontal or vertical impact, fall of person. In order to find the patterns, a method based on multivariate analysis was applied, measuring the variables' statistical cohesion with the Pearsons' chi-square test (χ^2). The associated variables formed the patterns. The latter were used as the basis of strategies to improve safety. For example, the variables forming the pattern concerning the second type of accident focused on human behavior. Accordingly, Silva & Jacinto [6] suggested that prevention measures should be behavior-based, such as specific training sessions and well-targeted information campaigns.

Rivas et al. [7] applied various data mining techniques to determine the capacity to predict an accident or incident, and to explain such an event. The techniques tested were association rules, decision trees, Bayesian networks, support vector machine, and logistic regression. Information about the occurrence of each accident and incident was gathered by means of a survey in two companies from the mining and construction sectors respectively. The data related to the variables describing the events came from the information declared in 62 completed questionnaires, i.e., 18 accidents and 44 incidents. Rivas et al. [7] found that the best-performing predictive models were the first four above-mentioned techniques. However, only the first three demonstrated good explanatory power, showing that the occurrence of accidents in the companies could be explained by 1) task duration in hours, and 2) company contractual status (i.e., subcontractor or main contractor).

In previous OHS studies, data mining has been used for decision support to help prevent accidents. Unfortunately, the algorithms that were used inferred knowledge without covering all the observations. For instance, in association rules, the knowledge inferred is based on the identification of frequent sets of variables values in the data that meet a certain threshold. When the threshold is not satisfied, the observations concerned are rejected even though they bring new information to the database. Moreover, except for Rivas et al. [7], these studies dealt with huge databases comprising hundreds or thousands of observations. Usually, data mining techniques require large amounts of data [6] in order to extract rules describing the trends in the data. But what about plants or industries where few accidents occur? When the amount of data is limited (i.e., small sample size), the frequent sets of variables values become rarer. Accordingly, the chances of finding strong rules characterizing the data decrease. As a result, there is a need to be able to extract hidden knowledge from scarce data with adequate classification accuracy. This paper proposes to apply LAD as a data mining algorithm that is able to infer such knowledge. Indeed, LAD allows pattern generation using scarce data, as long as there is at least one observation from one class that is different from one observation from another class. Of course, when the data are too few, the patterns cannot be generalized as it was possible in [6] with thousands of data. However, the patterns can describe or predict events only for the plants or industrial sectors concerned by the data, which is not always possible with other data mining techniques. Moreover, unlike [6] where statistical hypothesis was required, this paper proposes a study free of such hypothesis which eases the process of pattern generation. Another advantage of LAD over other data mining algorithms is the fact that all the observations are covered by patterns as long as the observations bring no contradiction into the database. Contradiction means having one class of events with an observation described with the same values of variables as an observation from another class.

Contrary to the case of Rivas et al. [7], where association rules and decision trees demonstrated high predictive performance in spite of insufficient data, these techniques performed poorly when the authors of the current paper applied them to scarce data (23×23 database) describing machinery-related accidents. For instance, only weak rules were obtained with the association rule and decision tree algorithms from Tanagra software [8]. That situation could be explained by the fact that Rivas et al. [7] might have had sufficient frequent sets of variables values in their database, which was not the case in the study reported on here. LAD has been used successfully in such diverse fields as medicine [2, 9-11], finance [1], and condition-based maintenance [12, 13]. LAD showed better prediction rates than other data mining techniques such as decision trees and neural networks [10, 12] and was in fact the most accurate data mining technique in those

cases. However, those studies dealt with huge databases, and the accuracy of LAD with small databases needed investigation.

The aim of this paper is thus two-fold. The first objective is to show that LAD is an algorithm able to characterize a narrow sample of machinery-related occupational accidents with adequate average classification accuracy. The second is to show that LAD can be used for prevention of machinery-related accidents. The use of LAD in machinery safety was suggested in a previous work [14] related to this study. The literature review from reference [14] highlighted some studies dealing with experience feedback using various data mining techniques to extract knowledge from events. One of this techniques, LAD, showed to outperform in medicine for disease diagnosis and prognosis when comes the time to distinguish and characterizing classes of events. The ability of LAD to perform on rare data was one of the main reasons why it was suggested in reference [14] for knowledge extraction suitable for machinery safety. In this paper, LAD is applied to machinery related accidents.

In the remainder of this paper, the LAD algorithm is described (section 2) and then the application of LAD to scarce data related to machinery accidents is presented (sections 3 and 4). The patterns generated from that application, as well as the average classification accuracy, are presented in section 5. The potential of LAD to prevent machinery-related accidents is discussed in section 6, based on the results.

2. Description of LAD

LAD is a data mining, pattern recognition, and classification algorithm. It is an optimization combinatorial process based on Boolean logic. LAD can deal with two-class or multi-class classification problems. In this paper, LAD is applied to a two-class problem. Generally, one of the two classes is called “positive” or “true,” whereas the other is called “negative” or “false.” A class can also have a precise name referring to a specific phenomenon or event, just as the example of section 2.1. It presents two classes: 1) Accidents resulting in harm versus 2) Accidents not resulting in harm. Moreover, as presented in section 3, the study regarding this paper deals with the two following specific classes: 1) a maintenance-related-accident class and 2) a production-related-accident class. As with any data mining technique, LAD is performed in two main phases: 1) training, then 2) testing.

The training phase consists of browsing any numerical database in order to extract knowledge naturally hidden in the database. The knowledge is generated in the form of patterns. The latter are sets of conditions that best characterize and distinguish specific classes of the phenomenon under study. Here, “condition” means an interval of values associated with the indicators of a class. Each indicator is a variable that is binary or not. The database is made up of a list of observations described by indicators. Each observation is a vector whose components are the indicators’ values. In this study, the indicators represent the potential causes or risk factors of accidents; each indicator has safety- or hazard-related possible values. The values describe the context in which each accident occurred. The LAD training phase comprises three stages: 1) Data Binarization, 2) Support Set Generation, and 3) Pattern Generation (Figure B1). These stages are explained in sections 2.1-2.3 by means of an example.

The testing phase, also named the validation phase, consists of measuring the pattern classification accuracy. That accuracy guides the selection of the appropriate patterns in order to form the theory [15, 16]. The testing phase consists of a single stage: the theory formation, which is the classification rule formation (Figure B1). That stage is explained in section 2.4.

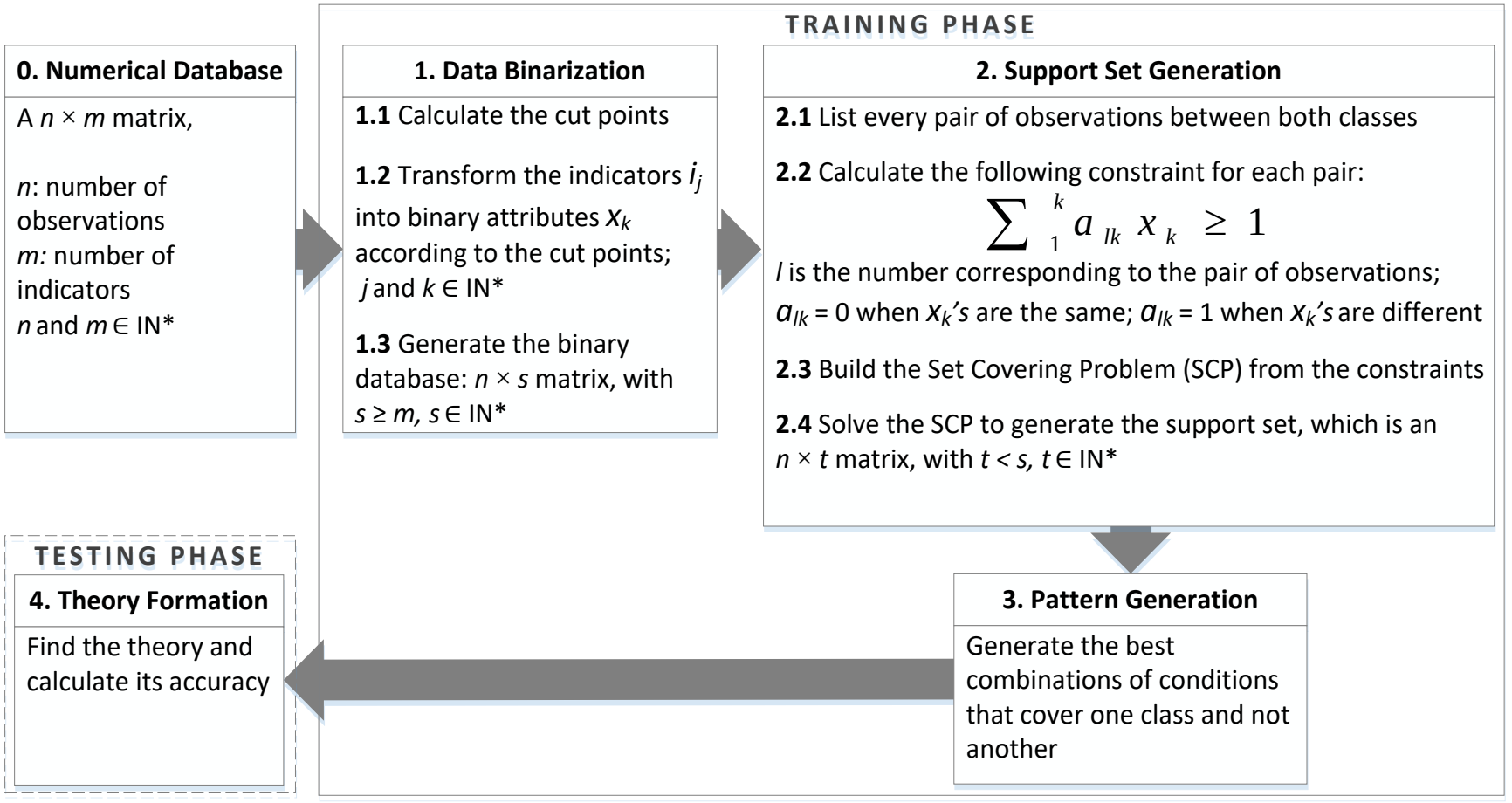


Figure B1: LAD general process

2.1 Data Binarization

The example used to explain LAD general process is based on the numerical database presented in Table B1. Two arbitrary classes are chosen for this database: C^+ and C^-. In this example, the positive class C^+ represents accidents that happened and resulted in harm for the worker. The negative class C^- represents accidents not resulting in harm for the worker. Class C^+ contains two observations: C_1^+ and C_2^+. Similarly, class C^- comprises two observations: C_1^- and C_2^-. The number of observations and indicators, as well as their values in Table B1 B1 were chosen arbitrarily.

Since LAD is based on Boolean logic and the database may not necessarily be binary, the data are first binarized before the patterns are extracted. This transformation is called "binarization": the non-binary indicators $i_1, i_2, \ldots, i_m$ are transformed into binary attributes $x_1, x_2, \ldots, x_s$ (m and s are defined at stages 0 and 1 of Figure B1). To undertake that transformation, the algorithm first aligns in increasing order the distinct values of each indicator throughout the database as seen in the first row of Table B2. Afterwards, it calculates the cut points mentioned in block 1 of Figure B1. A cut point is the average between each pair of consecutive values belonging to different classes as seen in the second row of Table B2. The cut points calculation is necessary in order to introduce binary attributes that distinguish the different classes. The number of binary attributes needed for each indicator equals the number of cut points found per indicator. Table B2 gives an example of how the cut points are estimated based on the data in Table B1. A cut point greater than zero entails that the value of the binary attribute produced is 1. A cut point lower than zero implies that the value of the binary attribute produced is 0. Table B3 shows how the transformation is performed to produce the binary database in Table B4.

Table B1: Example – A numerical database (stage 0 of Figure B1)

Class	Observation	i_1	i_2	i_3
C^+	C_1^+	1	3	1
	C_2^+	1	2	3
C^-	C_1^-	0	1	4
	C_2^-	0	1	2

Table B2: Example – Cut point calculation (stage 1.1 of Figure B1)

	i_1	i_2	i_3
Values of indicators in increasing order	0; 1	1; 2; 3	1; 2; 3; 4
Cut point (cp) = average between each pair of consecutive values belonging to different classes	$cp_1 = 0.5$	$cp_2 = 1.5$	$cp_3 = 1.5$ $cp_4 = 2.5$ $cp_5 = 3.5$

Table B3: Example – Transformation of indicators into binary attributes (stage 1.2 of Figure B1)

Observation	i_1		i_2		i_3					
	Logic for transformation	x_1	Logic for transformation	x_2	Logic for transformation	x_3	Logic for transformation	x_4	Logic for transformation	x_5
C_1^+	$(i_1 = 1) > cp_1 \rightarrow$	1	$(i_2 = 3) > cp_2 \rightarrow$	1	$(i_3 = 1) < cp_3 \rightarrow$	0	$(i_3 = 1) < cp_4 \rightarrow$	0	$(i_3 = 1) < cp_5 \rightarrow$	0
C_2^+	$(i_1 = 1) > cp_1 \rightarrow$	1	$(i_2 = 2) > cp_2 \rightarrow$	1	$(i_3 = 3) > cp_3 \rightarrow$	1	$(i_3 = 3) > cp_4 \rightarrow$	1	$(i_3 = 3) < cp_5 \rightarrow$	0
C_1^-	$(i_1 = 0) < cp_1 \rightarrow$	0	$(i_2 = 1) < cp_2 \rightarrow$	0	$(i_3 = 4) > cp_3 \rightarrow$	1	$(i_3 = 4) > cp_4 \rightarrow$	1	$(i_3 = 4) > cp_5 \rightarrow$	1
C_2^-	$(i_1 = 0) < cp_1 \rightarrow$	0	$(i_2 = 1) < cp_2 \rightarrow$	0	$(i_3 = 2) > cp_3 \rightarrow$	1	$(i_3 = 2) < cp_4 \rightarrow$	0	$(i_3 = 2) < cp_5 \rightarrow$	0

Table B4: Example – Binary attributes obtained after binarization (stage 1.3 of Figure B1)

Class	Observation	x_1	x_2	x_3	x_4	x_5
C^+	C_1^+	1	1	0	0	0
	C_2^+	1	1	1	1	0
C^-	C_1^-	0	0	1	1	1
	C_2^-	0	0	1	0	0

2.2 Support Set Generation

The purpose of this step is to minimize the dimension of the binary database in order to reduce the computational complexity of the pattern and theory formation [16]. It is an optimization problem solved here to determine the smallest sufficient subset of binary attributes required to distinguish an observation in one class from one in another class [15, 17]. All in all, the purpose of seeking for a support set is to reduce the size of the binary database by eliminating as many redundant attributes as possible, while preserving the following property in the data: no observation point is simultaneously true and false (i.e. positive and negative) [16]. Accordingly, the support set is a contradiction-free database [16].

Based on the property above, the very essence of LAD consists of distinguishing classes from one another. Consequently, LAD has to browse the database to make sure that every observation from one class is different from every observation from another class. In order to do that, LAD has to make up pairs of observations from different classes. That is the reason for stage 2.1 of Figure B1. Once the pairs of observations are listed, LAD compares the observations forming them. The equation at stage 2.2 of Figure B1 insures that no observation from one class is identical to an observation from another class.

Consequently, the list of all the possible pairs of observations to be compared between both classes of Table B4 can be found drawing the following trees in Figure B2.

Figure B2: Comparison of every observation from the positive class to all observations from the negative class

Subsequently, 4 pairs of observations are obtained: 1) $C_1^+C_1^-$, 2) $C_1^+C_2^-$, 3) $C_2^+C_1^-$, and 4) $C_2^+C_2^-$. Afterwards, the comparison between the observations of each pair is carried out. In the second pair for example, the value of x_1 in observation C_1^+ differs from the value of x_1 in observation C_2^- (see Table B4). Due to that difference, the corresponding a_{lk} value regarding that second pair of observations will be: $a_{21} = 1$ according to the explanation at stage 2.2 of Figure B1. Similarly, when comparing the values of x_2 and x_3 for the second pair of observation, one can notice that the x_k's values are different. Consequently, $a_{22} = 1$ and $a_{23} = 1$. On the contrary, similar values are noticed for x_4, as well as x_5 when comparing observation C_1^+ to C_2^-. Therefore, the corresponding a_{lk} values for the second pair of observations will be: $a_{24} = 0$, $a_{25} = 0$. The a_{lk} values for the other pairs of observations are obtained based on the same reasoning.

The summation $\sum_1^k a_{lk}x_k$ from stage 2.2 of Figure B1 can be represented by the product of the two following matrices: $[A]\times[X]$. Matrix A is made of the a_{lk} values. In the context of the example in Table B4, matrix A comprises 4 rows corresponding to the 4 pairs of observations. Then, the subscript $l = 1, 2, \ldots, 4$. Matrix A has 5 columns corresponding to the 5 binary attributes. Hence, the subscript $k = 1, 2, \ldots, 5$. Matrix X is the vector of the binary attributes x_k. Applying the constraint formula from stage 2.2 of Figure B1 to the example in Table B4, the expression $\sum_1^k a_{lk}x_k \geq 1$ is equivalent to the following system of inequalities (1):

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} \geq \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} \geq \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \quad (1)$$

The product of the matrices will lead to the following Set Covering Problem (SCP) mentioned in stage 2.3 of Figure B1:

$$\begin{aligned} &\textit{Minimize} && x_1 + x_2 + x_3 + x_4 + x_5 \\ &\textit{subject to} && x_1 + x_2 + x_3 + x_4 + x_5 \geq 1 \\ & && x_1 + x_2 + x_3 \geq 1 \\ & && (2) \\ & && x_1 + x_2 + x_5 \geq 1 \\ & && x_1 + x_2 + x_4 \geq 1 \end{aligned}$$

with $x_k \in \{0, 1\}$.

The SCP is an NP-complete problem. Its solution allows the generation of a support set (Table B5). That problem can be solved with the FOCUS-2 algorithm explained in [17].

Table B5: Example – Support set generated (stage 2.4 of Figure B1)

Class	Observation	x_1
C^+	C_1^+	1
	C_2^+	1
C^-	C_1^-	0
	C_2^-	0

Note that the Support Set Generation stage is not mandatory. In cases where all the indicators need to be considered in order to describe the phenomenon under study precisely, this stage can be omitted. However, pattern generation will then be computationally demanding. For instance, in large companies that can afford to track all indicators involved in machinery-related accidents, the Support Set Generation stage can be skipped. As a result, the patterns will be generated based on all the binary indicators contained in the initial binary database. In contrast, the Support Set Generation stage can be helpful to small-and-medium-sized enterprises (SMEs) that have fewer resources than large companies to allocate to accident prevention. By reducing the number of binary indicators and keeping the more relevant ones, SMEs will have fewer indicators to track. They will thus be able to prevent machinery-related accidents by focusing on the main relevant indicators. In the study reported on here, Support Set Generation was considered in the LAD process.

When the Support Set Generation stage is carried out, patterns are generated based on the Support Set. When that stage is skipped, patterns are generated from the binary database obtained at stage 1.3.

2.3 Pattern Generation

Patterns are the building blocks of the LAD technique. They are the extracted hidden rules that differentiate and characterize the classes. Each pattern is made up of one indicator meeting a certain threshold or an AND combination of indicators covering a proportion of observations from a class. Under LAD, every observation must be covered by at least one pattern. A pattern from one class never covers an observation from another class. Patterns can be generated based on various approaches. Ragab et al. [18] identified three:

- Enumeration-based approaches [16, 19]
- Heuristic approaches [20]
- Mixed 0-1 Integer and Linear Programming (MILP)-based methods [21, 22]

The cbmLAD software program [23] used for this paper generates patterns based on an optimized version of the MILP approach.

Below, the MILP approach is explained using the Support Set Generated in Table B5. In the interests of simplicity, the pattern generation method is shown for the positive class only, since a similar procedure applies to the negative class. The MILP approach generates positive patterns by solving the following set covering minimization problem:

Minimize $\sum_{h \in C^+}(z_h)$

subject to

$u_k + u_{t+k} \leq 1 \quad \forall\, k = 1,2,\ldots t$ (3)

$\sum_{k=1}^{2t}(v_{h,k}\, u_k) + t z_h \geq d \quad \forall\, h \in C^+$ (4)

$\sum_{k=1}^{2t}(v_{h,k}\, u_k) \leq d - 1 \quad \forall\, h \in C^-$ (5)

$\sum_{k=1}^{2t} u_k = d$ (6)

z_h is a component of the Boolean coverage vector Z. The coverage is the number of observations in which the conditions forming the pattern are encountered. The components of vector *Z* can only take the values 0 or 1. The size of that vector equals the number of positive observations in class C^+. Vector Z indicates the observations covered by the pattern. $z_h = 0$ means the observation numbered *h* is covered by the pattern. $z_h = 1$ means the observation numbered *h* is not covered by the pattern. In the example of Table B5, observation C_1^+ corresponds to observation numbered *h* = 1; C_2^+ to *h* = 2; C_1^- to *h* = 3; C_2^- to *h* = 4. The purpose of minimizing *Z* is to lessen the number of positive observations not covered by the positive pattern to be generated and maximize its degree *d* respecting constraints (3) to (6). Minimizing $\sum_{h \in C^+}(z_h)$ aims at getting the minimum possible sum of z_h's. The interest here is to get the biggest amount possible of $z_h = 0$. In other words, the algorithm seeks to get the maximum observations possible from the positive class C^+ that are covered by the pattern, subject to constraints (3) to (6).

Each pattern to be generated is associated with a Boolean pattern vector *U* (u_1, u_2, …, u_{2t}). The size of that vector equals 2*t*, where *t* is the number of binary attributes in the database. The value *2t* is the total amount of binary attributes and their corresponding negations. *k* is the number corresponding to the subscript of the binary attribute in the support set generated. The components of vector *U* are u_k's (i.e. from u_1 to u_t) and u_{t+k}'s (i.e. from u_{t+1} to u_{2t}). They only take the values 0 or 1. Component $u_k = 1$ means the binary attribute x_k is part of the pattern. Component $u_{k+t} = 1$ means the negation $\overline{x_k}$ of the binary attribute is part of the pattern. Otherwise, they are not included in the pattern. Such definition of vector *U*'s components yields to constraint (6): $\sum_{k=1}^{2t} u_k = d$. The degree *d* of the pattern indicates the number of binary attributes that compose it ($1 \leq d \leq t$). Consequently, the degree *d* of the pattern equals the number of 1's in the Boolean pattern vector *U*. The binary value of each component of the vector *U* acknowledges the presence or not of a binary attribute or its negation. Accordingly, the summation of the components values represents the total number of attributes and their negations composing the pattern.

Therefore, the constraint (6) associated with the Support Set generated in Table B5, will be: $u_1 + u_2 = d$, since $k = 1$ and $2t = 2$ in respect of that table. In the example of Table B5, $t = 1$ because there is only one binary attribute in that set: x_1. Consequently, and considering the meaning of *k*, the latter has only one possible value: $k = 1$.

Moreover, the definition of vector *U*'s components also implies that a pattern cannot include a binary attribute and its negation at the same time, which is the reason for constraint (3): $u_k + u_{t+k} \leq 1 \;\forall\, k = 1,2,\ldots t$. Indeed, based on the previous definition of u_k and u_{k+t}, if both have the

value 1 it would imply that the pattern comprises an attribute and its negation at the same time, which would make the sum $u_k + u_{k+t} > 1$. But, the algorithm requires the opposite, i.e. $u_k + u_{k+t} \leq 1$.

Considering the Support Set generated in Table B5 and the fact that $t = k = 1$ in that example, constraint (3) will be: $u_1 + u_2 \leq 1$.

Every observation of the Support Set can be associated with a Boolean observation vector: V_h ($v_{h,1}$, $v_{h,2}$,…$v_{h,t}$, $v_{h,t+1}$, $v_{h,t+2}$, … $v_{h,2t}$) such that $v_{h,k} = 1$ if $x_k = 1$, and $v_{h,t+k} = 1$ if $x_k = 0$. Otherwise, $v_{h,k} = 0$, and $v_{h,t+k} = 0$. In other words, $v_{h,k}$ (i.e. $v_{h,1}$ to $v_{h,t}$) equals the value that the k-th binary attribute takes in the observation numbered h. On the contrary, $v_{h,t+k}$ (i.e. $v_{h,t+1}$ to $v_{h,2t}$) is the negation of $v_{h,k}$.

Based on the Support Set Generated in the example of Table B5, one can obtain the associated Boolean observation vectors V_h shown in Table B6. In that example, the Boolean observation vectors will be: V_h ($v_{h,1}$, $v_{h,2}$) since $t = k = 1$.

Table B6: Example – Components of Boolean observation vectors V_h

		Support Set Generated	V_h	
Observation No.	**Class**	x_1	$v_{h,1} = x_1$	$v_{h,2} = \overline{x_1}$
$h = 1$	C^+	1	1	0
$h = 2$		1	1	0
$h = 3$	C^-	0	0	1
$h = 4$		0	0	1

In order to generate patterns that partition the data in $C^+ \cup C^-$ into two classes, the MILP approach introduces the following expression: $\sum_{k=1}^{2t}(v_{h,k}\, u_k)$ [22]. If a positive pattern covers a positive observation, the dot product of the Boolean pattern vector U and the Boolean observation vector V_h of each observation covered by that pattern must be equal to the degree d of that pattern [21]. That yields to the following equation:

$$\sum_{k=1}^{2t}(v_{h,k}\, u_k) = d \tag{7}$$

The positive pattern to be generated must cover at least one positive observation. At the same time, it is not mandatory for the positive pattern to cover all the positive observations. That condition is described by constraint (4): $\sum_{k=1}^{2t}(v_{h,k}\, u_k) + tz_h \geq d \;\forall\, h \in C^+$.

The positive pattern should not cover any negative observations. Therefore, the dot product of the Boolean pattern vector U and the Boolean observation vector V_h of each negative observation must be lower than the degree d of the positive pattern [21] (i.e. "$< d$" or "$\leq d - 1$"). That entails constraint (5): $\sum_{k=1}^{2t}(v_{h,k}\, u_k) \leq d - 1 \;\forall\, h \in C^-$.

The example in Table B6 and the understanding of the previous constraints enable the formulation, in Table B7, of the set covering minimization problem. Note that, unlike constraints

(3) and (6), constraints (4) and (5) are repeated in Table B7, because they depend on the observation number as specified in the previous set covering minimization problem. Constraint (4) must be developed for observations numbered $h = 1$ and $h = 2$ only, since that constraint concerns the positive class C^+ solely. Similarly, constraint (5) must be developed for observations numbered $h = 3$ and $h = 4$ only, since it regards the negative class C^- solely.

Here is an explanation of the development of constraints (4) and (5), taking observations numbered $h = 2$ and $h = 3$ as an example:

- Constraint (4) with $h = 2$:

  $\sum_{k=1}^{2t}(v_{h,k}\, u_k) + tz_h \geq d, \forall\, h \in C^+$

  Since $h = 2$ and $t = 1$, we have:

  $tz_h = 1 \cdot z_2 \rightarrow tz_h = z_2$

  and $\sum_{k=1}^{2t}(v_{h,k}\, u_k) = \sum_{k=1}^{2}(v_{2,k}\, u_k)$

  $\rightarrow \sum_{k=1}^{2}(v_{2,k}\, u_k) = (v_{2,1}\, u_1) + (v_{2,2}\, u_2)$

  Based on the values of $v_{h,k}$ in Table B6, we have:

  $\sum_{k=1}^{2}(v_{2,k}\, u_k) = (1 \cdot u_1) + (0 \cdot u_2)$

  $= u_1$

  Therefore, constraint (4) for observation numbered $h = 2$ becomes: $u_1 + z_2 \geq d$ as mentioned in Table B7.

- Constraint (5) with $h = 3$:

  $\sum_{k=1}^{2t}(v_{h,k}\, u_k) \leq d - 1, \forall\, h \in C^-$

  Since $h = 3$ and $t = 1$, we have:

  $\sum_{k=1}^{2t}(v_{h,k}\, u_k) = \sum_{k=1}^{2}(v_{3,k}\, u_k)$

  $\rightarrow \sum_{k=1}^{2}(v_{3,k}\, u_k) = (v_{3,1}\, u_1) + (v_{3,2}\, u_2)$

  Based on the values of $v_{h,k}$ in Table B6, we have:

  $\sum_{k=1}^{2}(v_{3,k}\, u_k) = (0 \cdot u_1) + (1 \cdot u_2)$

  $= u_2$

  Therefore, constraint (5) for observation numbered $h = 3$ becomes: $u_2 \leq d\text{-}1$ as mentioned in Table B7.

Table B7: Example – Set covering minimization problem

Observation No.	Problem		Constraint No.
	Minimize $z_1 + z_2$		
	Subject to:		
	$u_1 + u_2$	≤ 1	(3)
$h = 1$	$u_1 + z_1$	$\geq d$	(4)
$h = 2$	$u_1 + z_2$	$\geq d$	(4)
$h = 3$	u_2	$\leq d\text{-}1$	(5)
$h = 4$	u_2	$\leq d\text{-}1$	(5)
	$u_1 + u_2$	$= d$	(6)

That problem can easily be solved, knowing that $d = 1$, since $t = 1$ and $1 \leq d \leq t$, as previously mentioned. That entails $u_2 = 0$ from constraint (5), then $u_1 = 1$ from constraint (6). Therefore, from constraint (4) $z_1 = z_2 = 0$. Based on what was said earlier, $u_1 = 1$ means that the binary attribute x_1 represents the positive pattern found. That pattern is $P_1^+ = x_1$. It covers observations 1 and 2, since the coverage vector $Z(z_1, z_2) = Z(0, 0)$. If the positive pattern had not covered all the positive observations, it would have been necessary to repeat the set covering minimization problem for the remaining uncovered observations, in order to find a pattern that covers them all, just as Ragab et al. [24] did in their example.

2.4 Theory formation

At this stage, the classification accuracy of the patterns generated from a training database is estimated based on the observations from a testing database. The calculation of the following discriminant dictates the class of the tested observation denoted *y*:

$$\Delta(y) = \sum_{P_q^+ \in P^+}^{f} w_q p_q^+(y) - \sum_{P_r^- \in P^-}^{g} w_r p_r^-(y) \quad (8)$$

Let's denote:

- P^+, P^-: the set of positive or negative patterns, respectively
- P_q^+: a positive pattern belonging to P^+
- P_r^- : a negative pattern belonging to P^-
- $p_q^+(y)$: a value that equals to 1 if the generated positive pattern P_q^+ covers the testing observation *y*, and equals to zero otherwise.
- $p_r^-(y)$: a value that equals to 1 if the generated negative pattern P_r^- covers the testing observation *y*, and equals to zero otherwise.
- *f*, *g*: the total number of positive or negative patterns, respectively
- w_q, w_r: the weight associated with P_q^+ or P_r^-, respectively. The weight of a pattern is the ratio of the number of observations from a class that are covered by that pattern to the total coverage of the patterns of that class.

The value of the discriminant lies between -1 and 1. If $\Delta(y) > 0$, the tested observation *y* will be classified into the positive class. If $\Delta(y) < 0$, *y* will be classified into the negative class. If $\Delta(y) = 0$, *y* will remain unclassified. No calculation example is given in this subsection, since section 4 of this paper details how that stage is performed.

3. Application of LAD – Training Phase

The data mining process (step 4 of Figure B3) is part of a larger process called "knowledge extraction." In this study, the algorithm used at step 4 is LAD, and the software is cbmLAD [23]. As the arrows indicate, knowledge extraction requires going back and forth through some preliminary steps (steps 1-3 of Figure B3) in order to prepare the data so they are compatible with the data mining algorithm, as well as with the software used to infer the knowledge. These preliminary steps also "clean" and organize the data to ensure the relevant information is retained. As long as the interpretation of the patterns inferred (step 5 of Figure B3) is not satisfactory, some adjustments are required in the previous steps. The knowledge extraction process illustrated in Figure B3 is applied in this study for the purpose of extracting knowledge from machinery-related accident reports.

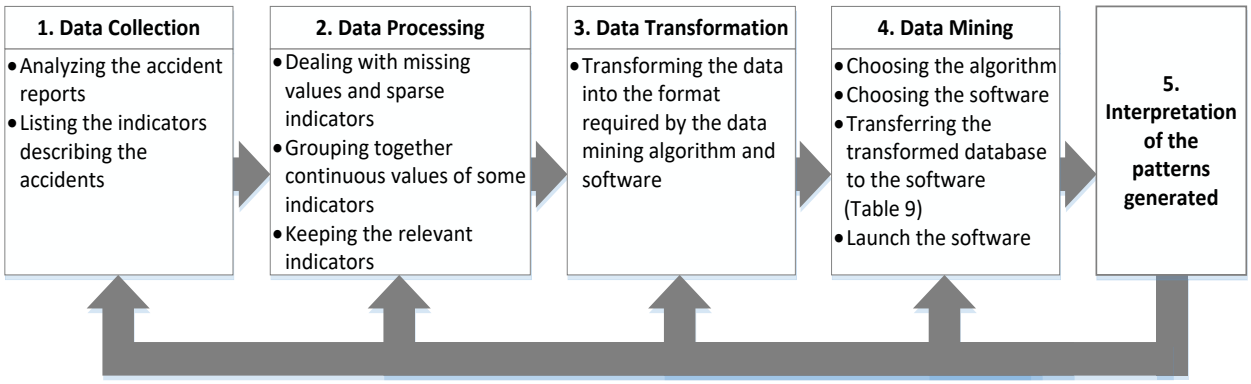


Figure B3: Knowledge extraction process

3.1 Data Collection: Choosing the Observations

To illustrate the application of LAD to machinery safety, one type of machine was selected: belt conveyors. This type of machine was chosen because it represented the highest proportion (8.5%) of serious and fatal injuries, according to an original database developed at the IRSST (Institut de recherche Robert-Sauvé en santé et en sécurité du travail), an OHS research institute. This database contains reports on 164 accidents that occurred between 1999 and 2007 involving non-mobile machinery (reports stored at the documentation center of the CNESST: Commission des normes, de l'équité, de la santé et de la sécurité du travail, www.csst.qc.ca). According to a recent search at the documentation center for the purposes of this study, belt conveyors were found to be the type of conveyor with the highest score: 23 out of 137 conveyor-related serious and fatal injuries from 1990 to 2011. Another study [25], which analyzed serious and fatal accidents linked to moving parts of machinery, revealed that many accidents were caused by the in-running nip points formed by the moving belt and rotating rollers on belt conveyors. Consequently, 23 belt-conveyor-related accident reports—20 regarding maintenance activities and 3 concerning production activities—were retained for the purpose of observation analysis by means of LAD. Table B18 from Appendix describes what kinds of activities were involved in each class of accidents analyzed. Accident reports available at the documentation center involved only serious and fatal injuries (i.e., no near misses or minor injuries).

LAD was therefore used to characterize a small sample of belt-conveyor-related occupational accidents in order to identify potential causes or risk factors for maintenance-related accidents and production-related accidents. This paper therefore seeks to characterize and distinguish two classes of accidents, 1) maintenance-related and 2) production-related, to illustrate two different contexts of use of the machine. By distinguishing the two contexts, safety practitioners will be able to take specific or targeted prevention measures, while considering the realities of their companies. The word “targeted” refers here to the detection (identification) of the risk factors that may be the potential accident causes or explanations that characterize and distinguish one type of accident from another, and the choice of appropriate risk reduction measures suggested by these risk factors.

3.2 Data Collection: Choosing the Indicators

The accident reports varied in length, content, and degree of detail. More indicators were taken into account for this study than for Chinniah's [25] accident analysis, which included worker's experience, activities (tasks) being carried out when accidents occurred, OHS committee and OHS management, machinery risk assessment, machinery safeguards, bypassing (defeating) safeguards, lockout programs and procedures, and safety control systems. A 32-indicator database was built using Microsoft Excel.

3.3 Data Processing: Surveying and pre-processing the data

Surveying helps reveal flaws in the data, such as missing data and outliers [26]. The former are gaps in the database, whereas the latter refer to inaccurate data that deviate from the normal situation [27]. No outliers were found in this study. However, there were missing data, as not all the reports contained the same indicators. This challenge was addressed as follows:

- 8 out of the 32 indicators were deleted to minimize bias because each of them had a number of instances of missing data, representing 40% or over. Here, bias means a distortion of LAD's classification accuracy. Before the 8 indicators were deleted, missing data represented a total proportion of 20.4% of the database. Bennane & Yacout [28] observed that LAD's classification accuracy drops rapidly when the proportion of missing data is greater than 20%. Indicators with high percentages of missing data were therefore deleted.
- 6 remaining indicators had missing data representing 4–26% for each. Since the new amount of missing data, for the database as a whole, totaled 4.2%, the authors decided to fill in the gaps in the data. That decision was based on Bennane & Yacout [28], who showed that LAD's classification accuracy remains the same in general for databases with missing data totaling 5% compared with the same database without missing data. That finding was arrived at when Bennane & Yacout [28] studied the effect on LAD's classification accuracy of five pre-treatment methods for dealing with missing data. In the current study, the gaps were filled in, preserving the variability relationships of the available values for each indicator. To do that, Pyle's method [26] was used, from section 8.2.2 of his book. The value allocated to each gap in the data was the one that would least disturb the initial standard deviation (S_0) of the indicator, i.e., the value that could make the final standard deviation (S_f) of the indicator be the nearest to S_0 (see the example of indicator I_1 in Table B8). When the final standard deviations (S_f) based on the possible values were equally distant from the initial standard deviation (S_0), the possible value that provided the nearest mean (M_f) to the initial mean (M_0) of the indicator was chosen (see the example of indicator I_8 in Table B8).

Table B8: Example of how missing data were filled in

Indicator	Initial mean and standard deviation of indicator with missing data	Possible values of indicator	Final mean and standard deviation with possible value	Choice of value to fill in gap
I_1	$M_0 = 0.60$; $S_0 = \mathbf{0.50}$	0 1	$M_f = 0.61$; $S_f = \mathbf{0.50}$ $M_f = 0.65$; $S_f = 0.49$	$\rightarrow I_1 = \mathbf{0}$
I_8	$M_0 = \mathbf{0.32}$; $S_0 = 0.48$	0 1	$M_f = \mathbf{0.30}$; $S_f = 0.47$ $M_f = 0.35$; $S_f = 0.49$	$\rightarrow I_8 = \mathbf{0}$

Finally, another indicator was deleted because it had the same value throughout all the observations. Therefore, it would not be useful to the learning process, as it would not help distinguish the classes. At the end of the data processing, the database consisted of the remaining 23 indicators.

3.4 Data Transformation and Data Mining

In order to use cbmLAD, the indicator values for the raw data collected were transformed into a numerical database as shown in Table B9, based on the indicator definitions given in Appendix. Once the database was built, cbmLAD was launched to generate the patterns for each class. They are presented in section 5.

Table B9: Truncated version of numerical database in cbmLAD software

Accident No.	Class	Indicators																					
		I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}	I_{11}	I_{12}	I_{13}	I_{14}	I_{15}	I_{16}	I_{17}	I_{18}	I_{19}	I_{20}	I_{21}	I_{22}	I_{23}
1	1	1	3	1	1	1	1	2	0	0	1	0	0	0	0	0	1	0	1	3	1	0	1	2
2	1	1	2	1	1	1	1	1	1	0	1	0	0	1	0	0	2	0	1	3	0	0	1	1
⋮																								
23	0	0	1	1	1	1	0	1	0	0	1	0	0	0	0	0	4	1	0	3	1	0	2	2

N.B. Having the number of indicators equaling the number of observations (accidents) is pure coincidence.

4. Application of LAD – Testing Phase

Once the interpretation of the patterns was satisfactory, the classification accuracy of LAD for the machinery-related accidents database was estimated. Instead of estimating accuracy based on a single test, Witten et al. [29] recommend repeating the testing process several times with different samples in order to obtain an average classification accuracy. As this paper deals with a very limited-data situation, the "Leave-One-Out Cross-Validation" procedure described in [29] was first used. That choice is based on the fact that in this procedure, the greatest possible amount of data is used for training in each case, which presumably increases the chance that the classifier is an accurate one [29]. According to Witten et al. [29], leave-one-out seems to offer a chance of squeezing the maximum out of a small dataset and getting as accurate an estimate as possible.
The "Leave-One-Out Cross-Validation" procedure consists of an *n*-fold cross-validation, where *n* is the number of observations in the database [29]:

- *n* training-testing iterations are done. For each iteration, a different observation is left out. That observation is the one tested, whereas the *n*-1 remaining observations serve as a training database. In our case, *n* = 23 observations. Therefore, each iteration has a 22-observation training database, and the chosen 23rd observation for testing.
- Patterns are generated for each of the *n* training databases.
- For the testing observation corresponding to each training database, the classification accuracy is calculated based on the patterns generated for each training database and the performance of the discriminant formula $\Delta(y)$ mentioned in section 2.4. Examples of classification of testing observations (Table B11) are given in Table B12, based on the patterns generated in Table B10 at iterations 1, 21 and 23.
- Finally, the average classification accuracy is estimated based on the mean of the *n* classification accuracies calculated.

Table B10: Leave-one-out – Patterns and weights related to training databases of iterations 1, 21 and 23

		Iteration No. 1		Iteration No. 21		Iteration No. 23	
	Pattern No.	**Pattern**	**Weight**	**Pattern**	**Weight**	**Pattern**	**Weight**
Maintenance accidents	P_1^+	$I_1 > 0.5$	$w_{q1} = 0.31$	$I_1 > 0.5$	$w_{q1} = 0.46$	$I_1 > 0.5$	$w_{q1} = 0.44$
	P_2^+	$I_6 > 0.5$ $I_{16} > 1.5$ $I_{23} < 6.5$	$w_{q2} = 0.40$	$I_6 > 0.5$ $I_{16} > 1.5$ $I_{23} < 6.5$	$w_{q2} = 0.54$	$I_{16} > 1.5$ $I_{23} < 6.5$	$w_{q2} = 0.56$
	P_3^+	$I_{22} < 1.5$	$w_{q3} = 0.29$	------	------	------	------
Production accidents	P_1^-	$I_1 < 0.5$ $I_8 < 0.5$ $I_{22} > 1.5$ $I_{23} > 1.5$	$w_{r1} = 1$	$I_1 < 0.5$ $I_8 < 0.5$ $I_{16} < 4.5$ $I_{22} > 1.5$	$w_{r1} = 1$	$I_1 < 0.5$ $I_{16} < 3.5$ $I_{22} > 1.5$	$w_{r1} = 1$

Table B11: Leave-one-out – Testing observations for iterations 1, 21 and 23

Accident No.	Class	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}	I_{11}	I_{12}	I_{13}	I_{14}	I_{15}	I_{16}	I_{17}	I_{18}	I_{19}	I_{20}	I_{21}	I_{22}	I_{23}
1	1	1	3	1	1	1	1	2	0	0	1	0	0	0	0	0	1	0	1	3	1	0	1	2
21	1	0	1	1	0	0	0	1	0	0	1	0	0	1	0	0	4	1	0	2	1	1	1	4
23	0	0	1	1	1	1	0	1	0	0	1	0	0	0	0	0	4	1	0	3	1	0	2	2

Table B12: Leave-one-out – Classification of previous testing observations based on discriminant value

Accident No.	Original class	$p_1^+(y)$	$p_2^+(y)$	$p_3^+(y)$	$p_1^-(y)$	$\Delta(y) = \sum_{P_q^+ \in P^+}^{f} w_q p_q^+(y) - \sum_{P_r^- \in P^-}^{g} w_r p_r^-(y)$	Classification by $\Delta(y)$
$y = 1$	1	1	0	1	0	$\Delta(1) = 0.60 \rightarrow$	1 $\rightarrow$ S
$y = 21$	1	0	0	------	0	$\Delta(21) = 0 \rightarrow$	Unclassified $\rightarrow$ F
$y = 23$	0	0	1	------	0	$\Delta(23) = 0.56 \rightarrow$	1 $\rightarrow$ F

S: success – the testing observation is correctly classified; F: failure – the testing observation is misclassified or unclassified.

Table B12 verifies whether each pattern from the positive (P_1^+, P_2^+, P_3^+) or the negative (P_1^-) class covers the accident y tested. If yes, $p_q^+(y)$ or $p_r^-(y) = 1$, otherwise $p_q^+(y)$ or $p_r^-(y) = 0$. Based on section 2.4, the accidents 1 and 23 are classified as class 1: the maintenance-related accident class, because their associated discriminant $\Delta(y)$ is greater than 0. The accident 21 is left unclassified, since $\Delta(y) = 0$. If the "Original class" column is compared with the "Classification by $\Delta(y)$" column, it can be seen in this example that only accident 1 is correctly classified by the discriminant, while accident 21 is unclassified and accident 23 is misclassified.

The testing phase based on the "Leave-One-Out Cross-Validation" procedure considers a single testing observation from one class per iteration. Thus, both classes are not represented at once in the testing database. Consequently, there is no possibility for the testing database to contain accidents from both classes in a proportion slightly similar to that of the initial database, i.e.

respecting the ratio: 3 production-related accidents for 20 maintenance-related accidents as Witten et al. [29] recommend. Therefore, the confidence of the average classification accuracy cannot be strong. To enhance that confidence, it would be interesting to have a bigger testing database allowing representation of both classes of accidents in a proportion that represents approximately that of the initial database. Thus, the "5-fold Cross-Validation" was used as another testing procedure, splitting the 23 accidents into two sets: an 11-accident training database, and a 12-accident testing database. Even though 10-fold cross-validation has become the standard procedure in practical terms, 5-fold or 20-fold cross-validation is likely to be almost as good as mentioned in [29].

Therefore, the "5-fold Cross-Validation" was applied to the original numerical database of this study:

- For each of the 5 training-testing iterations, a different training database and a different testing database were built;
- In both databases, the proportion of production-related accident versus maintenance-related accident was slightly maintained through the iterations;
- Patterns were generated with cbmLAD for each training database;
- For the testing database corresponding to each training database, the classification accuracy of each pattern generated was calculated based on the performance of the discriminant formula ($\Delta(y)$). An example of the classification accuracy calculation is given in Tables B13-B15 based on the patterns generated at iteration 1;
- Finally, the average classification accuracy is estimated based on the mean of the 5 classification accuracies calculated.

Table B13: 5-fold – Patterns generated from the training database of iteration 1 and their weights

	Pattern No.	Pattern	Weight
Maintenance accidents	P_1^+	$I_{22} < 1.5$ $I_{23} < 6.0$	$w_{q1} = 0.33$
	P_2^+	$I_{18} > 0.5$	$w_{q2} = 0.33$
	P_3^+	$I_{20} < 0.5$ $I_{23} < 6.0$	$w_{q3} = 0.33$
Production accidents	P_1^-	$I_7 < 1.5$ $I_{18} < 0.5$ $I_{22} > 1.5$	$w_{r1} = 1$

Table B14: 5-fold – Truncated version of testing database for iteration 1

Accident No.	Class	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}	I_{11}	I_{12}	I_{13}	I_{14}	I_{15}	I_{16}	I_{17}	I_{18}	I_{19}	I_{20}	I_{21}	I_{22}	I_{23}
$y = 7$	0	0	1	1	0	1	1	2	0	0	1	0	0	0	0	0	1	1	1	3	0	0	2	3
.																								
.																								
.																								
$y = 20$	1	1	1	1	0	1	1	1	1	0	1	1	0	1	0	0	4	1	1	3	1	0	6	1
$y = 22$	1	1	2	1	1	1	0	1	0	0	1	0	0	0	0	0	1	1	1	3	1	0	1	6

Table B15: 5-fold – Classification of the tested observations for iteration 1 based on the discriminant value

Accident No.	Original class	$p_1^+(y)$	$p_2^+(y)$	$p_3^+(y)$	$p_1^-(y)$	$\Delta(y) = \sum_{P_q^+ \in P^+}^{f} w_q p_q^+(y) - \sum_{P_r^- \in P^-}^{g} w_r p_r^-(y)$	Classification by $\Delta(y)$
$y = 7$	0	0	1	1	0	$\Delta(7) = 0.67$	1 → F
⋮							⋮
$y = 20$	1	0	1	0	0	$\Delta(20) = 0.33$	1 → S
$y = 22$	1	0	1	0	0	$\Delta(22) = 0.33$	1 → S

S: success – the testing observation is correctly classified; F: failure – the testing observation is misclassified.

For iteration 1, nine tested accidents out of the twelve were correctly classified. In other words, there were 9 successes (S) and 3 failures (F).

5. Results: Patterns Generated and Accuracy of Classification

This section presents the patterns generated at the training phase of the LAD application (Table B16) for the 23-machinery-related accident database. It also shows the importance of some indicators based on the frequency of their corresponding condition (Table B17). These conditions are the ones included in the patterns generated throughout the 23 training-testing iterations. They are arranged in the table by their total frequency of appearance in the patterns. Here, the 23 training-testing iterations were chosen over the 5 training-testing iterations. The higher number of observations in the training database of the first testing procedure increases the ability of the patterns to generalize compared to the second one. Lastly, this section presents the average classification accuracies calculated at the testing phase according to the testing procedures used.

Table B16 also gives the pattern interpretations, as well as the pattern coverage and relative prevalence. The coverage is the number of observations in which the conditions forming the pattern are encountered. The relative prevalence is that number in percentage terms. For example, in Table B16, pattern P_3^+ covers a specific set of 11 maintenance-related accidents out of the 20 analyzed. Therefore, in 55% of the maintenance-related accidents analyzed, the victim was crushed. The thresholds forming the patterns are automatically calculated by cbmLAD, based on the possible values listed in Table B19 for each indicator.

Table B16: Patterns generated from entire database, with coverage and relative prevalence

	Pattern No.	Pattern	Pattern interpreted	Coverage	Relative prevalence
Maintenance accidents	P_1^+	$I_1 > 0.5$	A prevention program exists	12/20	60%
	P_2^+	$I_6 > 0.5$	Accident due to poor working environment (e.g., cluttered work area)	14/20	70%
		$I_{16} > 1.5$	Accessible hazard zone where accident occurred: Pulley *OR* roller *OR* drive drum *OR* tensioning drum *OR* tail drum		
		$I_{23} < 6.5$	Part of body in hazard zone and location of injury: Body *OR* torso *OR* head *OR* left arm *OR* right arm *OR* both arms		
	P_3^+	$I_{22} < 1.5$	Accidental circumstance: Crushing	11/20	55%
Prod. acc.	P_1^-	$I_1 < 0.5$	No prevention program	3/3	100%
		$I_8 < 0.5$	No conveyor functional impairment		
		$I_{16} < 4.5$	Accessible hazard zone where accident occurred: Pulley *OR* roller *OR* drive drum		
		$I_{22} > 1.5$	Accidental circumstance: Entrapment *OR* strangulation		

The context in which the maintenance-related accidents happened is described by the disjunction of three patterns: P_1^+, P_2^+, P_3^+ (Equation 4). Pattern P_1^- (Equation 5) describes the context of the production-related accidents.

$$\textit{Maintenance accident context} = P_1^+ \textit{ OR } P_2^+ \textit{ OR } P_3^+ \quad (9)$$

$$\textit{Production accident context} = P_1^- \quad (10)$$

where:

$P_1^+ = (I_1 > 0.5)$

(11)

$P_2^+ = (I_6 > 0.5)$ AND $(I_{16} > 1.5)$ AND $(I_{23} < 6.5)$

(12)

$P_3^+ = (I_{22} < 1.5)$

(13)

$P_1^- = (I_1 < 0.5)$ AND $(I_8 < 0.5)$ AND $(I_{16} < 4.5)$ AND $(I_{22} > 1.5)$

(14)

Table B17: Conditions ordered by total frequency

	Maintenance-related accidents								Production-related accidents								
Condition	$I_{23} < 6.5$	$I_1 > 0.5$	$I_6 > 0.5$	$I_{16} > 1.5$	$I_{22} < 1.5$	$I_5 < 0.5$	$I_{16} > 2$	$I_{18} > 0.5$	$I_{22} > 1.5$	$I_1 < 0.5$	$I_8 < 0.5$	$I_{23} > 1.5$	$I_{16} < 4.5$	$I_{11} < 0.5$	$I_{13} < 0.5$	$I_{16} < 3.5$	$I_{18} < 0.5$
Total frequency	24	22	22	21	19	2	1	1	23	22	19	14	6	1	1	1	1

Average classification accuracy obtained with the "Leave-One-Out Cross-Validation" procedure:

Throughout the iterations, 17 tested observations out of 23 were correctly classified. In contrast, 1 tested observation was left unclassified and 5 others were misclassified. As a result, there were

17 successful classifications and 6 failures. Accordingly, the average classification accuracy of LAD for the narrow machinery-related accident database based on the "Leave-One-Out Cross-Validation" procedure is 74% (17/23). For the tested maintenance-related accidents, that testing procedure misclassified 2/20 accidents and unclassified 1/20. These ratios represent 13% of observations that failed the classification test. Concerning the tested production-related accidents, that procedure misclassified all of them (i.e. 3/3 or 100%).

Average classification accuracy obtained with the "5-fold Cross-Validation" procedure:
The classification accuracies obtained at each of the 5 iterations were:

- 9/12 for iterations 1, 2 and 4;
- 8/12 for iterations 3 and 5.

Taking the mean of these values, the average classification accuracy of LAD for the accident database using the "5-fold Cross-Validation" procedure is 72%. Concerning the tested maintenance-related accidents, that testing procedure misclassified 21% of them. In the case of the tested production-related accidents, that procedure misclassified 90% of them.

6. Discussion

It cannot be claimed that the patterns generated characterize all belt-conveyor-related accidents that occur during maintenance or production activities in general or in different companies because of the sparse data from which the patterns were derived. These patterns do, however, explain the context of occurrence of the accidents analyzed for the industries concerned by the sample studied. In contrast, the context of occurrence could not be explained by either association rules or the decision tree method. This study shows the power of LAD when dealing with sparse data: the algorithm is able to learn from past events without leaving any observation behind. Moreover, LAD can generate patterns characterizing a narrow sample of 23 machinery-related accidents with an adequate average classification accuracy:

- 74% based on the "Leave-One-Out Cross-Validation" procedure;
- 72% using a "5-fold Cross-Validation" procedure involving an approximated 50-50 split of the numerical database.

Such a level of accuracy must be considered adequate, given that some authors (see the Diabetes database mentioned in [16]) refer to "correct prediction rates" ranging from 71.4–74.4% when describing their classification accuracy. This is a significant point in light of the fact that each of the training databases from the testing phase had only 22 accidents to learn from in one case, and 11 accidents to learn from in the other case, whereas the Diabetes database had 768 observations.

Comparing the 23-accident database of this study to the Diabetes database previously mentioned, one can deduce that a larger number of observations available for training does not necessarily guarantee a higher accuracy in the testing phase. However, it allows a higher confidence on the ability of the patterns generated to characterize the sample of observations analyzed.

The two testing procedures used in this paper were applied to training databases having identical indicators. Considering that a higher classification accuracy was obtained with the "Leave-One-Out Cross-Validation" procedure, one can deduce that a bigger training database can increase the average classification accuracy when identical indicators are used from one testing procedure to another. However, let's take a closer look to the classification of the tested observations by the

two procedures applied. One can notice that the patterns from the "Leave-One-Out Cross-Validation" procedure misclassified all the production-related accidents tested, whereas 13% of the maintenance-related accidents tested were misclassified. On the contrary, the "5-fold Cross-Validation" was able to classify properly at least one production-related accident tested, and misclassified 21% of the maintenance-related accidents tested. The fact that the testing databases of the "5-fold Cross-Validation" procedure were well balanced (i.e. both classes were represented in proportions similar to that of the initial database) compared to those of the "Leave-One-Out Cross-Validation" explains such a difference in the prediction of each class of accidents. The classification accuracy is far better for the maintenance-related accidents because LAD has a bigger proportion of this type of accident to learn from. To increase the chances of improving the classification accuracy, more production-related accidents regarding belt-conveyors are necessary. Unfortunately, they were unavailable.

All in all, the average classification accuracies found are adequate overall, despite scarcity in the data. Nevertheless, when comparing the predictions class to class, LAD performs far better for maintenance-related accidents than production-related accidents. However, one must keep in mind that, despite sparse data, LAD was able to generate patterns characterizing the accidental portrait of the enterprises concerned by the 23 accidents analyzed unlike association rules and decision trees.

The patterns obtained (Table B16) serve as risk identification tools. Their coverage and relative prevalence can, in combination with the importance of their indicators (Table B17), be used as risk estimation tools when prioritizing accident-prevention initiatives. The relative prevalence gives an indication of the importance of the pattern.

6.1 Pattern Interpretation for Risk Identification

Risk identification consists in detecting the risk factors (indicators) that contribute to the occurrence of an accident and its severity. Since these factors can be numerous, it is not realistic to manage them all at once. One first step would therefore be to identify the main indicators. That is what the patterns generated do. They indicate what seems to be wrong or what needs improvement in order to reduce or eliminate the risk of accident. If the safety practitioner addresses these main risk factors, the main causes and explanations of accidents can be controlled or eliminated. The patterns reveal that accidents, whether maintenance-related or production-related, can still occur even though some conditions required for safety seem to have been met. Sections 6.1.1 and 6.1.2 discuss this issue.

6.1.1 Apparent Safe Value of Some Indicators versus Accident Occurrence

***Maintenance accidents**:* Pattern P_1^+ shows that a prevention program did exist for 60% of maintenance accidents. This fact is of great concern, as it would seem to indicate that having a prevention program is a characteristic of the context in which a majority of maintenance-related accidents occurred. But how could an accident occur if a prevention program was already in place? In the accident reports analyzed, three possible reasons were noted:

- The prevention program is adequate, but the presence of other indicators with unsafe values caused the accident. For example, the procedures described in the prevention program were not followed, or no one in the workplace promotes the values and implementation of the guidelines stated in the prevention program. It is not uncommon for

companies to adopt prevention programs to achieve regulatory compliance, but fail to implement the programs. Employers need to motivate their workers to become familiar with the prevention program and follow safety guidelines.

- The prevention program has not been updated. Updating is an important factor because it takes into account changes in the workplace, such as new risks, new machinery, new employees, changes in installations, changes in regulations, and results of audits.
- The prevention program exists but is not comprehensive enough. It has numerous flaws, such as failure to implement a lockout program or new workers' training, or unclear responsibilities.

The prevention program aims at eliminating hazards in order to protect workers' health and safety. Among other things, the program must identify the main hazards in the workplace, the personal protective equipment required, training requirements and various means to achieve them, and the preventive maintenance measures to be applied. Based on the observations made in the three above-mentioned points, employers must, as part of an accident prevention action plan, promote, update and revise their prevention programs, providing clear guidelines, in order to keep the workplace safe during maintenance.

***Production accidents**:* Pattern P_1^- indicates that the production accidents analyzed occurred in the context of proper functioning of the belt conveyor. When a machine is in operation, there are moving parts or other hazards from which the worker must stay away in order to avoid injury. Safety measures are required for production activities, such as 1) making sure that hazard zones are inaccessible while machines are running, and 2) following established safety procedures when working on machinery.

6.1.2 Hazard Value of Some Indicators versus Accident Occurrence

***Maintenance accidents**:* Pattern P_2^+ reveals that 70% of the belt-conveyor-maintenance accidents analyzed occurred due to 1) a poor working environment, such as a cluttered work area, combined with 2) an accessible hazard zone, combined with 3) body parts in contact or in close proximity with hazard zones. As a preventive action plan, it would therefore be desirable for companies to eliminate or safeguard critical hazard zones for maintenance activities. If access to a hazard zone is required, a lockout or equivalent procedure must be followed. Lockout procedures involve (i) stopping the machine, (ii) isolating the hazardous energy sources, (iii) applying locks to circuit breakers or valves, (iv) dissipating residual energies, and (v) verifying that the lockout procedure has been done correctly by performing a start-up test using the controls or measuring instruments. Under OHS regulations, a lockout procedure is required for maintenance activities in several countries, including the U.S.A., Canada, France, the U.K., and Australia. However, other methods can be used based on risk assessment if lockout procedures are impossible to apply or if the maintenance can only be done when there is power to the machinery (e.g., troubleshooting). One such method is the use of reduced speed or force by means of hold-to-run controls when performing maintenance. Finally, pattern P_3^+ indicates that in 55% of the maintenance accidents observed, the injuries resulted from a part of the body getting crushed. The safety practitioner, in conjunction with the workers, must therefore identify all the areas around the belt conveyors where it is possible to get crushed while performing maintenance. Subsequently, the crush-related hazard zones or situations they identify need to be eliminated or controlled.

Production accidents: Pattern P_1^- reveals that the production accidents analyzed were the consequences of 1) the absence of a prevention program combined with 2) an accessible hazard zone and 3) a body part being trapped or strangled while the belt conveyor was in normal operation. This shows that even when there is no malfunctioning, the combination of these three factors exposes workers to major occupational risks, as the accidents are serious or even fatal. As a result, measures must be taken to reduce the risk, such as developing and implementing a prevention program, using proper safeguarding, identifying all hazard zones or situations where someone could get trapped or strangled, and addressing such zones or situations. It is noteworthy that entrapment zones that need to be eliminated are not necessarily located on the machine itself, but may be between the machine and a separate structure or piece of equipment, as was the case in two of the production-related accident reports analyzed. Finally, patterns P_1^- and P_2^+ related to production and maintenance accidents support the statement made in the belt-conveyor safety guide [30]: the majority of belt-conveyor-related accidents occur on the drums and rollers.

6.2 Relative Prevalence of Patterns and Importance of Indicators for Risk Estimation

For each class of accidents, risk estimation first involves ranking the patterns interpreted according to their relative prevalence. Second, it involves ranking their indicators according to the importance of their corresponding condition (Table B17). The importance of a condition is given by its total frequency in the patterns obtained throughout the training-testing iterations. The higher the frequency, the more important the condition, and likewise the corresponding indicator. The pattern with the highest relative prevalence should be prioritized for risk reduction, and so should its indicators, beginning with the most important condition. That ranking provides risk-management decision-making support for safety practitioners at the operations and maintenance stage of machine use. For instance, in maintenance activities (Table B16) at the plants concerned by the accidents analyzed, risk reduction must be carried out by taking action on the indicators of pattern P_2^+ first, then P_1^+, then P_3^+. For pattern P_2^+, for example, such action should focus first on the most important indicator, and then work down to the least important: I_{23}, I_6, then I_{16} according to the ranking in Table B17 for this class. These risk factors and potential causes must be addressed by taking a range of safety measures. For instance, the workers will have to be trained properly, and supervision will be required to make sure they work safely in the hazard zone; the working environment will have to be free of obstacles by making sure it is cleaned regularly; hazard zones (e.g., rollers and drums) will have to be made inaccessible using nip-point guards, or access to them will have to be secured with a safety lockout procedure, or other means mentioned in 6.1.2.

The fact that the conditions from the patterns generated for the entire database were also noted in almost all the training-testing iterations shows that the indicators of these conditions are very important.

7. Conclusion

This paper presented the LAD algorithm and its application to a small sample of belt-conveyor-related accidents. The application of LAD to machinery safety in the workplace is innovative. This article has shown that LAD is capable of characterizing a small sample of machinery-related occupational accidents with an adequate average classification accuracy. Indeed, the 72% and 74% average classification accuracies obtained respectively with the “5-fold Cross-Validation” and the “Leave-One-Out Cross-Validation” procedures prove it. Nevertheless, when comparing

the predictions class to class, LAD performs far better for the maintenance-related accidents than the production-related accidents analyzed concerning belt conveyors. This can be explained by the fact that there were 3 production-related accidents for 20 maintenance-related accidents. It is expected that for the same indicators, a training database with more production-related accidents would improve prediction quality.

Contrary to other data mining techniques mentioned in the introduction, the patterns generated by LAD, as shown in this study, cover 100% of the data. This is another fact that demonstrates an advantage of LAD, in addition to its adequate average classification accuracy. The interpretation of the patterns, as well as their relative prevalence and the importance of their conditions, showed how useful LAD can be to machinery-related accident prevention. Indeed, this paper showed that LAD can be used by decision makers to prioritize risk factors and potential accident causes, along with corresponding safety measures. Having the prioritization based on such quantitative information, instead of personal judgment, brings greater objectivity to the hierarchization of risk factors and safety measures.

The patterns interpreted showed that belt-conveyor-related accidents can happen under hazardous conditions as well as apparently-safe conditions. As discussed earlier, one should be aware of the apparent safe value of some indicators forming the patterns. The safe values may seem strange in a pattern, since patterns are made here to characterize past accidents and predict future ones. However, these apparent safe values hide something that went wrong in the workplace. Thus, it is the responsibility of the safety practitioner to investigate the unsafe sub-factors underlying to the apparent-safe-valued indicators in order to perform continued improvement. For example, the pattern showing that 60% of the maintenance-related accidents analyzed took place in spite of the company having a prevention program. Three possible groups of sub-factors that could contribute to these accidents were:

1) When the prevention program was adequate:
    - no application of the procedures described in it, or
    - no promotion of the values and implementation of the guidelines stated in the prevention program.
2) No update of the prevention program.
3) Presence of some flaws in the prevention program (e.g. failure to implement a lockout program or new workers' training, or unclear responsibilities).

Having to perform that investigation manually instead of automatically with LAD makes it a limitation of this study.

The application of LAD to accident prevention in this paper was aimed at the two following classes of accidents: maintenance-related versus production-related. However, LAD can also be applied to other classes of events, such as accidents versus near-misses, or accidents on machine A versus accidents on machine B.

The paper has also shown the importance of tracking data on relevant accident-related indicators. The periodical analysis of such databases can update knowledge acquired about risk represented in the form of patterns. The patterns generated in this study represent high-risk situations because they are based on serious and fatal injuries. Analysis of non-serious accidents and near-misses, or of accidents causing material or property damage, could also be done.

Acknowledgment

The funding provided for this study by the IRSST and the NSERC (research grant #141111) is gratefully acknowledged. The authors also wish to thank the anonymous reviewers for their comments which have improved the paper substantially.

References

[1] Hammer PL, Kogan A, Lejeune MA. A logical analysis of banks' financial strength ratings. Expert Syst Appl 2012; 39:7808–21.

[2] Alexe S, Blackstone E, Hammer PL, Ishwaran H, Lauer MS, Pothier Snader CE. Coronary risk prediction by logical analysis of data. Ann Oper Res 2003; 119:15–42.

[3] Feng S, Li Z, Ci Y, Shang G. Risk factors affecting fatal bus accident severity: Their impact on different types of bus drivers. Accid Anal Prev 2016; 86:29–39.

[4] Verma A, Khan SD, Maiti J, Krishna OB. Identifying patterns of safety related incidents in a steel plant using association rule mining of incident investigation reports. Saf Sci 2014; 70:89–98.

[5] Cheng CW, Leu SS, Cheng YM, Wu TC, Lin CC. Applying data mining techniques to explore factors contributing to occupational injuries in Taiwan's construction industry. Accid Anal Prev 2012; 48:214–22.

[6] Silva JF, Jacinto C. Finding occupational accident patterns in the extractive industry using a systematic data mining approach. Reliab Eng Syst Safe 2012; 108:108–22.

[7] Rivas T, Paz M, Martín JE, Matías JM, García JF, Taboada J. Explaining and predicting workplace accidents using data-mining techniques. Reliab Eng Syst Safe 2011; 96:739–47.

[8] ERIC. Tanagra, http://eric.univ-lyon2.fr/~ricco/tanagra/fr/tanagra.html [accessed 16.05.17].

[9] Alexe G, Alexe S, Axelrod DE, Bonates TO, Lozina II, Reiss M, Hammer PL. Breast cancer prognosis by combinatorial analysis of gene expression data. Breast Cancer Res 2006; 1–20.

[10] Brauner MW, Brauner N, Hammer PL, Lozina I, Valeyre D. Logical analysis of computed tomography data to differentiate entities of idiopathic interstitial pneumonias. In: Pardalos PM, Boginski VL, Vazacopoulos A, editors. Data Mining in Biomedicine Vol. 7, New York: Springer optimization and its applications; 2007, p. 193–208.

[11] Lauer MS, Alexe S, Pothier Snader CE, Blackstone EH, Ishwaran H, Hammer PL. Use of the logical analysis of data method for assessing long-term mortality risk after exercise electrocardiography. Circulation 2002; 106:685–90.

[12] Yacout S. Fault detection and diagnosis for condition based maintenance using the logical analysis of data. In: IEEE 40th International Conference on Computers and Industrial Engineering (CIE). 2010; 1–6.

[13] Mortada M, Yacout S. cbmLAD – Using logical analysis of data in condition based maintenance. In: 3rd International Conference on Computer Research and Development (ICCRD). 2011; 4:30–4.

[14] Jocelyn S, Chinniah Y, Ouali M-S. Contribution of dynamic experience feedback to the quantitative estimation of risks for preventing accidents: A proposed methodology for machinery safety. Safety Science 2016; 88:64-75.

[15] Kim HH, Choi JY. Hierarchical multi-class LAD based on OvA-binary tree using genetic algorithm. Expert Syst Appl 2015; 42:8134–45.
[16] Boros E, Hammer PL, Ibaraki T, Kogan A, Mayoraz E, Muchnik I. An implementation of logical analysis of data. IEEE Trans Knowl Data Eng 2000; 12:292–306.
[17] Almuallim H, Dietterich TG. Learning Boolean concepts in the presence of many irrelevant features. Artif Intell 1994; 69:279–306.
[18] Ragab A, Yacout S, Ouali MS. Intelligent data mining for automatic face recognition. Online J Sci Technol 2013; 3:97–101.
[19] Hammer PL, Kogan A, Simeone B, Szedmák S. Pareto-optimal patterns in logical analysis of data. Discrete Appl Math 2004; 144:79–102.
[20] Hammer PL, Bonates TO. Logical analysis of data—an overview: from combinatorial optimization to medical applications. Ann Oper Res 2006; 148:203–25.
[21] Mortada MA, Yacout S, Lakis A. Diagnosis of rotor bearings using logical analysis of data. J Qual Maint Eng 2011; 17:371–397.
[22] Ryoo HS, Jang IY. MILP approach to pattern generation in logical analysis of data. Discrete Appl Math 2009; 157:749–61.
[23] Yacout S. Tool and method for fault detection of devices for condition based maintenance. Provisional Patent PCT/CA2011/000876, 2010.
[24] Ragab A, Yacout S, Ouali MS. Interpretable pattern-based machine learning for condition based maintenance. In: RAMS, 2015 The 61st Annual Reliability & Maintainability Symposium. 2015; 1–17.
[25] Chinniah Y. Analysis and prevention of serious and fatal accidents related to moving parts of machinery. Saf Sci 2015; 75:163–73.
[26] Pyle D. Data preparation for data mining. San Francisco, CA, USA: Morgan Kaufmann Publishers, Inc. 1999.
[27] Bennane A, Yacout S. LAD-CBM; new data processing tool for diagnosis and prognosis in condition-based maintenance. J Intell Manuf 2012; 23:265–75.
[28] Bennane A, Yacout S. Processing missing and inaccurate data in a condition based maintenance database. In: IEEE 40th International Conference on Computers and Industrial Engineering (CIE). 2010; 1–5.
[29] Witten IH, Frank E, Hall MA. Chapter 5 – Credibility: Evaluating what's been learned. In: Data Mining Practical Machine Learning Tools and Techniques, 3rd edition, Burlington, MA, USA: Morgan Kaufmann Publishers; 2011, p. 147–190.
[30] Giraud L, Massé S, Dubé J, Schreiber L, Turcot A. Sécurité des convoyeurs à courroie – Guide de l'utilisateur. 2nd Edition. Quebec : CSST (Commission de la santé et de la sécurité du travail). 2003.

Appendix

Table B18: Definition of classes involved

Class	Task in progress when accident occurred	Value
Production-related accident	Production (operation, e.g., sort the material conveyed, pick up an object that fell from the conveyor)	0
Maintenance-related accident	Unjamming, repair or cleaning	1

Table B19: Definition of 23 indicators used in database

Indicator	Definition	Values
I_1	A prevention program exists	No = 0; yes = 1
I_2	Time in position	0 to 4 years = 1; 5–10 years = 2; 20–24 years = 3
I_3	Worker from the company or subcontractor	Company = 1; subcontractor = 2
I_4	Worker's regular activity (task)	No = 0; yes = 1
I_5	Worker is specially trained to use this machine	No = 0; yes = 1
I_6	Accident due to a poor working environment (e.g., cluttered work area)	No = 0; yes = 1
I_7	Causal agent	Nip point = 1; entrapment zone = 2
I_8	Machine functional impairment	No = 0; yes = 1
I_9	Accident happened when machine started unexpectedly	No = 0; yes = 1
I_{10}	Accident happened when machine was functioning in automatic mode	No = 0; yes = 1
I_{11}	Safeguarding was in place at time of accident	No = 0; yes = 1
I_{12}	Flawed or deficient warnings and markings involved in accident	No = 0; yes = 1
I_{13}	Personal protective equipment (PPE), work clothing or tool involved in accident	No = 0; yes = 1
I_{14}	Lockout procedure was applied	No = 0; yes = 1
I_{15}	Control system was involved in accident	No = 0; yes = 1
I_{16}	Accessible hazard zone where accident occurred	Zone between belt or chassis and another structure = 1; pulley = 2; roller = 3; drive drum = 4; tensioning drum = 5; tail drum = 6
I_{17}	Flawed occupational health and safety (OHS) management	No = 0; yes = 1
I_{18}	An OHS committee exists	No = 0; yes = 1
I_{19}	Harm	Crushed forearm = 1; amputated arm = 2; death = 3
I_{20}	Another piece of equipment (in addition to belt conveyor) was involved in accident	No = 0; yes = 1
I_{21}	A missing accessible emergency stop device was involved in accident	No = 0; yes = 1
I_{22}	Accidental circumstance	Crushing = 1; entrapment = 2; impact = 3; strangulation = 4; wrenching = 5; entanglement, with suffocation from engulfment = 6
I_{23}	Part of body in hazard zone and location of injury	Body = 1; torso = 2; head = 3; left arm = 4; right arm = 5; both arms = 6; neck = 7

ANNEXE C – ARTICLE 3

"ESTIMATION OF PROBABILITY OF HARM IN SAFETY OF MACHINERY USING AN INVESTIGATION SYSTEMIC APPROACH AND LOGICAL ANALYSIS OF DATA"

Sabrina Jocelyn[ab][2], Mohamed-Salah Ouali[b], Yuvin Chinniah[b]

[a]*Institut de recherche Robert-Sauvé en santé et en sécurité du travail (IRSST), 505 de Maisonneuve Blvd. West, Montreal, Quebec, Canada H3A 3C2*

[b]*Department of Mathematical and Industrial Engineering, Polytechnique Montréal, 2500 chemin de Polytechnique, Montreal, Quebec, Canada H3T 1J4*

Abstract

In machinery safety, the estimation of the probability of occurrence of harm is a recurrent problem. This paper proposes and applies a new method to estimate that probability. Information regarding accidents involving machinery that is gathered and analyzed by experts is formatted based on a systemic-inspired model using the MELITO concept. Then, Logical Analysis of Data (LAD) is used to extract knowledge automatically to characterize the accidents. MELITO describes the context in which the accident has occurred, gathering information about the moment (M), equipment (E), location (L), individual (I), task (T) and organization (O). LAD is a data mining algorithm that infers knowledge learning from a database. In this paper, a case study consisting of twenty-three fatal and serious accident reports involving belt conveyors is presented. Data about these accidents is classified according to MELITO. The inferred knowledge is presented in the form of interpretable patterns that characterize and distinguish fatalities from non-fatal harms. Each pattern consists of a Boolean equation of MELITO and covers a subset of accidents. Based on each pattern, the probability of occurrence of harm related to a hazardous situation is estimated. Such probability is useful in monitoring the risk behavior after the occurrence of a new accident for instance.

Keywords

Machinery safety; Quantitative risk estimation; Logical Analysis of Data (LAD); Data formatting.

1. Introduction

Risk can be defined in various ways, even within the same field. Villa et al. [1] reviewed five different ways to describe "risk" through the process industry: (1) an expected loss, (2) the

[2]Corresponding author at Tel.: +1 514 288 1551, ext. 407.

Email address: sabjoc@irsst.qc.ca (S. Jocelyn)

probability of an undesirable event, (3) the measurement of an outcome of uncertainty, (4) the potential for a negative consequence, or (5) the combination of an event estimated frequency or probability and its consequence. These authors summarized all of these meanings into one concept made of two notions: consequence and probability. In the safety of machinery, the general ISO 12100:2010 [2] standard states a definition of risk comparable to the one proposed by Villa et al. [1]. Risk is a combination of the probability of occurrence of harm and the severity of that harm [2].

In the workplace, machinery causes fatal and minor or serious non-fatal injuries (e.g. fractures). These injuries happen for different reasons that can be of human, organizational or technical origin. Chinniah [3] analyzed 106 serious and fatal accident investigation reports highlighting various causes related to these aspects, such as: inexperience of workers, unsafe working methods, absence of lockout procedures, lack of risk assessment poor machinery design, and easy access to the moving parts of machinery. A hazard is the potential source of harm [2]. A moving part of the machinery is a mechanical hazard. Other kinds of hazards exist on machines such as thermal (hot or cold temperature), chemical (gases), biological (bacteria), sharp edges, non-compliance to ergonomic principles, radiation and noise [4][6], [5].

To prevent machinery-related accidents, risk management is required. In this paper, an “accident” refers to an unexpected event that leads to harm. To manage the risk, risk assessment is required with the participation of the stakeholders concerned: the employer and the workers, all helped by the safety practitioners. Risk assessment encompasses two main steps: risk analysis and risk evaluation. Risk analysis is particularly based on (1) identification of hazards and risk factors, followed by (2) risk estimation. According to ISO 12100:2010, risk evaluation is the judgment, on the basis of risk analysis, of whether the risk reduction objectives have been achieved [2]. A detailed explanation of risk assessment is available in [6].

Jocelyn et al. [6] notice that machinery-related risk estimation was mainly based on qualitative and static tools. Qualitative tools lead to subjective decision-making. Some criticisms such as: the vagueness in their terminology and the challenges of comparing the degree of protection achieved with that for other everyday risks can explain their subjectivity [1], [7]. With a similar mindset, Hubbard et al. [8] claim that risk assessment should involve the use of mathematical probabilities instead of scoring methods made of verbal labels describing probabilities or impact. For instance, “very likely”, “likely”, “unlikely” and “very unlikely” are proposed to estimate the probability thresholds and “high”, “medium”, and “low” to characterize the impact levels [8]. Hubbard et al. [8] shared an experience in which subjects had to allocate a quantitative value to the verbal labels of probability. They concluded that regardless of whether the verbal labels were detailed, the probability assigned would vary due to different interpretations of the labels. Some rules to design more robust qualitative risk estimation tools that reduce variability in the results have been proposed [9], [10].

A machinery safety report, related to the practical experimentation of risk estimation parameters and tools, revealed that the parameter “Probability of occurrence of harm” estimation is a problematic aspect of the risk estimation and requires special care [9]. Gauthier et al. [9] observed a considerable divergence in the results regarding that parameter. It revealed that probability seems difficult to estimate by the stakeholders [9]. Quantitative risk estimation allows the reduction of subjectivity in the risk management process. Indeed, Villa et al. [1] acknowledge

that, in the field of process industry, quantitative risk assessment proved to be the best available, analytic, predictive tool for risk assessment despite the fact that it does not give an exact description of reality. Hence, as Apostolakis [11] explained, probabilities cannot be realistically calculated.

On the other hand, static tools that are dealt with in machinery-related risk estimation do not allow for the modifications brought to the machine and its surroundings (organizational and environmental) to be captured. Inspired by various techniques from diverse fields like process industry, aerospace, finance, medical, Jocelyn et al. [6] highlighted the importance of updating occupational risks in the field of machinery safety based on quantitative risk estimation. Updating risks is of paramount importance to take any changes in the workplace into account and avoid making outdated decisions. To continually improve the representation of "reality", a dynamic risk assessment method should be favored. Accident analysis and probability updates are key to dynamic risk assessments [1]. However, applying the update of risk is a subject that will be dealt with in a further study. For now, this paper paves the way to such application by proposing a framework with a structure that will allow an update of the risk portrait when a new accident is registered in a database. The risk portrait is given by patterns, their probabilities, the probability of occurrence of harm and the severity of that harm. Patterns are combinations of the main risk factors and possible accident causes. Those combinations constitute knowledge inferred from a database of machinery-related accidents analyzed. The knowledge is inferred by launching a data mining algorithm. Knowledge inference from reported events (e.g. accidents) is dynamic experience feedback. When a new accident happens, the patterns will be updated.

This paper aims to analyze belt-conveyor-related accident investigation reports through dynamic experience feedback to estimate the risk on a quantitative basis. The quantitative aspect based on probability calculations brings objectivity to the risk management process. The data-mining algorithm chosen for the dynamic experience feedback is Logical Analysis of Data (LAD) due to the fact that a small sample of accidents is available to this study. LAD is a pattern-recognition technique that was shown to be suitable for application to scarce data in machinery safety [12]. LAD is fully explained in [12]. The quantitative risk estimation considered in this paper aims to calculate the probability of occurrence of harm using some specific extracted patterns.

Section 2 contains a brief review establishing a benchmark to compare this paper to what exists in the domain. Accordingly, that review lays the foundation for the choice of an adequate method to achieve the accident model. An overall view of the method is described in section 3. Section 4 details how the data formatting of the belt-conveyor-related accident database used for LAD is performed. That database comprises fatal and non-fatal accident data. Section 5 deals with the knowledge inference by showing the patterns generated with LAD. Section 6 explains the calculation of the probability of occurrence of harm associated with a hazardous situation represented by a pattern. Section 7 gives a summary of the results including the probabilities obtained for the occurrence of harm. That section also presents another outcome of the study: a hierarchy of the patterns according to their probability. Such hierarchy is useful to prioritize risk reduction measures. At last, sections 8 and 9 discuss the results and draw conclusions respectively.

2. Review of accident-based quantitative risk estimation in machinery safety

In machinery safety, few studies ([13], [14]) exist on quantitative risk estimation based on accidents. Raviv et al. [13] performed a quantitative risk estimation based on a database of tower crane-related incidents. Initially, the database was qualitative. They transformed it into a quantitative one to apply the k-means data mining algorithm to identify clusters in the data. The *n* clusters to be found enable classifying the incidents (the *n* is not predefined). In the case of their tower crane-related study, Raviv et al. [13] found five clusters. Moreover, they enable calculating the partial risk potential (PRP) of each incident per cluster. The PRP is the potential of an incident for escalating into a circumstantial event of a specific outcome severity: (1) severe damage only (material damage), (2) minor injury, (3) major injury or (4) fatality. The Analytic Hierarchy Process (AHP) was used for expert knowledge elicitation. AHP is a method that breaks down a main issue into smaller ones to facilitate its management. The main issue is about estimating the potential for overall damages to the construction company due to accidents on the construction sites. That issue was subdivided into two groups of issues: (1) overall monetary damage and (2) company reputation damages. Finally, each of these groups was subdivided into the aforementioned four outcome severity levels. The contribution of 16 senior experts with a total of 300 years of experience allowed for the attribution of weights quantifying the impact of the outcome severity levels on the monetary and reputation damages. As risk is defined as the combination of the probability of occurrence of harm and the severity of that harm, the quantitative aspect of this study relies in the product of two matrices: the PRP matrix and the matrix of the aspect grades for the four outcome severity levels. The former has *n* lines corresponding to the *n* clusters and 4 columns regarding the four outcome severity levels. The latter comprises 4 lines corresponding to the 4 outcome severity levels and 3 columns: the first one is related to material damage aspect grade, the second one to the company reputation aspect grade and the third one to the overall grade. Combining the experts' weights with the partial risk potential for incidents of every cluster leads to the total risk potential, used as an index for quantitative risk estimation. Raviv et al. [13] acknowledge that the quantitative aspect does not lead to the ultimate truth about the risk. However, it enables the construction companies to know about their overall safety levels. Also, it can be used "to promote a competitive system for assessing safety-related achievements [13]".

Unlike the study of Raviv et al. [13], this paper excludes damages and focuses only on bodily injuries and fatalities. The data mining algorithm used in this paper to characterize accidents is a supervised learning algorithm called LAD since the classes of accidents were already known but not characterized. On the other hand, Raviv et al. [13] used an unsupervised learning algorithm (k-means) because they were seeking classes (clusters) of accidents to characterize them. Unlike Raviv et al. [13] who estimated the impact of outcome severity levels through experts' elicitation, this paper does not because the levels of severity of harm involved (fatal harm versus non-fatal harm) and their impact on the occupational risk related to a machine are non-ambiguous. It is obvious that fatal harm is more severe than non-fatal harm. Just as reference [13] presented a method transferable to other industries, this paper, based on accidents involving belt conveyors, proposes a method transferrable to any type of machinery in any industry.

Another study on accident-based quantitative risk estimation in machinery safety is that of Aneziris et al. [14]. They propose a method based on a bowtie diagram describing the pathways to its center event, as well as the different scenarios leading to diverse consequences. The center

event is contact with the moving parts of a machine. The consequences of this include: fatalities or recoverable or permanent injuries. The bow tie principle is described as follows: the center event is the undesired event. It is the center of the bow tie. The center event is induced by the failure of elimination barriers (or proactive safety barriers) forming a fault tree on the left side of the bow tie. If the center event happens and is not controlled due to failure of prevention barriers, it will generate further events affecting the population, the environment or the worker. The advantage of this representation is the mapping of diverse scenarios that can lead to a potential accident.

Also, Aneziris et al. [14] calculated the probability of the center event and of its consequences based on the failure probabilities associated with the elimination barriers and prevention barriers preceding the occurrence of the center event and its consequences. They also calculated the risk decrease index of different prevention measures to evaluate their impact in risk mitigation. The uncertainties associated with the risk quantification process are dealt with in another study [15]. The bowtie is built using a software program called *Storybuilder* [16]. The bowtie is a predefined one. When the user reads an accident report, he can go through the bowtie and select the barriers associated with the accident report. The bowtie of Aneziris et al. [14] is based on 3000 accidents related to the center event. Their accident-related data have been supplemented with data describing exposure of the worker to the hazard [14]. Their bow-tie diagram gives a generic accidental sequence. It is a sequential accident model.

Unlike the study by Aneziris et al. [14], this paper proposes a method that can have its parameters changed by the safety practitioner. The method is not static in that it is not fixed in a predefined model. The proposed model can change according to the user's needs, based on the context in which an accident happened, described by a systemic approach. The model is inferred from analyzed accident investigation reports using an artificial intelligence algorithm called LAD, while the one proposed by Aneziris et al. [14] was designed manually based on the analysis of accident investigation reports. Once their bow-tie conceptual model was built, further accident reports were forced to be described according to the model frame. However, the proposed method suggests a model (patterns) which conceptual part can be changed as new accidents are added to the initial database. Moreover, the method does not follow a sequential approach, as the one proposed by Aneziris et al. [14]. In *Storybuilder*, only the declared causes, having a cause-and-effect relationship with the accident, were selected to build their bowtie conceptual model. In this paper, all aspects of the workplace at the moment of the accident were considered in the conceptual model, whether the accident investigator identified it as a proven cause, a possible cause, or not a cause at all. Such a choice allows for a closer portrait of reality. Moreover, the decision to embrace all aspects of the accidental situation in a non-sequential approach is based on the visions of Dekker [17], Leveson [18] and Hollnagel [19].

Like Dekker [17] and Leveson [18], Hollnagel [19] is critical about sequential approaches to describe accidents. Hollnagel [19] underlines that the cause-and-effect assumption is perhaps the least attractive option [19]. Accordingly, cause-and-effect accident models are restrictive. Moreover, Hollnagel defines accident models as a frame of reference, which is a stereotypical way of thinking about how an accident occurs [19]. He states three types of accident models:

1) **Sequential accident models.** In that type of model, "a sudden, unexpected event initiates a sequence of consequences where the last one is the accident" [19]. The aim of the accident analysis is to eliminate or contain accident causes after identifying the cause-

and-effect links [19]. Fault trees, domino and network models (e.g. Bayesian networks) are some examples of such an accident model.

2) **Epidemiological accident models.** This is reminiscent of sequential accident models, but some features differ in this type of model from the sequential one. For instance, the barrier notions and latent conditions are some of these features. Epidemiological accident models describe an accident as a spreading phenomenon in which the consequence is "a combination of factors, some manifest and some latent, that happen to exist together in space and time" [19]. The goal of this type of model is to make defenses and barriers stronger [19].
3) **Systemic accident models.** These "models endeavor to describe the characteristic performance on the level of the system as a whole, rather than on the level of specific cause-and-effect 'mechanisms' or even epidemiological factors" [19]. The purpose of a systemic accident model is to monitor and control the performance variability of a system [19], which fits the aim of this paper in allowing risk update.

In this paper, a systemic-inspired accident model is used. The chosen model is based on a concept widely used by accident investigators in Quebec: the MELITO concept [20]. MELITO is used because inspectors investigated the accidents from the reports analyzed in this study according to that concept. MELITO consists of explaining the context in which the accident occurred by describing the moment (M), the equipment (E), the location (L), the individual (I), the task (T) and the organization (O) related to the accident. According to reference [20]:

- "M" includes, for example: the time of the accident, whether the employee was working overtime;
- "E" concerns the type of machine, tool, substance or any object involved in the accident. It considers, for example, whether the machine was in adequate condition, and whether there is an inspection program for that piece of equipment at the enterprise involved in the accident;
- "L" includes the place where the accident happened, whether that place was an usual one for the employee to be working, whether the area was cluttered, and the environmental temperature;
- "I" includes, for example, whether a certification is required for the activity, or if the worker was performing his usual or a new activity at the time of the accident, or what was his level of experience regarding this activity;
- "T" refers to the activity the worker was performing at the time of the accident. It also includes whether the worker was trained to perform the job, or whether the worker applied the safety rules regarding his job.
- "O" includes, for example, the supervision means in the enterprise and the existence of safety standards or a training program at the enterprise.

Basing this paper on MELITO makes the approach clearer to safety organizations already familiar with that way of considering accidents. However, concepts other than MELITO exist for accident investigation reports. For instance, the Canadian Centre for Occupational Health and Safety (CCOHS) [21] suggests another concept consisting of grouping accident causes into five categories: (1) Task, (2) Material, (3) Environment, (4) Personnel, and (5) Management. It is similar to MELITO except that the "moment" aspect is missing. In the United States of America (U.S.A.), the Occupational Safety and Health Administration (OSHA) suggests the "Who? Where? What? Why? When? How?" concept for investigating accidents [22]. In the United

Kingdom (U.K.), the Health and Safety Executive (HSE) presents a concept similar to that in the U.S.A. based on the six previous questions [23]. In France, the Institut national de recherche et de sécurité (INRS), an occupational health and safety research institute, proposes the ITAMaMi concept [24]: (1) *I*: *individu* which means the individual, (2) *T/A*: *tâche/activité* for task/activity, (3) *Ma*: *matériel* for material, and (4) *Mi*: *milieu* for environment. Since "Mi" considers not only the physical environment where the accident occurred, but also the psychosocial one, compared to MELITO the only aspect excluded is the moment one.

The core of the systemic-inspired accident model proposed relies on the data formatting in the accident database from which patterns are to be generated. The data formatting method relying on MELITO is transferable to any accident investigation concept. For example, transposing the MELITO formatting to ITAMaMi will require gathering data in the group "Mi" pertaining to the "L" and "O" of MELITO. Furthermore, in that case, data describing the "M" in MELITO will disappear because the moment is absent in the ITAMaMi concept.

3. Case study and methodological framework

To reach the aim of this paper, the following four-stage framework is proposed (Figure C1).

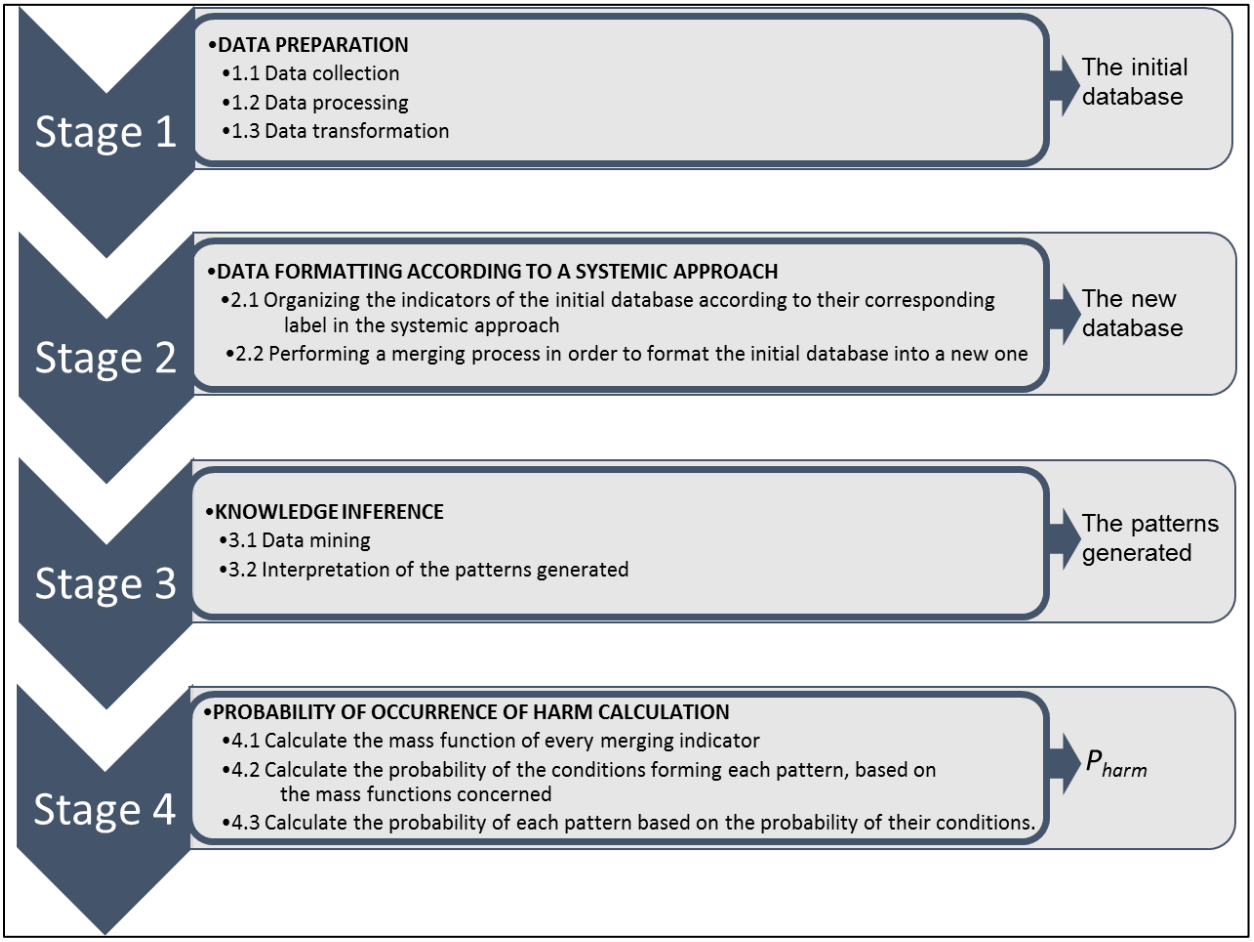


Figure C1: An overall view of the four-stage method

Stage 1. Data preparation. Prior to implementing the MELITO systemic approach, data had to be collected, processed, then transformed. The data collection involves gathering 23 occupational accident investigation reports regarding belt conveyors: 19 fatal accidents and 4 non-fatal ones. The fatal and non-fatal accidents are disjoint sub-spaces forming the space of belt-conveyor-related accidents as shown in Figure C2. Every number mentioned in these sub-spaces represents one of the 23 accidents. The inscription: "…" means that every sub-space is made of unknown belt-conveyor-related accidents that happened in addition to those registered in the database.

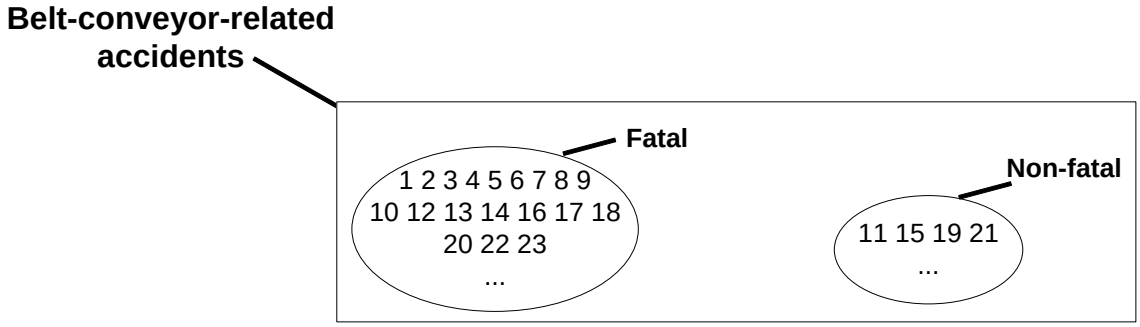


Figure C2: The space of belt-conveyor-related accidents made of fatal and non-fatal accidents

The reports are from the Documentation Center of the CNESST (*Commission des normes, de l'équité, de la santé et de la sécurité du travail*) [25]. The challenge in the data collection is three-fold:

- To document all aspects of the accidental context described textually in the reports, the reader must analyse, then translate the textual information into data. The data are the values attributed to the indicators of the accidents analyzed. Indicators are variables whose values describe each accident. These variables represent accident causes or risk factors describing the accidental context. Even though the main causes of the accidents are highlighted in the reports, sub-causes and risk factors are hidden in the text. The goal of the analysis is to retrieve those causes as well as the implicit sub-causes and risk factors describing the accidental context and formalize them in a database. Information analysis from the reports is processed manually in the form of data structured in a database (see the initial database in Figure C3).
- Furthermore, the analysis reveals which aspect of the MELITO concept an indicator belongs to. Although the inspectors use MELITO as an accident investigation tool, the information contained in the accident investigation reports is not presented in a MELITO format. Usually, the reports are comprised of the following sections: an abstract, a description of the work organisation, a description of the activity undertaken during the accident, the facts and analysis of the accident, and a conclusion.
- Although the reports cover the six aspects of the MELITO concept, the level of detail from one report to another varies depending on the year of publication or the inspector who wrote it. Indeed, each report does not always share the same indicators as the others. Therefore, a common basis had to be found by processing the data by dealing with missing values and keeping the relevant indicators, namely those shared by all of the reports.

The data processing is similar to that of reference [12]. In the database, the accidents are divided into two classes: “Fatal” and “Non-fatal” defined in Table C7. The class “Fatal” is labeled with number 1 in the database, whereas the class “Non-fatal” is labeled with 0. The database used for this paper initially encompassed 23 indicators (having the number of indicators equaling the number of accidents is pure coincidence). At the very beginning, every indicator in the database had discrete qualitative possible values. Since LAD requires numerical data to generate patterns, each discrete qualitative value had to be transformed into an ordinal scale.

Stage 2. Data formatting according to a systemic approach. The systemic approach used is the MELITO concept. The data formatting consists of:

- Organizing the 23 initial indicators according to their corresponding label: “M”, “E”, “L”, “I”, “T” or “O” (see Table C8). These labels are merging indicators;
- Performing a merging process in order to format the initial 23×23 database into a new 23×6 one (23 accidents and 6 merging indicators, see Table C8).

Stage 3. Knowledge inference with LAD. Knowledge inference consists of generating interpretable patterns from capitalized data. To alleviate the variability from the content of the 23 reports, knowledge inference gives a common and overall description of the context of the accidents analyzed, extracting the essentials from these reports. In this study, the patterns are inferred using LAD. That algorithm is chosen for data mining because it was shown in [12] to be adequate for knowledge extraction with scarce data using the 23 accidents dealt with in the current paper.

A pattern is a set of conditions that characterizes and distinguishes a specific class of a phenomenon against others (e.g. class of accidents). The conditions are intervals of values associated with the indicators selected by the data mining algorithm.

Stage 4. Probability of occurrence of harm calculation. To enable risk calculation and risk monitoring, the database allows the two principal parameters of risk to be taken into account:

- Probability of occurrence of harm;
- Severity of that harm.

The severity of harm per hazardous situation is represented by the class allocated to the accident in the database. The probability of occurrence of harm is the probability of each hazardous situation represented by a pattern generated with LAD. Such probability depends on the distribution of each indicator’s values per class. Section 6 explains the probability calculation process.

4. Data formatting

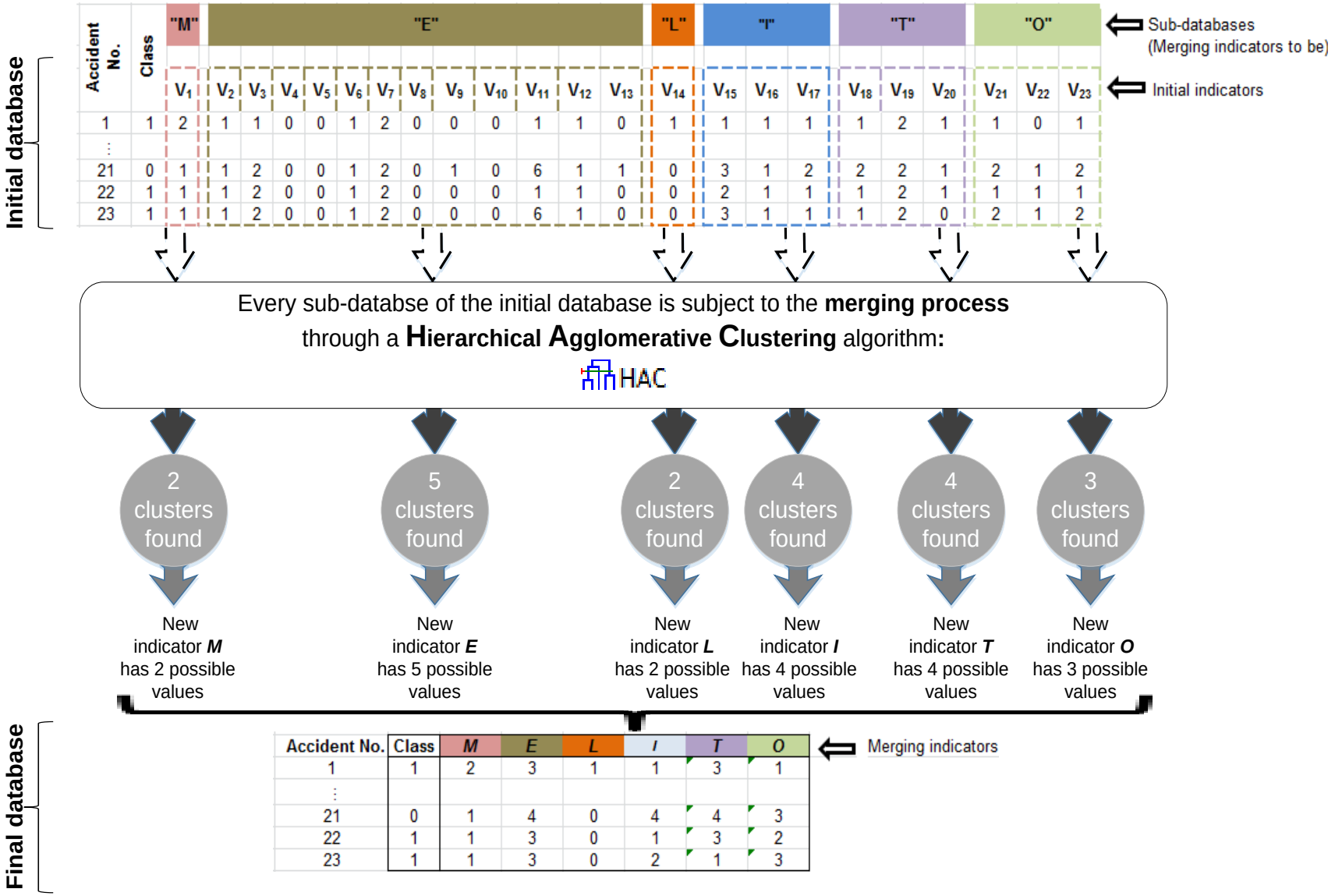
The numerical values in the initial database were allocated according to the following logic for the purpose of this study (see the Appendix for details):

- The highest numerical value was attributed to the most dangerous discrete qualitative value of the indicator;
- The lowest numerical value was allocated to the least dangerous one.

Having the values of the 23 initial indicators allocated in respect to that logic systematically guides the allocation of values to the merging indicators as will be illustrated in Table C1.

As this study pursues a systemic-inspired accident model based on the MELITO concept, the number of indicators had to drop from 23 to 6. Addressing the issue of merging 23 indicators with known values into 6 new indicators with unknown values is not obvious to accomplish manually. Therefore, the following two-step procedure is adopted:

- The first step consists of grouping the indicators (see the 6 sub-databases in Figure C3) with respect to a definition according to the MELITO concept described in section 2 and detailed in [20]. For instance, indicator V_1 expresses the shift when the accident occurred and constituted the sub-database describing the moment "M". Then, indicators V_2 to V_{13} described in Table C8 are grouped to form the sub-database defining the equipment "E", and so on.
- The second step seeks the allocation of possible values to each merging indicator of M-E-L-I-T-O. To do that, the number of possible values per merging indicator is assumed to be equal the number of clusters composing each sub-database. Every cluster characterizes the merging indicator in a specific way, based on the values of the initial indicators. However, the number of clusters per sub-database was unknown and difficult to detect visually due to the amount of indicators and their various values, except for the M and L indicators, with only one initial indicator and two possible values. Consequently, an unsupervised learning algorithm called Hierarchical Agglomerative Clustering (HAC) is used to identify the clusters in each sub-database. The HAC algorithm is launched on *Tanagra* [26].



Accident No.	Class	M	E	L	I	T	O
1	1	2	3	1	1	3	1
⋮							
21	0	1	4	0	4	4	3
22	1	1	3	0	1	3	2
23	1	1	3	0	2	1	3

Figure C3: Data formatting – Organization of the initial indicators and the merging process towards the MELITO concept

The allocation of values to the 6 merging indicators is detailed in Table C1, through the example of initial indicators V_{15} to V_{17} describing the "I" of MELITO. *Tanagra* found 4 clusters forming the sub-database regarding "I": *HAC_1* to *HAC_4*. Afterwards, how can a meaning be given to these four clusters? Knowing that the values of the initial indicators were ranked from the least to the most dangerous, so would the values of each merging indicator. The following artifice is used to address that issue:

- The mean of the values per initial indicator was calculated for each cluster (see $\overline{V_i}$ in Table C1). These mean values are given by *Tanagra* as components of the centroid of the cluster. The centroid is the middle point of a cluster;
- The summation of these means per cluster would guide how to rank the cluster (see $\sum \overline{V_i}$ in Table C1).

Table C1: Allocation of values to the merging indicator *I*

Accident No.	Class	V_{15}	V_{16}	V_{17}	Cluster	Corresponding ordinal value
7	1	3	1	2		
10	1	3	1	2		
13	1	3	1	2		
15	0	3	1	2	*HAC_1*	**4**
16	1	3	1	2		
20	1	3	1	2		
21	0	3	1	2		
		$\overline{V_{15}} = 3$	$\overline{V_{16}} = 1$	$\overline{V_{17}} = 2$	$\sum \overline{V_i} = 6$	
3	1	3	2	1		
4	1	2	2	1	*HAC_2*	**3**
11	0	3	2	1		
17	1	3	2	1		
		$\mathbf{\overline{V_{15}} = 2.75}$	$\mathbf{\overline{V_{16}} = 2}$	$\mathbf{\overline{V_{17}} = 1}$	$\mathbf{\sum \overline{V_i} = 5.75}$	
6	1	3	1	1		
8	1	3	1	1		
12	1	3	1	1	*HAC_3*	**2**
14	1	3	1	1		
19	0	3	1	1		
23	1	3	1	1		
		$\mathbf{\overline{V_{15}} = 3}$	$\mathbf{\overline{V_{16}} = 1}$	$\mathbf{\overline{V_{17}} = 1}$	$\mathbf{\sum \overline{V_i} = 5}$	
1	1	1	1	1		
2	1	2	1	1		
5	1	2	1	1	*HAC_4*	**1**
9	1	2	1	1		
18	1	2	1	1		
22	1	2	1	1		
		$\mathbf{\overline{V_{15}} \approx 1.83}$	$\mathbf{\overline{V_{16}} = 1}$	$\mathbf{\overline{V_{17}} = 1}$	$\mathbf{\sum \overline{V_i} \approx 3.83}$	

Based on those quantities, the clusters can be ranked from the least to the most dangerous: *HAC_4* < *HAC_3* < *HAC_2* < *HAC_1*. Hence, an ordinal value is allocated to each cluster. Consequently, the merging indicator "I" has 4 possible ordinal values. The meaning of these new values as well as those of the other merging indicators are available in Table C2. Their meaning is expressed in terms of the rounded mean values of their corresponding initial indicators.

For instance, Table C1 has shown the following mean values associated with cluster *HAC_2* representing $I = 3$: $\overline{V_{15}} = 2.75$, $\overline{V_{16}} = 2$, $\overline{V_{17}} = 1$. The rounded mean values of these three initial indicators are: $V_{15} = 3$, $V_{16} = 2$, $V_{17} = 1$. Therefore, having the merging indicator $I = 3$ can be interpreted as follows: the individual (i.e. the victim of the accident) had worked for 0 to 4 years ($V_{15} = 3$) as a subcontractor ($V_{16} = 2$) and was doing a regular activity when the accident happened ($V_{17} = 1$).

The meaning of the rounded mean values of corresponding initial indicators can be interpreted using the corresponding qualitative values described in Table C8 (see the Appendix). Table C2 presents some examples of cluster interpretations.

Table C2: Meaning of the merging indicators possible values

Merging indicator	**Possible values**	$\sum \overline{V_i}$	**Meaning in terms of the rounded mean values of corresponding initial indicators**	**Interpretation examples: the accident happened…**
M	1	1.00	$V_1 = 1$	…during day shift
	2	2.00	$V_1 = 2$	
E	1	8.00	$V_2 = 3$; $V_3 = 1$; $V_4 = 0$; $V_5 = 0$; $V_6 = 1$; $V_7 = 2$; $V_8 = 0$; $V_9 = 0$; $V_{10} = 0$; $V_{11} = 1$; $V_{12} = 0$; $V_{13} = 0$	…on a vertically mobile belt conveyor, in an entrapment zone (a zone between belt or chassis and another structure), while the machine had no functional impairment and no unexpected start, was running in automatic mode with no safeguarding in place, while there were no flawed or deficient warnings and markings involved, no personal protective equipment, work clothing or tool involved, no control system involved, no other piece of equipment (in addition to belt conveyor) involved, no missing accessible emergency stop device involved.
	2	10.00	$V_2 = 1$; $V_3 = 1$; $V_4 = 0$; $V_5 = 1$; $V_6 = 0$; $V_7 = 1$; $V_8 = 0$; $V_9 = 0$; $V_{10} = 1$; $V_{11} = 3$; $V_{12} = 0$; $V_{13} = 0$	
	3	10.67	$V_2 = 1$; $V_3 = 2$; $V_4 = 1$; $V_5 = 0$; $V_6 = 1$; $V_7 = 2$; $V_8 = 0$; $V_9 = 0$; $V_{10} = 0$; $V_{11} = 4$; $V_{12} = 1$; $V_{13} = 0$	
	4	12.83	$V_2 = 1$; $V_3 = 2$; $V_4 = 0$; $V_5 = 0$; $V_6 = 1$; $V_7 = 2$; $V_8 = 0$; $V_9 = 1$; $V_{10} = 0$; $V_{11} = 5$; $V_{12} = 0$; $V_{13} = 1$	
	5	13.00	$V_2 = 1$; $V_3 = 2$; $V_4 = 0$; $V_5 = 1$; $V_6 = 0$; $V_7 = 2$; $V_8 = 1$; $V_9 = 0$; $V_{10} = 1$; $V_{11} = 5$; $V_{12} = 0$;	

			$V_{13} = 0$	
L	0	0.00	$V_{14} = 0$	…despite the absence of a poor working environment
	1	1.00	$V_{14} = 1$	
I	1	3.83	$V_{15} = 2$; $V_{16} = 1$; $V_{17} = 1$	…while the victim was a worker of the company with 5-10 years in the position and performing a regular activity
	2	5.00	$V_{15} = 3$; $V_{16} = 1$; $V_{17} = 1$	
	3	5.75	$V_{15} = 3$; $V_{16} = 2$; $V_{17} = 1$	
	4	6.00	$V_{15} = 3$; $V_{16} = 1$; $V_{17} = 2$	
T	1	3.33	$V_{18} = 1$; $V_{19} = 2$; $V_{20} = 0$	…while the worker was specially trained to use the machine, while no lockout procedure was applied, which is sensible since a production activity was being performed.
	2	4.00*	$V_{18} = 2$; $V_{19} = 1$; $V_{20} = 1$	
	3	4.00*	$V_{18} = 1$; $V_{19} = 2$; $V_{20} = 1$	
	4	5.00	$V_{18} = 2$; $V_{19} = 2$; $V_{20} = 1$	
O	1	2.00	$V_{21} = 1$; $V_{22} = 0$; $V_{23} = 1$	…despite the existence of a prevention program, despite the absence of a flawed occupational health and safety management, and despite the presence of an OHS committee.
	2	3.30	$V_{21} = 1$; $V_{22} = 1$; $V_{23} = 1$	
	3	4.89	$V_{21} = 2$; $V_{22} = 1$; $V_{23} = 2$	

* Even though these two clusters have the same $\sum \overline{V_i}$ they are different, based on their corresponding initial indicators values. $T = 2$ means that an accident occurred under the following conditions related to the task: the worker was not specially trained to use the machine (i.e. $V_{18} = 2$), a lockout procedure was applied (i.e. $V_{19} = 1$) under maintenance operation (i.e. $V_{20} = 1$). Whereas $T = 3$ means that an accident occurred while the worker was specially trained to use the machine (i.e. $V_{18} = 1$), no lockout procedure was applied (i.e. $V_{19} = 2$) under maintenance operation (i.e. $V_{20} = 1$). The comparison between the values of the corresponding initial indicators allows for the ranking of these clusters based on the fact that it is normally safer to work on a machine free of hazardous energy thanks to a lock-out procedure than on one where no lockout procedure is applied.

5. Knowledge inference

After the merging process, the MELITO database for knowledge extraction is built. Afterwards, LAD-WEKA [27] is launched to generate the patterns characterizing and distinguishing the two classes of accidents: “Fatal” was considered as the positive class and “Non-fatal” as the negative class. Tables C3 and C4 present the positive and negative patterns obtained for these two classes. The extracted patterns constitute the knowledge hidden within the MELITO database. The patterns are ranked by their coverage, which calculates a ratio of the number of accidents sharing the same set of conditions over the total number of accidents.

Table C3: Patterns generated by LAD-WEKA for the "Fatal" class ranked by their coverage

Pattern					Accidents covered	Coverage
P_6^+ =		($E \leq 4.5$) AND		($O \leq 2.5$)	1 2 3 4 5 7 12 13 16 17 20 22	12/19
P_5^+ =		($E \leq 4.5$) AND	($I \leq 2.5$)		1 2 5 6 8 9 12 14 18 22 23	11/19
P_2^+ =		($E \leq 3.5$) AND		($O \leq 2.5$)	1 2 3 4 5 7 13 16 17 20 22	11/19
P_4^+ =	($M \leq 1.5$) AND	($E \leq 3.5$)			3 5 7 9 13 16 17 18 22 23	10/19
P_3^+ =		($E \leq 3.5$) AND	($I \leq 2.5$)		1 2 5 8 9 18 22 23	8/19
P_7^+ =			($I \leq 2.5$) AND	($O > 2.5$)	6 8 9 14 18 23	6/19
P_1^+ =		($E > 4.5$)			10	1/19

Table C4: Patterns generated by LAD-WEKA for the "Non-fatal" class ranked by their coverage

Pattern					Accidents covered	Coverage
P_3^- =		($E \leq 4.5$) AND	($I > 2.5$) AND	($O > 2.5$)	11 21	2/4
P_7^- =	($M \leq 1.5$) AND	($E > 3.5$) AND		($O \leq 2.5$)	15 19	2/4
P_1^- =		($E = 4$) AND	($I > 2.5$)		15 21	2/4
P_5^- =	($M \leq 1.5$) AND	($E = 4$) AND		($O \leq 2.5$)	15 19	2/4
P_{11}^- =		($E \leq 3.5$) AND	($I > 2.5$) AND	($O > 2.5$)	11	1/4
P_4^- =	($M \leq 1.5$) AND	($E \leq 4.5$) AND	($I > 2.5$) AND	($O > 2.5$)	21	1/4
P_6^- =		($E > 3.5$) AND	($I > 2.5$) AND	($O \leq 2.5$)	15	1/4
P_{12}^- =	($M \leq 1.5$) AND	($E > 3.5$) AND	($I > 2.5$) AND	($O \leq 2.5$)	15	1/4
P_2^- =	($M \leq 1.5$) AND	($E = 4$) AND	($I \leq 2.5$) AND	($O \leq 2.5$)	19	1/4
P_{10}^- =	($M > 1.5$) AND		($I > 2.5$) AND	($O > 2.5$)	11	1/4
P_8^- =		($E = 4$) AND	($I > 2.5$) AND	($O > 2.5$)	21	1/4
P_9^- =	($M > 1.5$) AND	($E \leq 3.5$) AND	($I > 2.5$) AND	($O > 2.5$)	11	1/4

6. Probability of occurrence of harm estimation

The probability of occurrence of harm associated with a hazardous situation is the probability of the pattern representing that situation and is given by the probability of the intersection of its conditions. For example:

$$\mathcal{P}[P_6^+] = \mathcal{P}[(E \leq 4.5) \cap (O \leq 2.5)] \quad (1)$$

The probability of each condition forming a pattern is calculated based on the mass function $p_Y(y)$ of its corresponding merging indicator. y represents the value of the merging indicator. Y represents the merging indicator as a random variable associated with that value. $p_Y(y) = \mathcal{P}[Y = y]$ is the mass distribution of the value y in the database. Figure C4 illustrates the mass distributions of the six merging indicators. Every bar in each histogram corresponds to the probability of each possible merging indicator's value. That probability equals the relative frequency of the merging indicator's value. In other words, it is the ratio between the absolute frequency of that value in the database and the number of accidents in the database. For instance, the value 2 of merging indicator E appears three times in the database of 23 accidents. Therefore: $\mathcal{P}[E = 2] = 3/23$ which explains the value 0.13 on the top of the histogram corresponding to $E = 2$.

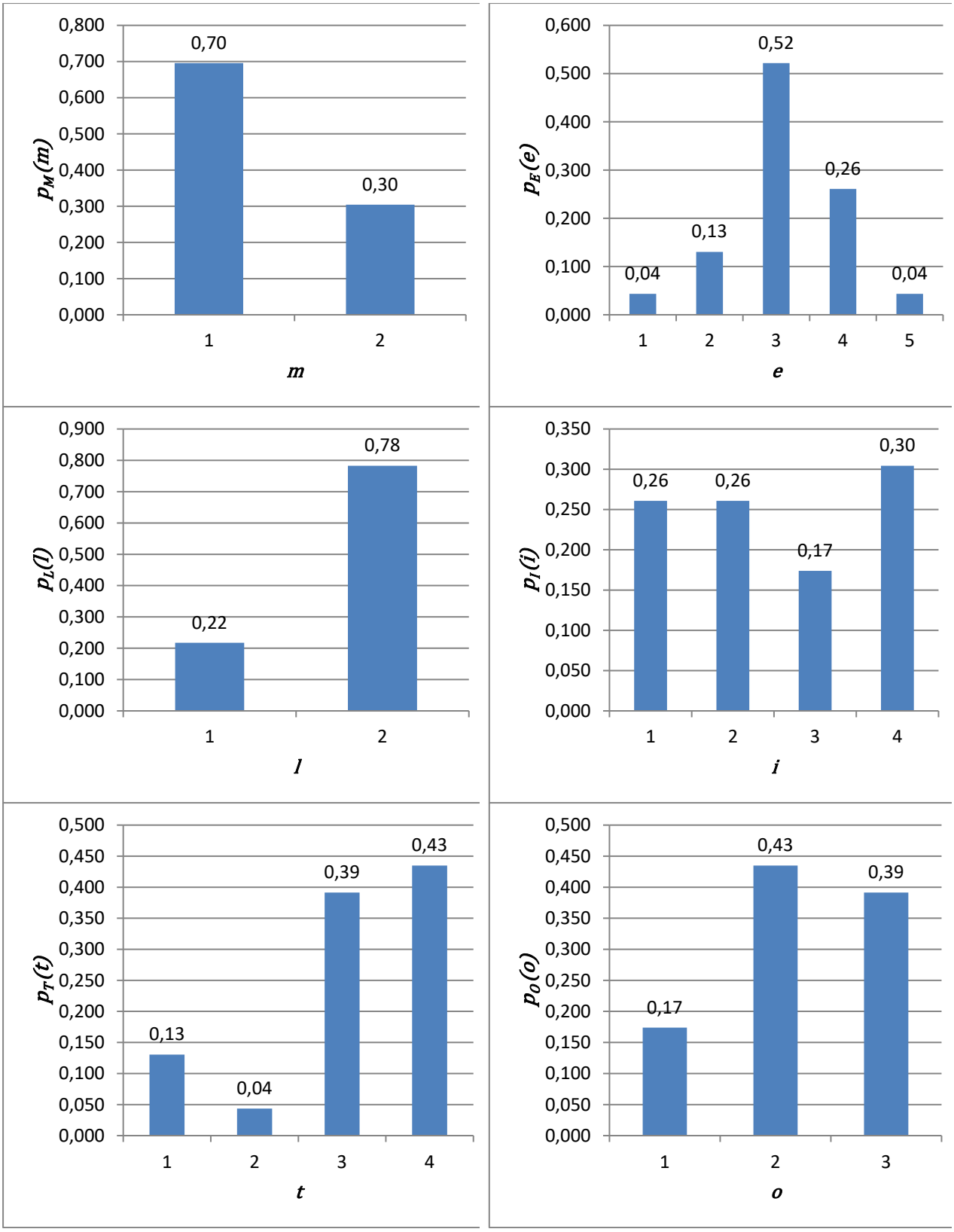


Figure C4: Mass distributions of the six merging indicators *M*, *E*, *L*, *I*, *T*, and O^{**}

[**]The sum of the probabilities in a mass distribution must equal 1. When adding up the probabilities displayed on histograms "E", "I", "T" and "O", their summation lacks 0,01 for each chart. The rounded values of the probabilities explain that lack. However, adding up the full values of the probabilities per histogram leads to 1.

Developing equation (1) based on the mass distributions of merging indicators *E* and *O*, the probability of pattern P_6^+ is:

$$\mathcal{P}[P_6^+] = \mathcal{P}\left[\left((E = 1) \cup (E = 2) \cup (E = 3) \cup (E = 4)\right) \cap \left((O = 1) \cup (O = 2)\right)\right]$$

It is noted that:

- The intersection of the values of a merging indicator is empty because the values are disjoint (i.e. they cannot happen at the same time). Hence, the probability of the union of these values is a simple summation as shown hereinafter;
- It is considered that the intersection of the conditions forming a pattern is an empty set. That explains the simple product between the probabilities of the following conditions: $E \leq 4.5$ and $O \leq 2.5$ instead of applying a conditional probability formula.

$$\mathcal{P}[P_6^+] = (\mathcal{P}[E = 1] + \mathcal{P}[E = 2] + \mathcal{P}[E = 3] + \mathcal{P}[E = 4]) \times (\mathcal{P}[O = 1] + \mathcal{P}[O = 2])$$
$$\approx (0.043 + 0.130 + 0.522 + 0.261) \times (0.174 + 0.435)$$
$$\mathcal{P}[P_6^+] \approx 0.582$$

The calculation of the other patterns' probabilities is carried out according to the same reasoning as for pattern P_6^+. The probability of occurrence of fatal harm and that of non-fatal harm associated with every hazard situation (i.e. pattern) are presented in section 7. Apart from allowing a calculation of the probability of occurrence of harm, the pattern probability supports their ranking in case they have the same coverage. The importance of a pattern is given by its coverage. However, when patterns have the same coverage and, from a risk management perspective, if one needs to determine which risk factor combination is the priority, another criterion is required to establish a hierarchy. The authors chose the pattern probability presented in section 7 as the other criterion.

7. Summary of the results

Data formatting based on a merging process of a 23×23 belt-conveyor-related-accident database into a new 6×6 database using the MELITO concept is proposed in section 4. From the new database, LAD-WEKA generated 7 patterns P_i^+ characterizing the fatal accidents and 12 patterns P_i^- characterizing the non-fatal accidents. The patterns generated are presented in section 5. The average classification accuracy of the patterns generated is 78% according to LAD-WEKA. That percentage is calculated automatically by the software program that used a "10-fold Cross-Validation" procedure to test the accuracy of the patterns.

The probability of every pattern is presented in Table C5 and Table C6, and represents the probability of harm related to a hazardous situation.

Table C5: “Fatal” patterns ranked from the most (left) to the least (right) important.

P_i^+	P_6^+	P_5^+	P_2^+	P_4^+	P_3^+	P_7^+	P_1^+
$\mathcal{P}[P_i^+]$	0.582	0.499	0.423	0.484	0.363	0.204	0.043

Table C6: “Non-fatal” patterns ranked from the most (left) to the least (right) important

P_i^-	P_3^-	P_7^-	P_1^-	P_5^-	P_{11}^-	P_4^-	P_6^-	P_{12}^-	P_2^-	P_{10}^-	P_8^-	P_9^-
$\mathcal{P}[P_i^-]$	0.179	0.129	0.125	0.110	0.130	0.125	0.089	0.062	0.058	0.057	0.049	0.040

8. Discussion – Contribution to risk management

8.1 What do the patterns reveal?

Considering the enterprises involved in the accidents associated with the database of this paper, it appears that the moment (*M*), the equipment (*E*), the individual (*I*) and the organization (*O*) explain mainly the occurrence of accidents. The location (*L*) and the task (*T*) are absent from the patterns. However, the absence of these two merging indicators does not mean that they have nothing to do with the accidents, nor that no risk reduction measures are required for them. The conditions forming a pattern represent the essentials of the knowledge that explain the accidents. Therefore, conditions related to *L* and *T* are not part of that essential knowledge for the 23 accident investigation reports analyzed. The patterns generated for these accidents indicate that in order to carry out a targeted risk prevention, one must tackle merging indicators *M*, *E*, *I* and *O* first, because they discriminate between fatal and non-fatal accidents. The word “targeted” refers here to (1) the detection (identification) of the indicators (risk factors and potential accident causes) that characterize and distinguish one type of accident from another, and (2) the choice of appropriate risk reduction measures suggested by these indicators [12]. The patterns inform the safety practitioner what indicators in the workplace to focus on.

Analyzing the recurrence of the merging indicators in the set of positive and negative patterns generated, one can notice that the merging indicators forming them can be ranked in descending order of their frequency *f*:

- For the “Fatal” class: *E* (*f* = 6) > *I*, *O* (*f* = 3) > *M* (*f* = 1);
- For the “Non-fatal” class: *E*, *O* (*f* = 11) > *I* (*f* = 10) > *M* (*f* = 7).

The equal frequencies of *I* and *O* in the “Fatal” class, and of *E* and *O* in the “Non-fatal” class disturb the hierarchy sought after. Indeed, it is impossible to take care of two indicators at once. To overcome this obstacle, the hierarchy of the indicators among the whole set of patterns will be preferred: *E* (*f* = 17) > *O* (*f* = 14) > *I* (*f* = 13) > *M* (*f* = 8). This hierarchy reveals that to reduce the belt-conveyor-related risks in the enterprises where the analyzed accidents occurred, the safety practitioner must tackle the indicators describing the merging indicator *E* first, then the one describing the organization (*O*), then the individual (*I*), and at last, the moment (*M*). Nevertheless, concerning the moment (*M*), it is inconceivable to decide that workers will be off on the shifts revealed by the patterns. However, the safety practitioners can alert the workers of the enterprises concerned by the accident analysis that accidents happen more often on day shifts in their case. Since night shifts were considered to be more dangerous than day shifts in this study, the safety practitioners must question the higher recurrence of day shifts as a condition to fatal and non-fatal accidents. He or she must question whether the accident happened at a time

when the worker had to readapt himself to day shifts after a period of night shifts, for example. If yes, a transition facilitating the worker's adaptation to shift switches must be planned by the enterprise. After all, the switch from night to day shifts takes more than 6 days and depends on the individual [29].

Merging indicator *E* represents all aspects related to the equipment involved in the accident. Having *E* as the most recurrent in the set of patterns reveals that the risk reduction process shall start decreasing the value of the indicators describing *E* (i.e. V_2 to V_{13} defined in table C8) in the enterprises concerned by the accidents. That result for belt-conveyor-related accidents concurs with the finding of Raviv et al. [13] in the tower-crane-related domain, noticing that technical factors appear to be the most hazardous. It also reveals the compliance with the risk reduction hierarchy established by the machinery safety standard ISO 12100:2010 [2]. That standard states that the risk reduction process shall start with an inherently safe design of the machine (that implies taking action in order to reduce the values of equipment-related indicators: V_2 to V_6, V_{10}, and V_{11} whenever possible), followed by safeguarding (reducing V_7's value) then the complementary protective measures (reducing V_{13}'s value) if the residual risk cannot be eliminated at the design step. Merging indicator *E* emphasizes the importance for the designer and the integrator to design, respectively, a machine (here, belt conveyors) and a work environment (e.g. working on V_{12} for a safe factory layout) free of hazards, providing the user with information for using the machine (working on V_8). Once the designer and the integrator have done their part, it is the user's duty to implement further protective measures such as the provision and use of additional safeguards (V_7) and use of personal protective equipment (V_9) [2]. Afterwards, the user will implement protective measures of organizational then individual types, such as:

- Implementing and updating a comprehensive prevention program (V_{21}) that is well understood by the workers, with safety procedures, training and responsibilities thoroughly explained in it. These recommendations are supported by a previous study [12]. It showed that the majority ($^{12}/_{20}$) of maintenance-related accidents owing to belt conveyors happened while there was a prevention program. The fact that the prevention program was not updated or had some flaws such as: failure to implement a lockout program or new workers' training, or unclear responsibilities [12] were some of the reasons explaining the accidents;
- Preferring workers from the company than subcontractors to do the job. If impossible, the subcontractors must be taken care of by informing them about the communication policies and risks inherent to their job in the enterprise.

One should note that less efficient risk reduction measures regarding the ISO 12100:2010 [2] standard are organizational and human measures. Therefore, effort must be emphasized on risk reduction measures related to the equipment as revealed by the frequency of merging indicator *E* in the set of patterns and the efficiency stated by ISO 12100:2010 [2] concerning safeguarding.

8.2 Guidance for a more objective risk reduction

As explained in section 1, to reduce the risk, risk estimation is required. Since risk estimation in machinery safety is mainly based on qualitative tools, the risk reduction process is subjective. To help overcome that subjectivity, a previous study [12] suggested a risk estimation process based first on the relative prevalence (i.e. coverage in percentage) of LAD-generated patterns and

second on the frequency of appearance of their conditions. However, in this paper, according to Table C3 and Table C4, many patterns have the same coverage. Accordingly, to enable the ranking of the patterns when they have identical coverage, the authors use the probability concept to improve the method suggested in [12]. Such hierarchy is the risk estimation. It is useful to guide the risk reduction process.

The risk estimation process suggested here consists of:

1) Ranking the patterns based on their coverage and their probability (see the hierarchy in Table C5 and Table C6);
2) For every pattern, ranking the merging indicators based on their frequency in the whole set of patterns generated.

For instance, in the "Fatal" class, the most important pattern is P_6^+. For that pattern, condition $E \leq 4.5$ is more important than condition $O \leq 2.5$ since E is more frequent than O in the set of patterns, as previously said.

The safety practitioner will therefore reduce the risk in respect of that logic: from the most to the least important pattern per class, starting from the most important class. When some patterns have the same probability, one must consider the ability of the pattern to generalize. The fewer or less strict conditions a pattern has, the more generalizable it is. Consequently, the more important it is. For instance, P_1^- and P_4^- have the same probability but a different number of conditions: 2 versus 4. In that example, P_1^- has the ability to generalize more than P_4^- because the former has fewer conditions than the latter.

The importance of a class of accident is obvious: "Fatal" is more severe than "Non-fatal". Moreover, "Fatal" patterns have a higher probability of occurrence of harm than "Non-fatal" ones. Nevertheless, the acceptability of the risk should be determined by the stakeholders concerned by the risk management process; namely, the worker, the employers and safety practitioner. The level of acceptability will suggest how deep the risk reduction measures should be. The authors cannot determine the acceptability of the risk, because the level of acceptability depends on the stakeholders' judgment, which is based on the reality of their enterprise and personal factors such as: values, culture, background [30], [31]. Depending on these reasons, the stakeholders will state to themselves the kind of risk they can accept living with. The consensus required from these stakeholders will be especially important when facing particular risk scenarios such as: high severity – high probability, high severity – low probability, low severity – high probability, low severity – low probability.

8.3 Interpretation of the probabilities values

Having a probability of occurrence of harm for non-fatal harm that is lower than for fatal harm can be explained by the fact that:

- The total non-fatal accidents were inferior to the amount of fatal accidents in the database;
- The probabilities obtained are conditional probabilities: the probabilities obtained consider the fact there was a belt-conveyor-related accident. In other words, the probabilities found are the probabilities of occurrence of harm knowing that fatal or non-fatal belt-conveyor-related accidents have happened in the industries concerned.

Since the accident investigation reports available at the CNESST's Documentation Center are only about serious and fatal injuries, many non-fatal accidents that are not serious are not

publicly reported. In reality, there are more non-fatal (serious and non-serious) injuries than fatal ones. If all of those regarding belt conveyors were available, the probabilities of occurrence of harm would have been different and maybe would reflect reality more in the sense that the probability of occurrence of fatal harm associated with a hazardous situation would have been lower than that of non-fatal ones.

Another aspect contributing to the misrepresentation of reality are the uncertainties related to the choice of the number of clusters deducted from the merging process in the data formatting. A higher or lower number of clusters may have been picked according to the needs of the user of the model and would impact the probabilities calculated.

Despite that misrepresentation, the probabilities of the occurrence of harm obtained represent a basis for comparison serving as a frame of reference to:

- Evaluate the effectiveness of risk reduction measures when mitigating risks.
- Monitor the impact of the occurrence of new accidents on the probability of occurrence of harm.

Indeed, such basis for comparison may be subject to change after a future accident, or after the integration of risk reduction measures.

As a result of insufficient data, no other accident was available to update the patterns (i.e. the update of the risk identification). However, having a 24th accident added to the database may have yielded to a new set of patterns, and, accordingly, different probabilities of occurrence of harm.

Applying a risk decrease index of a risk reduction measure similar to that of Aneziris et al [14] will decrease the probability of occurrence of harm by a certain percentage. For instance, in Aneziris et al [14], they show graphically that lock-out and tag-out decrease the risk of maintenance-related accidents by nearly 15%. For continual improvement, the new basis for comparison to be will serve as a new reference to evaluate further risk reduction measures.

Finally, the probability calculation method proposed to estimate the probability of occurrence of harm can also be applied to estimate the probability of occurrence of a hazardous event. In that context, the classes of the database would be types of hazardous events (e.g. unexpected release of energy, or unexpected start). The indicators would be variables describing such an event (e.g. a faulty energy dissipation process during lockout, or a failure in the control system). For qualitative risk estimation matrices comprising the parameter “Probability of occurrence of hazardous event”, knowing the quantitative value of that parameter would bring objectivity to the risk estimation by describing the qualitative levels assigned to that parameter in the matrices.

9. Conclusion

This paper proposes a systemic-based approach integrating Logical Analysis of Data (LAD) to estimate the probability of occurrence of harm. For this purpose, a database is used. It contains 23 fatal and non-fatal belt-conveyor-related occupational accidents from Quebec and 23 indicators (i.e. variables). The systemic-inspired accident model suggested in this paper uses the MELITO concept. That concept consists of describing the context in which the accident occurred, focusing on the moment (*M*) of the accident, the equipment (*E*) involved, the location (*L*) of the accident,

the individual (*I*) concerned, the task (*T*) undertaken at the time of the accident, and the organization (*O*) in the workplace. The 23 indicators are organized according to the letter of MELITO they are related to. These six letters are called “merging indicators” in this paper. The MELITO model presented in this study can be personalized if one wants to include further indicators to describe the merging indicators. Moreover, the method itself is transferable to any systemic-based concept. An advantage of the systemic approach over the sequential one is the fact that all kinds of aspects of the workplace are considered, whether or not they have been identified as a cause by the accident investigator. Proceeding in this way ensures some causes or risk factors that contributed to the accident are not left behind. Furthermore, systemic models have the ability to allow monitoring of the performance variability of a system.

The knowledge inferred by LAD in the form of patterns and the probabilities obtained are useful to achieve a prevention action plan. The patterns highlight the main risk factors (here, the merging indicators, and consequently their inherent indicators,) that may yield a fatal or a non-fatal belt-conveyor-related accident. The set of patterns generated showed that risk factors and potential causes of accidents came primarily from the equipment, followed by the organization, the individual, then the moment, in the accidents of the 23 enterprises analyzed. Since the “Fatal”class reflects the most severe harm and has patterns with higher probabilities of occurrence of harm than that of the non-fatal class, the class in which the risk reduction process must start first is the “Fatal” class. The probabilities of occurrence of fatal or non-fatal harm related to a hazardous situation concurrently with the severity level of a class of accidents enable the stakeholders to make a decision on the acceptability of the risk in order to decide how far to go in the risk reduction process. Once the acceptability of the risk is determined, the risk reduction can be undertaken in respect of the hierarchy of the patterns based on their coverage and probability, followed by the hierarchy of the merging indicators involved.

The application of the method proposed to estimate the probability of occurrence of harm requires skills in knowledge extraction process, as well as in probability calculation. It is suggested that the method be automated to promote its dissemination in all kinds of enterprises and allow its usage by any safety practitioner. Once automated, all that the safety practitioners will have to do is to enter in the database a new accident or incident that happened then click an “Update” button. Clicking it will:

- launch the knowledge inference algorithm to update the patterns in order to identify new risks;
- estimate the new probabilities of the patterns, therefore, the new probabilities of occurrence of harm;
- prepare a preliminary action plan showing the patterns and their indicators ordered according to their importance dictated by their coverage, probability and frequency.

The safety practitioner will complete the prevention action plan by assigning risk reduction measures to every indicator forming the patterns from the most to the least important. The choice of the risk reduction measure will be performed with the end-users (e.g. workers) to make sure that these measures do not hinder their activity.

The safety practitioner will be able to use the probabilities of the occurrence of harm calculated as a basis for comparison to consider the impact of a new belt-conveyor-related accident or integration of safety recommendations on these probabilities. Accordingly, these bases for

comparison are means to monitor the occupational belt-conveyor-related risks. Apart from updating the risk based on new accidents registered in a database, risk updating also concerns a new level of the risk resulting from changes to the probability of occurrence of harm or the severity of that harm after applying risk reduction measures. In the current state of the framework, only a risk update after a new accident is possible. The risk update after the implementation of a risk reduction measure is dealt with later, exploiting expert elicitation. Experts will be consulted to gather their judgment on the impact, in percentage, of some risk reduction measures on the mass distributions of the indicators forming the patterns.

As a result of the risk evaluation step, risk management will always be subjective. Indeed, risk evaluation consists of determining the acceptability of the risk based on the stakeholders' judgment. However, it is possible to bring objectivity to that process by making its previous steps, namely, risk estimation, more objective. In order to make risk estimation more objective, quantitative parameters such as pattern coverage, pattern probability, and merging indicator frequency are used in this paper to help safety practitioners make decisions on facts (accidents reported).

Of course, depending on the definition given to the word "risk", the risk assessment as well as its management may vary [1]. Similarly, the choice of accident model (i.e. the frame of reference) impacts how risk assessment is done [14].

Acknowledgment

The funding provided for this study by the IRSST is gratefully acknowledged.

References

[1] Villa V, Paltrinieri N, Khan F and Cozzani V. Towards dynamic risk analysis: A review of the risk assessment approach and its limitations in the chemical process industry. Saf Sci 2016; 89: 77–93.

[2] ISO (International Organization for Standardization). Safety of machinery - General principles for design - Risk assessment and risk reduction. ISO 12100. Geneva, Switzerland: International Organization for Standardization. 2010

[3] Chinniah Y. Analysis and prevention of serious and fatal accidents related to moving parts of machinery. Saf Sci 2015; 75:163–73.

[4] Poisson P, Chinniah Y. Managing risks linked to machinery in sawmills by controlling hazardous energies: Theory and practice in eight sawmills. Saf Sci 2016; 84:117–130.

[5] Bluff E. Safety in machinery design and construction: Performance for substantive safety outcomes. Saf Sci 2014; 66: 27–35.

[6] Jocelyn S, Chinniah Y, Ouali M-S. Contribution of dynamic experience feedback to the quantitative estimation of risks for preventing accidents: A proposed methodology for machinery safety. Saf Sci 2016; 88: 64–75.

[7] Buncefield Major Investigation Board. The Buncefield incident 11 December 2005. Bootle, United Kingdom. 2008.

[8] Hubbard D, Evans D. Problems with scoring methods and ordinal scales in risk assessment. IBM J Res & Dev 2010; 54(3): 2:1–2:10.

[9] Gauthier F, Chinniah Y, Burlet-Vienney D, Aucourt B, Larouche S. Sécurité des machines – Expérimentation pratique de paramètres et d'outils d'estimation du risque (Research report R-940). Montreal, Quebec, Institut de recherche Robert-Sauvé en santé et en sécurité du travail (IRSST), 2016.

[10] Chinniah Y, Gauthier F, Lambert S, Moulet F. Experimental analysis of tools used for estimating risk associated with industrial machines (Rapport n° R-684). Institut de recherche Robert-Sauvé en santé et en sécurité du travail, Montreal, Quebec, Canada, 2011.

[11] Apostolakis GE. How useful is quantitative risk assessment? Risk Anal 2004; 24(3): 515–520.

[12] Jocelyn S, Chinniah Y, Ouali M-S, Yacout S. Application of logical analysis of data to machinery-related accident prevention based on scarce data. Reliab Eng Syst Saf 2017; 159: 223–236.

[13] Raviv G, Shapira A, Fishbain, B. AHP-based analysis of the risk potential of safety incidents: Case study of cranes in the construction industry. Saf Sci 2017; 91: 298–309.

[14] Aneziris ON, Papazoglou IA, Konstandinidou M, Baksteen H, Mud M, Damen M, Bellamy LJ, Oh J. Quantification of occupational risk owing to contact with moving parts of machines. Saf Sci 2013; 51: 382–896.

[15] Papazoglou IA, Aneziris O, Bellamy L, Ale BJM, Oh JIH. Uncertainty assessment in the quantification of risk rates of occupational accidents. Risk Anal 2015; 35(8): 1–26.

[16] Bellamy LJ, Ale BJM, Geyer TAW, Goosens LHJ, Hale AR, Oh J, Mud M, Bloemhof A, Papazoglou IA, Whiston JY. Storybuilder—A tool for the analysis of accident reports. Reliab Eng Syst Saf. 2007; 92: 735–744.

[17] Dekker S. The field guide to understanding human error. Burlington: Ashgate Publishing; 2006.

[18] Leveson NG. Engineering a safer world: Systems thinking applied to safety. Cambridge: The MIT Press; 2011.

[19] Hollnagel E. Barriers and accident prevention – or how to improve safety by understanding the nature of accidents rather than finding their causes. Burlington: Ashgate Publishing; 2004.

[20] Centre patronal de santé et sécurité du travail du Québec (CPSST). Vite, on enquête ! Convergence 2004; 20(2): 6–7.

[21] Canadian Center for Occupational Health and Safety. Accident investigation, http://www.ccohs.ca/oshanswers/hsprograms/investig.html [accessed 17.01.15]

[22] United States Department of Labor. Incident [Accident] investigations: a guide for employers – A systems approach to help prevent injuries and illnesses. Occupational Safety and Health Administration (OSHA), 2015.

[23] Health and Safety Executive (HSE). Investigating accidents and incidents – A workbook for employers, unions, safety representatives and safety professionals (Workbook HSG245). Health and Safety Executive (HSE), 2004.

[24] Compiègne I, Curry X, Duval C, Andéol-Aaussage B. L'analyse de l'accident du travail – La méthode de l'arbre des causes (Guide ED 6163). Institut national de recherche et de sécurité (INRS), 2013.

[25] Commission des normes, de l'équité, de la santé et de la sécurité du travail (CNESST), www.csst.qc.ca

[26] ERIC. Tanagra, http://eric.univ-lyon2.fr/~ricco/tanagra/fr/tanagra.html [accessed 16.05.17].
[27] Bonates TO, Gomes VSD. LAD-WEKA Tutorial Version 1.0. 2014.
[28] Jourdain B. Probabilités et statistiques. Paris, France : École des Ponts. 2013.
[29] Parkes KR. Shift schedules on North Sea oil/gas installations: a systematic review of their impact on performance, safety and health. Saf Sci 2012; 50(7): 1636–1651.
[30] Hietikko M, Malm T, Alanen J. Risk estimation studies in the context of a machine control function. Reliab Eng Syst Saf 2011; 96: 767–774.
[31] Nanda Tchiehe D, Gauthier F. Classification of risk acceptability and risk tolerability factors in occupational health and safety. Saf Sci 2017; 92: 138–147.
[32] Giraud L, Massé S, Dubé J, Schreiber L, Turcot A. Sécurité des convoyeurs à courroie – Généralités, protection contre les phénomènes dangereux – Guide de l'utilisateur. 2nd Edition. Quebec : CSST (Commission de la santé et de la sécurité du travail). 2003.
[33] Childress S. How subcontracting affects worker safety. Frontline Enterprise Journalism Group. 2012.
[34] Walter J. Safety management at the frontier: cooperation with contractors in oil and gas companies. Saf Sci 2017; 91: 394–404.
[35] Massé S, Giraud L, Dubé J, Vernoux G, Schreiber L, Desrochers Y. Sécurité des convoyeurs à courroie – Principes de conception pour améliorer la sécurité – Guide du concepteur. 2nd Edition. Quebec : CSST (Commission de la santé et de la sécurité du travail). 2003.

Appendix

Table C7: Definition of classes involved

Class	Definition	Value
Non-fatal	This class comprises non-fatal but serious accidents. Some of the injuries happened after a crushing, an entrapment, a wrenching, an impact to/of a part of the body or the entire body. The others happened after a strangulation or entanglement with suffocation from engulfment.	0
Fatal	This class encompasses fatal accidents.	1

Table C8: Definition of the 23 indicators used in the initial database

Initial indicator	Corresponding merging indicator	Definition of the initial indicator	Values of the initial indicator ordered from the less to the most dangerous
V_1	M	Shift	Day shift (6:00 AM-5:59PM) = 1; Night shift (6:00PM-5:59AM) = 2
V_2	E	Machine	Fixed belt conveyor = 1; horizontally mobile belt conveyor = 2; vertically mobile belt conveyor = 3
V_3	E	Causal agent	Entrapment zone = 1; nip point = 2
V_4	E	Machine functional impairment	No = 0; yes = 1
V_5	E	Accident happened when machine started unexpectedly	No = 0; yes = 1
V_6	E	Accident happened when machine was functioning in automatic mode	No = 0; yes = 1
V_7	E	Safeguarding was in place at time of accident	Yes = 1; no = 2
V_8	E	Flawed or deficient warnings and markings involved in accident	No = 0; yes = 1
V_9	E	Personal protective equipment (PPE), work clothing or tool involved in accident	No = 0; yes = 1
V_{10}	E	Control system was involved in accident	No = 0; yes = 1
V_{11}	E	Accessible hazard zone where accident occurred	Zone between belt or chassis and another structure = 1; pulley = 2; roller = 3; tensioning drum = 4; tail drum = 5; drive drum = 6
V_{12}	E	Another piece of equipment (in addition to belt conveyor) was involved in accident	No = 0; yes = 1
V_{13}	E	A missing accessible emergency stop device was involved in accident	No = 0; yes = 1
V_{14}	L	Accident due to a poor working environment (e.g., cluttered work area)	No = 0; yes = 1
V_{15}	I	Time in position (experience of worker)	20 to 24 years = 1; 5 to 10 years = 2; 0 to 4 years = 3
V_{16}	I	Worker from the company or subcontractor	Company = 1; subcontractor = 2
V_{17}	I	Worker's regular activity (task)	Yes = 1; no = 2
V_{18}	T	Worker is specially trained to use this machine	Yes = 1; no = 2
V_{19}	T	Lockout procedure was applied	Yes = 1; no = 2
V_{20}	T	Activity	Production = 0; Maintenance = 1
V_{21}	O	A prevention program exists	Yes = 1; no = 2
V_{22}	O	Flawed occupational health and safety (OHS) management	No = 0; yes = 1
V_{23}	O	An OHS committee exists	Yes = 1; no = 2

Table C8 shows the ordinal numerical values attributed to the discrete qualitative values of the initial indicators. For some of them, the order of the values was based on the presence of some

hazards. For instance, the indicator V_2 referring to the machine can take values from 1 to 3. The value "1" is the least dangerous and refers to a fixed belt conveyor. The value "2" represents a horizontally mobile belt conveyor. The value "3" is the most dangerous of the three and corresponds to a vertically mobile belt conveyor. The vertically mobile belt conveyor is the most dangerous because in addition to the common hazards related to a fixed belt conveyor, it moves vertically. Its vertical motion makes it more dangerous than a horizontally mobile belt conveyor because no matter where it is installed, there is a risk of being crushed between the belt and the surface it is installed on. On the other hand, when a horizontally mobile belt conveyor moves from side to side, there is a risk of striking someone without crushing if it is installed far enough from an obstacle.
Regarding the causal agent (V_3), as the majority of belt-conveyor-related accidents occur on drums and rollers [32], nip points related to these mobile parts are more involved in these accidents than entrapment zones.

Concerning the accessible hazard zone where an accident occurred (V_{11}), the ordering of its 6 values was based on the fact that:

- According to reference [32], among 85 belt-conveyor-related accidents analyzed, 48% happened on the drive drums, 13% on the rollers.
- Furthermore, in the database, accidents happened the most according to the order presented in Table C8.

Regarding indicator V_{16}, having a worker from the company is considered to be less dangerous than having a subcontractor. According to reference [33], subcontracting affects a worker's safety. "The more subcontracting, the more 'fissuring' of the work, the greater the risk for health and safety" [33]. Indeed, as outsourcing was made a priority to reduce costs, riskier activities were transferred to subcontractors instead of transferring risk management technologies [34]. That can make subcontractors less familiar with the workplace where they intervene than workers from the company.

For the activity (V_{20}), maintenance is considered as being more dangerous than production because a belt conveyor design guide [35] declares that the majority of accidents occurring on belt conveyors happen during maintenance activities.

Concerning the shift (V_1), two values are allocated: day shift and night shift. Night shifts are considered more dangerous than day shifts even though the safety problem relies also in the switch from one shift to another. For instance, the switch from day to night shifts requires 5-6 days to adapt physiologically and psychologically [29]. Readaptation to day shifts takes more time [29]. Parkes [29] affirms that field study findings generally reveal that sleep, alertness and performance are relatively stable across day-shift. Initial night shifts are adversely affected by circadian disruption, she says [29]. That affirmation explains why the "night shift" value is considered more dangerous than the "day shift" value of V_1.

For the other indicators, the ordering was obvious and sensible. For example, it is obvious that having a flawed occupational health and safety (OHS) management (V_{22}) is more dangerous than having a non-flawed OHS management.

ANNEXE D – RÉSULTATS DU LANCEMENT DE LA FOUILLE DE DONNÉES AVEC LE LOGICIEL *TANAGRA* ET LA TECHNIQUE « RÈGLES D'ASSOCIATION »

« M » désigne la classe « Accident du travail sur un convoyeur à courroie lors d'activités de maintenance ».

« P » désigne la classe « Accident du travail sur un convoyeur à courroie lors d'activités de production ».

Dans cette annexe, on retrouve les paramètres : « Support », « Confiance » et « Conviction » qui définissent la qualité des règles d'association trouvées. Notez que, dans *Tanagra*, « *Lift* » désigne la conviction.

Accident en maintenance

Spv Assoc Rule 1

Parameters

A-Priori parameters	
Support min	0,8
Confidence min	0,7
Max rule length	4
Lift filtering	1,1
Learning set ratio	1
Repetition	1
Value to predict	M
VT Cut Value	2

Algorithme utilisé :
Supervised Association Rule

Valeurs minimales choisies pour le paramétrage

Results

Sample characteristics

Samples size	
Training	23
Test	0

ITEMS

Transactions	23
Counting items	
All items	62
Filtered items	11
Counting itemsets	
card(itemset) = 2	2
card(itemset) = 3	1
Rules	
Number of rules	**0**

RULES

Filtered = 0 rules

Aucune règle d'association caractérisant la classe « M »

Rules evaluation

N	Antécédent	Conséquent	n	n[A]	n[C]	n[A^C]	Support	Confiance	Lift

All rules

Rules evaluation

N	Antécédent	Conséquent	n	n[A]	n[C]	n[A^C]	Support	Confiance	Lift

Accident en maintenance

Spv Assoc Rule 1

Algorithme utilisé :
Supervised Association Rule

Parameters

A-Priori parameters	
Support min	0,04
Confidence min	0,6
Max rule length	4
Lift filtering	1,1
Learning set ratio	1
Repetition	1
Value to predict	M
VT Cut Value	2

Valeurs minimales réduites pour le paramétrage

Results

Sample characteristics

Samples size	
Training	23
Test	0

ITEMS

Transactions	23
Counting items	
All items	62
Filtered items	62
Counting itemsets	
card(itemset) = 2	60
card(itemset) = 3	1276
card(itemset) = 4	14054
Rules	
Number of rules	0

RULES

Aucune règle d'association caractérisant la classe « M »

Filtered = 0 rules

Rules evaluation

N	Antécédent	Conséquent	n	n[A]	n[C]	n[A^C]	Support	Confiance	Lift

All rules

Rules evaluation

N	Antécédent	Conséquent	n	n[A]	n[C]	n[A^C]	Support	Confiance	Lift

Accident en production

Spv Assoc Rule 1

Algorithme utilisé :
Supervised Association Rule

Parameters

A-Priori parameters	
Support min	0,1
Confidence min	0,7
Max rule length	4
Lift filtering	1,1
Learning set ratio	1
Repetition	1
Value to predict	P
VT Cut Value	2

Valeurs minimales choisies pour le paramétrage

Results

Sample characteristics

Samples size	
Training	23
Test	0

ITEMS

Transactions	23
Counting items	
All items	62
Filtered items	51
Counting itemsets	
card(itemset) = 2	21
card(itemset) = 3	184
card(itemset) = 4	939
Rules	
Number of rules	**3**

3 règles d'association caractérisant la classe « P »

RULES

Filtered = 3 rules

Rules evaluation

N	Antécédent	Conséquent	n	n[A]	n[C]	n[A^C]	Support	Confiance	Lift
1	"I5=Oui" - "I13=Non" - "I22=Deux"	"Class=P"	23	2	3	2	0,08696	1	7,66667
2	"I2=Un" - "I5=Oui" - "I13=Non"	"Class=P"	23	2	3	2	0,08696	1	7,66667
3	"I1=Non" - "I13=Non" - "I22=Deux"	"Class=P"	23	2	3	2	0,08696	1	7,66667

All rules

Rules evaluation

N	Antécédent	Conséquent	n	n[A]	n[C]	n[A^C]	Support	Confiance	Lift
1	"I1=Non" - "I13=Non" - "I22=Deux"	"Class=P"	23	2	3	2	0,08696	1	7,66667
2	"I2=Un" - "I5=Oui" - "I13=Non"	"Class=P"	23	2	3	2	0,08696	1	7,66667
3	"I5=Oui" - "I13=Non" - "I22=Deux"	"Class=P"	23	2	3	2	0,08696	1	7,66667

ANNEXE E – RÉSULTATS DU LANCEMENT DE LA FOUILLE DE DONNÉES AVEC LE LOGICIEL *TANAGRA* ET LA TECHNIQUE « ARBRE DE DÉCISIONS »

« M » désigne la classe « Accident du travail sur un convoyeur à courroie lors d'activités de maintenance ».

« P » désigne la classe « Accident du travail sur un convoyeur à courroie lors d'activités de production ».

Supervised Learning 1 (C4.5)

Algorithme utilisé : *C4.5*

Parameters

Decision tree (C4.5) parameters	
Min size of leaves	2
Confidence-level for pessimistic	0,1

Results

Classifier performances

Error rate				0,1304		
Values prediction				Confusion matrix		
Value	Recall	1-Precision		M	P	Sum
M	1	0,1304	M	20	0	20
P	0	1	P	3	0	3
			Sum	23	0	23

Classifier characteristics

Data description

Target attribute	Class (2 values)
# descriptors	23

Tree description

Number of nodes	1
Number of leaves	1

Aucun arbre permettant de classer « M » ni « P »

Decision tree

then Class = **M** (86,96 % of 23 examples)

Supervised Learning 1 (ID3)

Parameters

Algorithme utilisé : *ID3*

ID3 parameters	
Size before split	200
Size after split	50
Max depth of leaves	10
Goodness of split thresold	0,03

Results

Classifier performances

Error rate			0,1304		
Values prediction			Confusion matrix		
Value	Recall	1-Precision		M	P	Sum
M	1	0,1304	M	20	0	20
P	0	1	P	3	0	3
			Sum	23	0	23

Classifier characteristics

Data description

Target attribute	Class (2 values)
# descriptors	23

Tree description

Number of nodes	1
Number of leaves	1

Aucun arbre permettant de classer « M » ni « P »

Decision tree

then Class = **M** (86,96 % of 23 examples)