# An Ontology of Scientific Uncertainty: Methodological Lessons from Analyzing Expressions of Uncertainty in Food Risk Assessment

Sandrine Blanchemanche*, Akos Rona-Tas**,
Antoine Cornuéjols***, Antonin Duroy***, Christine Martin***

*Met@risk, INRA, Paris (France)
sandrine.blanchemanche@paris.inra.fr
http://www.une-page.html
** University of California, San Diego (USA)
aronatas@ucsd.edu
http://www.une-autre-page.html
*** AgroParisTech-INRA UMR-518 MIA
16, rue Claude Bernard – F_75231 Paris Cedex 05 (France)
antoine.cornuejols,christine.martin@agroparistech.fr

## Abstract

The relationship between scientific knowledge and uncertainty in science has been a central question in risk analysis. There have been several conceptualizations of uncertainty but most have been normative efforts to construct an ontology on the basis of theoretical considerations and there have been few empirical attempts to build such an ontology through textual analysis. Studies investigating how such ontologies work are equally scarce.

We developed an ontology to investigate uncertainty in risk assessment in food safety comparing the EU and the US, and the two main domains of food safety: biohazards and contaminants. The ontology gauges expressions of uncertainty in two ways: one looks for stylistic clues of judgment, the other registers the content of the uncertainty expressed in the documents. We have built a large data base where English language risk assessment documents by the European Food Safety Agency and the three US agencies primarily responsible for food safety in the US are coded according to our ontology. We are also in the process of creating software that uses machine-learning algorithm to code risk assessment documents in our database.

In our paper, we lay out our approach to scientific uncertainty, then describe the ontologies we developed to assess expressions of scientific uncertainty in risk assessment documents in food safety. We then discuss the results from supervised Machine Learning and the implications for both the method of machine coding and the insights we gained for human coding. Finally, we discuss some findings using the ontology testing three sets of hypotheses. The first one looks at differences in national styles in applied scientific research, the second differences in epistemic cultures between scientific subfields in food safety and the third about the effect of research and new knowledge on uncertainty.

## 1. Introduction: Mapping Scientific Uncertainty

Since the landmark report by the National Research Council (NRC) that codified the basic paradigm of risk analysis (NRC 1983), science has become an integral part of policy making. By separating risk assessment, a predominantly scientific review of the existing relevant knowledge on an issue of interest that is to embody the "scientific findings and judgments" of expert scientists, and risk management, that is to reflect "political, economic and technological considerations," the paradigm sought to establish the autonomy of science in the process. This separation was designed to allow science to make an independent contribution following its own methods and logic, yet it also created a disjunction. Science and policy making work in different epistemological cultures. Scientific knowledge is always provisional, open to skepticism and further inquiry, while policy requires knowledge that is final, certain and permits making decisions that are timely and often irreversible, and this difference is often not easy to bridge.

This epistemological mismatch became painfully obvious in recent years, and traumatic events like the mad cow disease or the L'Aquila earthquake, highlighted the need for scientists to communicate clearly not just what they believe to be the most likely truth on a topic, but to also express the limits of their knowledge. Since the mid-1990s, pressure began to mount on the scientific community to report not just what they know but also to clarify what they think they do not know, calling attention to uncertainties in the current corpus of the pertinent scientific knowledge.

In the area of food safety, our focus of interest, various international agencies began to emphasize the importance of expressing scientific uncertainty in risk assessment documents, analytic surveys of the scientific literature on a particular food hazard. Some, like WHO, EFSA, US EPA, US OMB, issued guidelines working towards a system that both identifies the type of uncertainty scientist perceive and the extent to which our knowledge is uncertain in a particular respect. These agencies recognized that their decision makers must have a clear sense of how much confidence they can place in various scientific findings in order to take the best decision. They also have to understand the nature of the weakness in the evidence experts present. Scholars studying science itself also took up the issue of scientific uncertainty and formulated various normative frameworks to guide experts in future risk assessment reports.

Our approach to scientific uncertainty is not normative, but empirical and comparative. We want to describe and understand how scientists express uncertainty in scientific reports assessing food risk. In our larger project, we look at English language risk assessment documents produced for food safety regulators in the United States and the European Union between 2000 and 2010. We investigate two main, distinct areas of food hazards in the food risk field: contaminants and biohazards. [1] As the two fields draw on different subdisciplines, they differ in the way they make use of various scientific methodologies (experiments, observational studies, statistical analyses, analytic modeling etc.) and thus may have different understandings of scientific uncertainties. Our ultimate interest is the relationship between uncertainty and accumulation of knowledge, as well as the effect of scientific uncertainty on policy outcomes.

To answer this we developed inductively two complementary ontologies: the first one is built with a linguistic approach, and the second one reflects the content or source of uncertainty. Both are implemented in a database containing coded risk assessment and management documents. Because the quality and the accuracy of the ontology are crucial, we developed a learning algorithm in order to evaluate the reproducibility of the coding. As risk assessments are numerous and documents can be quite long (up to 500 pages), this software may help coders in the future. The software can also improve our ontology, pointing out inconsistencies and imprecisions. In this paper, we present our first results from our analysis of our ontology with Machine Learning algorithms.

In the rest of the paper we lay out our approach to scientific uncertainty, then describe the ontologies we developed to assess expressions of scientific uncertainty in risk assessment documents in food safety. We then discuss the results from supervised Machine Learning and the implications for both the method of machine coding and the insights we gained for human coding. Finally, we discuss some findings using the ontology testing three sets of hypotheses. The first one looks at differences in national styles in applied scientific research, the second differences in epistemic cultures between scientific subfields in food safety and the third about the effect of research and new knowledge on uncertainty.

---

[1] Contaminants are any substance, such as arsenic, cadmium or lead, not intentionally added to food which is present as a result of the production, manufacture, or other steps while holding food or as a result of environmental contamination. Biological hazards include pathogenic viruses and bacteria.

## 2. Uncertainty in Science

Uncertainty is an indispensable part of science. Positivist approaches depict science a progressive conquest of uncertainty, whereby science with each step reduces the realm of unknowns. These approaches see knowledge and uncertainty as a zero sum game. As research moves successfully forward, knowledge expands and uncertainty contracts. From a positivist perspective, measurement of the extent of uncertainty is possible. Since uncertainty is the objectively existing entities still not known, their quantification is a difficult but meaningful exercise.

By contrast, the cognitivist conception of science posits uncertainty not as the opposite but as a product of knowledge (Popper 1976). Uncertainty, after all, assumes that we know we don't know something, thus uncertainty presupposes some knowledge. Consequently, increase in knowledge not only does not make uncertainties shrink, to the contrary, it adds to it. As uncertainty is always framed by existing knowledge, with new knowledge not only does it grow but also its nature shifts. What counts as uncertainty, therefore, depends on the context of what we believe is certain which changes as we learn more. As we cannot predict knowledge we not yet have, we also cannot foresee what uncertainties we face in the future. Measuring uncertainty can accomplish much less, at best it can account for our ignorance at a particular point in time as shaped by a very carefully specified scientific puzzle. Because uncertainty reflects existing knowledge, different fields with different methodologies will have different standards for uncertainty as well.

Finally, social constructivists believe that what is considered knowledge and uncertainty both built in social processes, where scientists are interested parties in a wider social context (Conway and Oreskes, 2010, Shackley and Wynne, 1996, Proctor and Schiebinger 2008). Uncertainty, they claim, is manufactured (or at least emphasized or de-emphasized) on purpose. They point to such debates as the health effects of tobacco, evolution or the human causes of climate change where uncertainty has been deliberately created. In those cases, uncertainty has been strategically produced by interested parties. As the institutional environment plays a key role, uncertainty will be perceived and identified differently in the US and EU.

Our ontology is compatible with all three theories, which means that we can later compare their empirical predictions. The typologies we offer assume, as all three theories do, that science is a discourse with its own rules of expression, and even climate change deniers or creationists must comply with them if they want to participate in this type of a discussion. Because our job is not to point to or evaluate claims of uncertainty, but only to describe those claims, our ontology need not take a prior position on these theories, but it can test their implications.

## 3. Ontology of Uncertainty in Food Risk Analysis

Ontology is "an explicit specification of a conceptualization" (Gruber 1993:199). In full-fledged ontologies, concepts and their relationships are organized in a system that is an abstract representation of a world with a certain purpose and at a specific level of granularity. Ontologies are powerful, because they can clarify and – to some degree – automate various cognitive processes that manipulate meaning. We set out to develop a conceptualization of uncertainty in scientific documents, to identify textual expressions of uncertainty and then to sort and analyze documents according to the amount and type of uncertainty they voice.

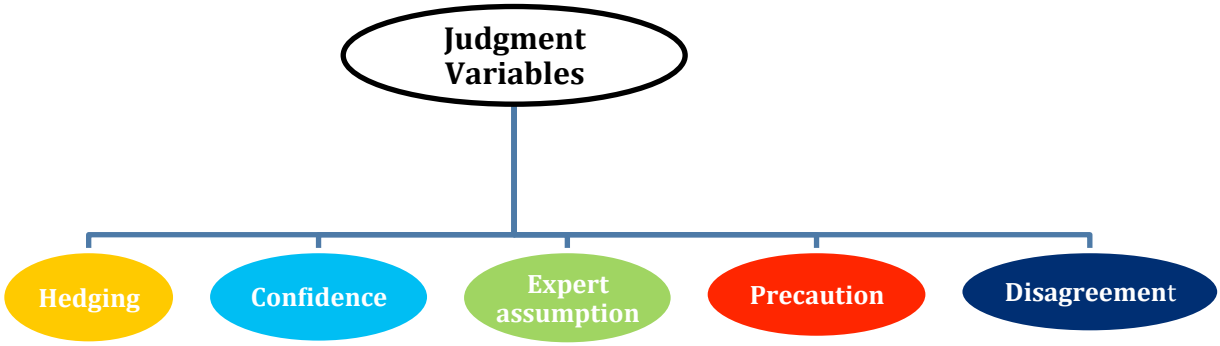
We developed two systems of classification or simple ontologies. The first, a simpler ontology is designed to capture the nature of the judgment the scientists make about the uncertainty of their conclusion. This ontology is a typology, a multidimensional way of classifying the expert's judgment of the evidence. The second, a hierarchical system gauges the content of uncertainty. The categories identify

the problems that give rise to uncertainty about our current state of knowledge as perceived by the authors. It is taxonomy, because the categories are arranged in a genealogical hierarchy, where "ancestry" can be seen as successive levels of generality. For both ontologies, our smallest coded unit is the sentence (our data point). Categories can be attached to one or more consecutive sentences. One sentence can contain multiple expressions of uncertainty and can be sorted into multiple categories.[2] Because we were interested in the final verdict of the experts, just as policy makers are, we coded only summaries and conclusions of each document to capture the uncertainties that the experts thought remained after their reviewed the available research on the topic.[3]

### *A. Judgment typology : A linguistic approach based ontology*

Our first ontology was designed to capture various aspects of the judgment of the experts in their conclusions. It describes how the panel judges the weight of the evidence and it follows more closely the language they use to do so. This ontology consists of five categories. They are conceptually distinct. Three of them express uncertainty, two (confidence and expert assumption) communicate the opposite.

**Figure 1. The structure of the ontology of uncertainty based on judgment**




*Hedging* is a way of indicating that experts have doubt about or a lack of total commitment to a proposition they present. There is a large literature on hedging. Hedging, a way of making things fuzzier (Lakoff 1972),[4] expresses a "lack of complete commitment to the truth value of an accompanying proposition" (Hyland 1998:1). It suggests that the speaker is not committed entirely to a proposition because he or she is uncertain about the truth of its content. The hedge signals this uncertainty without laying out its causes in detail there in the sentence, albeit the causes may be explained elsewhere in the text.[5] Hedging ill serves risk managers because it makes the topic of interest less clear. To identify

---

2 When a sentence contains multiple expressions of uncertainty, each is represented by different phrases or clauses. Therefore, in principle, we could break up those sentences and pick out the words relevant to each.

3 If the experts were left with uncertainty after one study reviewed in their report, but were then satisfied that another study presented resolved the issue, uncertainty would be reported for the first one, yet that uncertainty would not mater anymore.

4 Lakoff's original article that set off research on hedges makes the claim that making propositions fuzzier is actually making them more accurate, because the world is fuzzy and truth is a matter of degree. Hedges allow us to move beyond the stark and misconceived binary distinction between truth and untruth.

5 The literature attempts to classify hedges depending on how it deals with uncertainty, whether it serves to protect the author, or whether it just indicates that information is incomplete or that the validity or reliability of the proposition is not fully accepted. We did not make these distinctions.

hedging we ask the question: "Can the proposition be restated in such a way that it is not changed but that the author's commitment to it is greater than at present? If yes, then the proposition is hedged." (Crompton, 1997: 281). For instance, dropping "likely to be" in the sentence: "The … panel concluded that … the risk is likely to be conservative…" would make it more definite.

Our second category is *confidence*. Here we wanted to capture the opposite of uncertainty, an emphatic commitment to a proposition. Often referred to as boosters, expressions of certainty, assurance and conviction provide a crucial clue for risk managers (Myers 1989, Vazquez and Giner 2009). Boosters play an important role in persuasion in risk assessments. They stress finality and absence of doubt. While there are many words that are commonly used as boosters (e.g., undoubtedly, clearly, well-known, demonstrate, proven) whether they express confidence in the relevant scientific knowledge can be judged only from the wider context. Experts, for instance, can be confident that no good data are available on a topic or report that it was demonstrated that the statistical models cannot answer the crucial question. In such cases, there is uncertainty and confidence is to emphasize that it is there.

Our third category is *expert assumption*. This is another form of confidence. The expert is aware that studies or models make certain assumptions about the world. These assumptions are not directly supported by evidence, but according to the expert, this is not a problem. These are the best assumptions an expert can make or, at least, these are not assumptions that the report questions.

We coded *precaution* as our fourth variable. Precaution is a way of dealing with uncertainty. Making conservative assumptions or building conclusions around "worst case scenarios" is a way of creating certainty where data and models fail to provide it. There is a large literature on the precautionary principle in food safety and the presumed differences in precaution between the EU and the US that developed mostly in the context of genetically modified organisms (Lynch and Vogel 2001, Hammitt et al 2005).

Our final category is *disagreement*. Disagreement is a staple of science, but here were interested in only disagreements that the report treats as unresolved. This happens either when experts on the panel find unanimously that contradicting evidence on the topic is equally strong, or when the panel splits, and some members disagree with others and voice their dissent

### *B. Uncertainty taxonomy: An ontology based on the source of uncertainty*

To build our second ontology focusing on content, we began with the general literature on scientific uncertainty (Morgan and Henrion 1990, Hattis and Burmaster 1994, Pate-Cornell 1996, van Asslet and Rotmans 2002, van der Sluijs et al. 2005, Walker et al. 2003) and papers addressing uncertainties in the different disciplines involved in the food risk assessment process, such as epidemiology, microbiology, toxicology and exposure assessment, (Grandjean & Budz Jorgensen 2007, Kang & Kodell et al. 2000, Nautta 2000, Dorne & Renwick 2005 and Kroes & Muller et al. 2002). Beside this literature, we drew upon two main institutional documents: the opinion of the Scientific Committee of EFSA entitled Uncertainties in Dietary Exposure Assessment (EFSA, 2006) and the WHO Draft guidance document on Characterizing and Communicating Uncertainty in Exposure Assessment (WHO, 2007). We simplified and adapted the basic structure of these classification systems through a series of test coding of European, US and international food safety risk assessments arriving at a 28 item hierarchical ontology defined by a decision tree. As one moves down the tree one gets to more specific content. The coder had to code at the most specific (lowest) level possible.

**Table 1. Decision Tree for Uncertainty Taxonomy Coding**

<table>
<tr><td rowspan="15">Is it uncertainty that is irreducible?<br>OR<br>Is it that new information can resolve</td><td colspan="5">Ontic Uncertainty/Variability</td></tr>
<tr><td rowspan="14">Epistemic Uncertainty<br><br>Is it due to the absence of good data about the hazard?<br>OR<br>Is it due to the way the model is built?</td><td rowspan="3">Model<br>Is it due to arbitrary model assumptions?<br>OR<br>Is it due to some problem in our causal understanding i.e. what generates the hazard?</td><td colspan="3">Arbitrary assumptions of model</td></tr>
<tr><td colspan="2" rowspan="2">Causal inference<br>Is uncertainty due to ignoring synergism (combination effects)?<br>OR<br>Is it due to the inability to separate the effects of related causes?</td><td>Combination effects</td></tr>
<tr><td>Correlated causal factors</td></tr>
<tr><td rowspan="11">Data<br>Is it due to the complete absence of data<br>OR<br>Is it due to the lack of the exact kind of data we need and the fact that we have to use proxies (surrogates)?<br>OR<br>If we have the right kind of data, does some quality of the data create uncertainty ?</td><td colspan="3">Missing factor</td></tr>
<tr><td rowspan="3">Measurement<br>Data from different sources are incomparable and they point in different directions.<br>OR<br>We don't know enough about how it was measured to trust the data.<br>OR<br>Is it due to how it is measured?</td><td colspan="2">Comparability of data</td></tr>
<tr><td colspan="2">Reporting</td></tr>
<tr><td rowspan="2">Measure<br>Was the measurement poorly done?<br>OR<br>Does the methodology used in measuring have inevitable limitations?</td><td>Flawed measure</td></tr>
<tr><td>Limited analytic method</td></tr>
<tr><td rowspan="7">Surrogate Data<br>Is it a sampling problem?<br>OR<br>Is there a discrepancy between the hazard we have data for and the hazard of interest?<br>OR<br>Is there a discrepancy due to context?<br>OR<br>Is there a discrepancy between the population of interest and the population we have information on?</td><td rowspan="2">Sampling<br>Was the sample too small?<br>OR<br>Was it selected improperly?</td><td>Small sample size /few samples</td></tr>
<tr><td>Non-representative sample</td></tr>
<tr><td colspan="2">Surrogate hazard</td></tr>
<tr><td rowspan="3">Surrogate context<br>Are the data from the wrong context?</td><td>Inference in time</td></tr>
<tr><td>Scenario inference</td></tr>
<tr><td>Range inter- or extrapolation</td></tr>
<tr><td rowspan="2">Surrogate Population<br>Are the data from the wrong population</td><td>Inference from in vitro to in vivo</td></tr>
<tr><td>Inference from animal to human</td></tr>
<tr><td>Inference from general to sensitive population</td></tr>
</table>

The tree was a decision tool to aid our coders. Sentences coded at branches, rather than leafs or terminal nodes (at the right column in Table 1 or the light blue label in Figure 2) were either unspecified at a lower level or were specified but the specification was so rare that it did not deserve a separate category at the next level.

Coding sentences for content that can be quite complex raises the problem of context much more so than the categories of our first ontology. The meaning of sentences is often influenced by text that is not adjacent. Comprehending the source of a particular uncertainty often required following a long exposition in the body of the report that the coder read but did not code. In fact, while we annotated sentences, here it would be more accurate to say that we were classifying the entire document and flagged the sentences that provided the best clue.

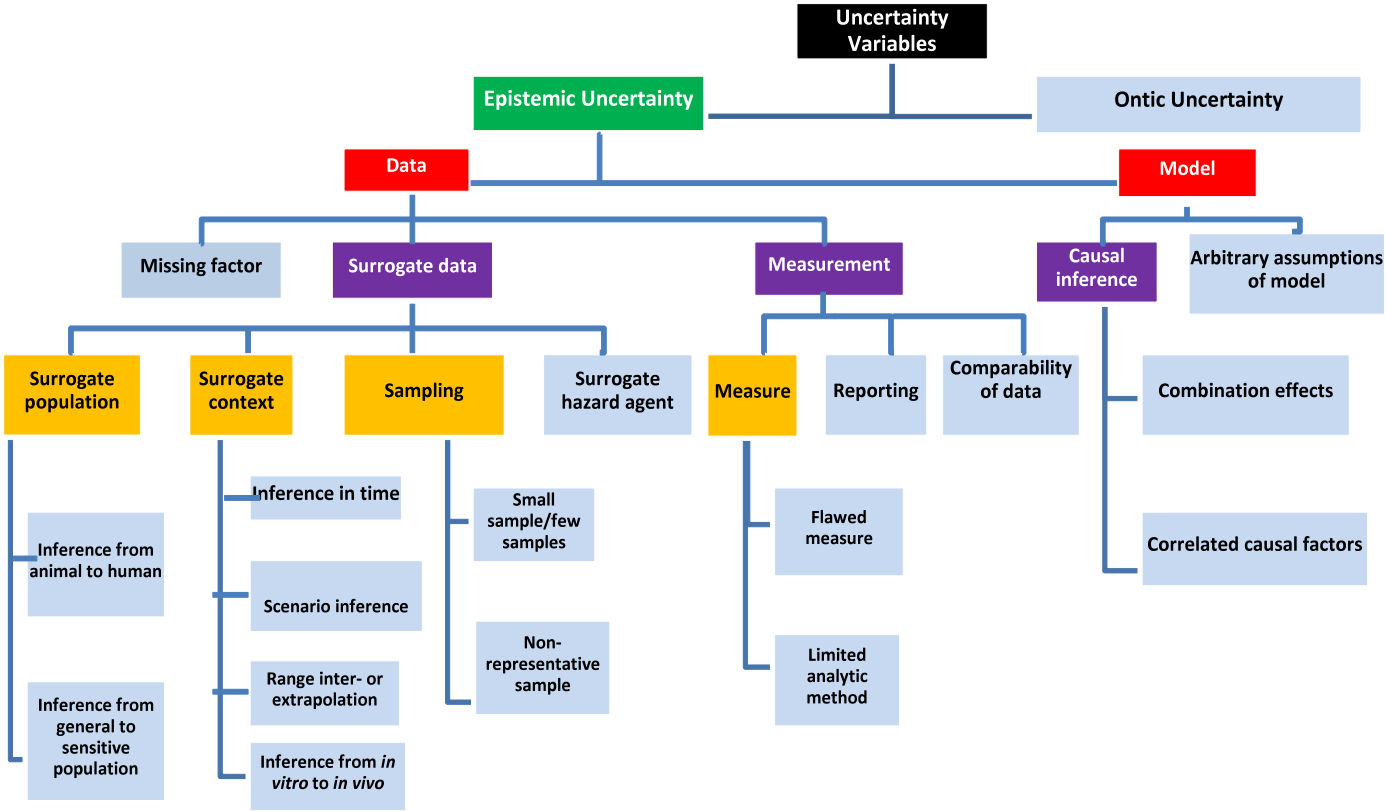
Our ontology begins with the common distinction between Epistemic and Ontic Uncertainty (also known as Natural Variability). Epistemic uncertainty is the kind that points to missing or incomplete

information. Ontic uncertainty is the inherent, random variation among cases that no further research can reduce. Epistemic uncertainty then is divided into problems that relate to Data and those that relate to the Model that we use to understand data. Each, in turn, except for one, is subdivided into lower level, more specific categories.

Data problem can be that the data (factors/variables) are simply missing. This is, however, rarely where the report stops. In the absence of good data, it reaches for surrogate data that are not exactly what we want and require some inference, or data we want but the measurement is somehow imperfect. Surrogate data can be imperfect because we have the wrong population, the wrong context, the wrong hazard or an imperfect sample. Measurement can be faulty because it was measured incorrectly, it was not properly reported or reported in a way that creates problems of comparison.

For models, we distinguish between causal and other, formalized models. Causal inference problems are further specified.

**Figure 2. The structure of the ontology of uncertainty based on content**




This ontology is, thus hierarchical by specificity. But another way of thinking of this tree structure is that the categories are organized in groups of similar content, whereby "children" of the same "parent" show more family resemblance than "children" of different "parents" or "grandparents."

## 4. Analyzing Our Empirical Ontology with Machine Learning

To examine the two ontologies, developed from data inductively, we tested them with Machine Learning. Two questions, in particular, arise prominently. First, do the nodes in a given ontology, labeled by some concept name (e.g. "ontic uncertainty"), truly correspond to some "natural category", or

are they in fact artificial distinctions imposed by the ontology maker with no or weak ground? Second, are the links between the nodes, making the structure of the ontology, associated with the right semantic relationships? For instance, do the child nodes of a parent node correspond to sub-categories of the category associated with the parent node? And are the child nodes associated with mutually exclusive categories?

Machine Learning is a systematic method to assess the validity of the proposed ontology with measurable results. The idea is to measure the performance of a learning algorithm when the data are organized along the lines of the evaluated ontology. If the learning algorithm is highly successful in predicting the node to which each sentence belongs, then the ontology is, at least, internally consistent. Poor fit can reveal problems. It can be the sign that the distinctions imposed by the ontology on the data are contrived and divide the domain poorly. It can point to categories that are not sufficiently distinct and have a semantic overlap, or that are improperly located in the taxonomy.

Because we were interested in the relationships among the categories of the ontologies, we did not include in our analysis sentences that did not contain expressions of uncertainty.[6]

### *A. A methodology for using Machine Learning to assess the validity of an ontology*

The general approach that we chose to assess the value of a given ontology was to measure how well decision rules could be learned by a learning algorithm in order to separate the data points, here sentences, between the categories imposed by the ontology. The protocol is the one of supervised learning. A set of sentences (796 in all) representing 112 risk assessment reports requested by official food safety agencies in the US and the EU, have been coded by experts, using the categories of the ontologies, thus providing a learning sample with pairs (*part of text*, *associated label*) a.k.a. examples. A subset of this learning sample was extracted to form a *training set* that would be used by the learning algorithm to produce decision rules, while the remaining learning examples formed the *test set*. The value of the learned decision rules was measured by how well they predicted the labels of the test examples. Usually the performance criterion was the error rate.

Several choices had to be made in order to realize the experiments:

1- Choice of the learning algorithm and of the kind of learning rule that can be learned. We focused on classical and state of the art supervised learning techniques, namely: naïve Bayes classifier, SVM (Support Vector Machines), Decision Trees and k-Nearest-Neighbors (see [Flach,2012]).

2- Choice of the coding of the sentences. Most learning techniques cannot take raw sentences as input. These must first be coded as vector of attribute-value pairs. Several techniques exist that aim at reducing the dimension of the input space which, otherwise, could easily attain $10^4$ up to $10^5$ features without pre-processing. These techniques include the removal of *stop-words*, i.e., words such as 'the', 'a', 'are', 'is', 'to', 'by' etc., that are supposed not to carry information about the label of the sentence. *Stemming* is a way to further reduce the vocabulary by retaining only the stems of the words (e.g. 'artificially' -> 'artificial'). One can also use information about synonymy, then retaining one word for each category of words. In our experiments, we removed 146 stop-words and used the Porter stemmer (see [Perkins,2010], [Porter,1980]), yielding a final vocabulary of size 2,530. This is still a very large space, but the results shown below suggest that this was already a good choice. At this point, one can adopt one out of several existing coding strategies. For instance, in the "bag of words" approach, each word present in a sentence will be associated with a 1 in the input vector and the other, absent, words with a 0. Or one can count the number of times this word is present. E.g. the sentence "*In addition to the limitations already listed, there are also limitations introduced by the methods used to analyze data inputs to the risk assess-*

[6] This is our next step. We are obviously interested in finding expressions of uncertainty in the text, not just finding out what kind of uncertainty it is once we identified an uncertainty statement.

*ment*." would become "*addit limit alreadi list limit introduc method analyz data input risk assess*" after stop-words have been removed and after stemming. Then, it can be coded either with a 1 for the feature 'limit' or with a 2 since this stem appears twice in this sentence. Other coding techniques involve the computation of relative frequencies (e.g. tfidf for "Term Frequency/Inverse Document Frequency"). Here, we report experiments with the 0/1 coding method.

3- Choice of a prediction strategy. Some algorithmic learners are, at the core, designed to separate two classes. If one wants then to learn multi-class classification, for instance, separating the three class '*missing variables*', '*surrogate context*' and '*sampling*', some scheme must be devised to turn two-class classification rules into multi-class ones. In our experiments, we used the all-versus-all technique (see [Aly,2005]). The best overall learners were Naïve Bayes and SVM. Below, we report the results obtained with the naïve Bayes classifier since they are easy to use and intrinsically adapted to multi-class classification.

4- One problem when learning to separate data from different classes is that their frequency can be significantly dissimilar. If, for instance, one class is ten times more highly represented than another one, a simple majority rule will yield a 90% successful prediction rate without any learning taking place. In order to circumvent this opportunistic but uninformative behavior, one method is to balance class sizes. When data are plentiful, it is sufficient to sample the over-represented classes in order to reduce their size to that of the least represented one. In our experiments, however, data are already rather scarce and another technique must be used (e.g. 1 sentence falls under the '*Measure*' node, while 71 fall under the '*Missing factors / variable* node). For each under-represented class, we chose to generate artificial data points (sentences) by mixing the characteristics of existing data points from this class. Specifically, we randomly drew two actual data points and made up an artificial data point by retaining for each feature either the one encountered in both actuals if they agreed, or by randomly choosing a value of one actual if they disagreed on this feature.

5- In order to measure the prediction performance, we used a five-fold cross-validation technique (see [Japkowicz et al.,2011]).

It is important to recall our premises *viz.* that a good predictive performance suggests that the ontology correspond to 'natural categories' and that therefore it has some merits. However, a poor predictive performance would not by itself suffice to reject an ontology. Indeed, several reasons could explain such poor performances. The experts could have mislabeled a significant proportion of the sentences used for training and/or for testing. The learning algorithm and/or the coding strategies could be inappropriate for the task at hand. Finally, the classes could be badly represented by the learning examples. For instance, some classes in the uncertainty variable ontology have a very small number of training instances associated with them, e.g. '*inference from general to sensitive population*' has only 6 training instances. Even if they are good representatives, this can be insufficient to learn a good decision rule. Furthermore, a significant number of instances rightly belong to several categories. This renders the computation of a fair prediction rate more involved, while at the same time implying that decision rules associated with different nodes can in fact appear as closely related, and therefore difficult to distinguish.

### *B. Experimental results*

We tested both ontologies: the *ontology on judgment variables* with its single level and five nodes, and the *ontology on the uncertainty variables* with its depth of five levels and its 28 nodes. In the latter case, we tested first the distinction between all nodes that can be used by human coders to label the fragments

of texts. This allowed us, in particular, to compare the distinction between siblings (children of the same node) and between nodes not belonging to the same branch. Then, we studied further the quality of the ontology by looking more closely at the recognition rate of each internal node when subnodes are aggregated up in the tree. For instance, '*sampling*' was first considered as a node by itself, independently of its child nodes, and then as category when the instances of its child nodes were aggregated under its flag. Finally, we measured also the prediction rate when for each category (internal node) only its child nodes were taken into account. The motivation was to see if the internal nodes, that correspond to categories obtained by merging subcategories, were truly homogeneous and distinct, rather than resulting from somewhat arbitrary aggregation of the subcategories.

Below, we report the prediction performance in confusion matrices. Each row corresponds to a "true category", that is to the category to which an example was ascribed by the expert, and each column corresponds to the category predicted by the learning algorithm. The final column sums up the predictive performance. For instance, in Table 3, the category '*arbitrary assumption*' contains 53 instances of which 37 are well predicted (corresponding to a prediction rate of 69.81%). The other 16 instances are predicted as belonging to one of the remaining 17 categories (of which only 7 are presented in Table 3).

*Results for the judgment variables*

The experiments on the judgment variables show that a Naïve Bayes classifier achieves a 93% successful prediction rate on average on the 5 classes (measured by a 5-fold cross-validation)[7]. A further study measures the contrast between *pairs* of classes, resulting in the confusion matrix of Table 2. The labels '*Confidence*' and '*Expert assumption*' appear to be the most likely to be mislabeled.

**Table 2 Confusion matrix of successful predictions distinguishing pairs of judgment variables**

| | Confidence | Disagreement | Expert assumption | Hedged language | Precaution |
|---|---|---|---|---|---|
| Confidence | | 100,00 | 77,52 | 90,94 | 91,43 |
| Disagreement | | | 100,00 | 100,00 | 95,24 |
| Expert assumption | | | | 94,56 | 88,11 |
| Hegded language | | | | | 96,23 |
| Precaution | | | | | |

This is not surprising. Both '*Confidence'* and '*Expert assumption'* reflect a certainty about findings and thus there could be some semantic overlap.

*Results for the uncertainty variables*

Table 3 reports the confusion matrix for a subset of 8 of the terminal nodes[8]. The overall prediction rate measured is 75.57%, but there are variations between the categories. For instance, '*Limits of analytic methods*' and '*Poor data quality / flawed measurement*' are siblings, both children of the '*Measure*' node. However, one, '*Poor data quality…*' is perfectly predicted with 100% (with only 7 instances), while '*Limits of …*' is more poorly predicted with a rate of ~58% (with 58 actual instances). Conversely, the

[7] By comparison, SVM reaches 92% and *k*-NN 89%. Decision trees only reach 76% of good prediction which might be explained by the high dimension of the description space and by the sparseness of the data, each sentence being described by less that a few tens of attributes out of 2,530.

[8] For lack of space, we did not include here the whole confusion matrix with its 18 rows and columns. There are 18 terminal nodes.

latter seems to have a porous boundary with the '*Missing factors / variables*' node since there are 5 + 8 misclassified instances between the two nodes (out of 129 actual instances).

**Table 3 Confusion matrix of successful predictions distinguishing pairs of 8 uncertainty variables**

| | Arbitrary assumptions of… | Combination effects | Correlated causal factors | Limits of analytical methods | Missing factors … | Non-random … | Ontic uncer-tainty… | Poor data quality… | *Prediction* |
|---|---|---|---|---|---|---|---|---|---|
| Arbitrary assump-tions of … | **37** | 1 | 0 | 2 | 3 | 0 | 2 | 0 | *69.81* |
| Combina-tion effects | 0 | **19** | 0 | 0 | 0 | 0 | 0 | 0 | *100.00* |
| Correlated causal factors | 0 | 0 | **15** | 0 | 0 | 0 | 0 | 0 | *100.00* |
| Limits of analytical methods | 3 | 1 | 0 | **34** | 5 | 2 | 3 | 0 | *58.62* |
| Missing factors … | 5 | 1 | 2 | 8 | **40** | 5 | 0 | 0 | *56.34* |
| Non-random … | 0 | 0 | 0 | 2 | 1 | **38** | 1 | 0 | *80.85* |
| Ontic uncertainty … | 2 | 0 | 0 | 2 | 0 | 1 | **23** | 0 | *74.19* |
| Poor data quality … | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | *100.00* |

| | |
|---|---|
| *overall* | 75.57 |

Overall, it appears that the distinction between siblings (children of the same parent node) is often not higher than with nodes of distant branches in the ontology. In part, this may reflect the fact that the same instances can be ascribed to more than one category in the ontology, and even though we were careful to eliminate from the learning set instances that were classified in several categories, it remains that some categories can be associated with decision rules that accommodate the same sub-populations of examples, and are therefore prone to prediction "errors". (Figure 3 shows the difference in prediction rate when multiple classification is taken into account and when it is not).

The second set of experiments on the recognition rate of categories, the one corresponding to internal nodes, shows that, overall, recognition improves when subcategories are aggregated to form categories (see Figure 4). This suggests that the categories that were decided upon are rather homogeneous, and are not aggregating things that do not belong together. So, even though there are nodes belonging to different branches (categories) that exhibit a high rate of confusion, this is no longer the case when the categories themselves are compared.

**Figure 3. Difference in prediction rate with and without multiclass classification.**

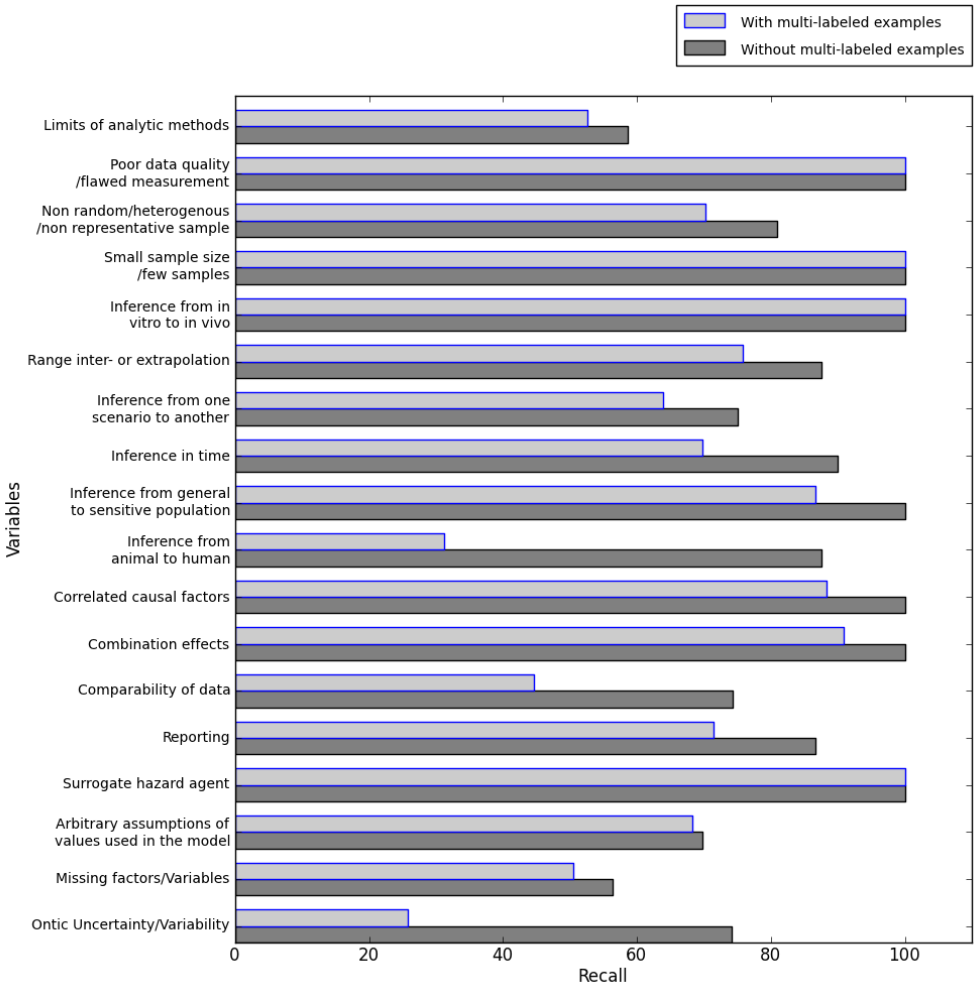
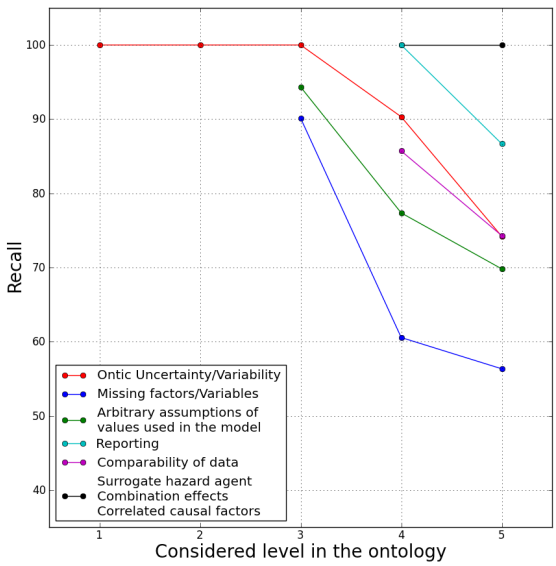
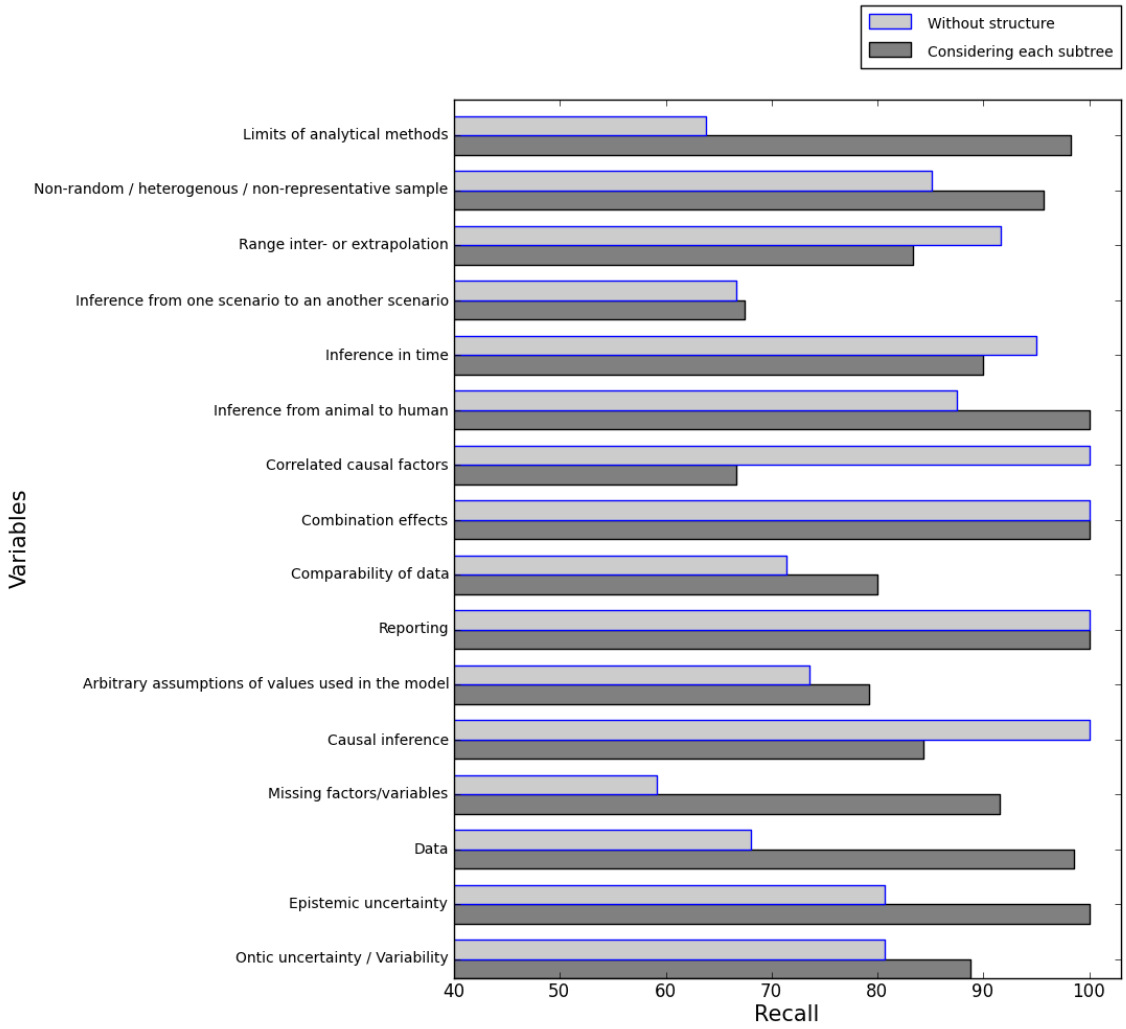



**Figure 4. Evolution of the recognition rate of some categories when aggregating subcategories level by level. The $x$ axis corresponds to the depth in the ontology. Going left corresponds to the aggregation up the levels towards the root of the ontology.**



Inversely, while the distinction between sibling nodes can sometime be hazardous when they are considered among all nodes, it is generally much easier to distinguish them when they are taken in isolation. For instance, Figure 5 shows that the recall of the class '*Limits of analytical methods*' goes from approximately 63% when all classes are considered to 98% when only the children of '*Measure*': '*Poor data quality* …' and '*Limits of analytical methods*' are to be distinguished. In general, the average prediction rate significantly increases if only the children of a category are considered as possible classes (one exception is the '*Causal inference*' class). This points to two consequences. First, this reinforces the view that the hierarchical organization of categories encoded in the ontology is not arbitrary. Second, this may suggest that coders should adopt a top-down coding process, by first deciding upon high level categories and then going down in the classification tree rather than trying to identify directly the class of a new piece of text.

**Figure 5. Difference of prediction rate with a direct labeling process and using a top-down labeling strategy**



## 5. Conclusion

The experiments reported here with Machine Learning techniques offer a first glimpse at what can be gained by testing ontologies in the making with these methods. In the context of this study, the experimental results seem to vindicate the soundness of both ontologies: one on the judgment variables and the other on the uncertainty variables. There is no doubt, however, that further experiments are needed in order to better assess the merits of ontologies, possibly leading to improved ones. Future developments will include the choice of a richer representation than the "bag of words" one, as well as experiments geared towards the optimization of the structure of the hierarchical ontology.

We will also compare the performance of Machine Learning techniques for the US and EU documents, as well as for contaminants and biohazards. Because in the US, risk assessments are generated by multiple agencies and by multiple expert panels, while in the EU a single bureaucratic structure is responsible for the reports, we expect more standardization and better prediction in the EU. As for the two types of hazards, while we expect them to have different types of uncertainties, we do not expect differences in the performance of Machine Learning: the logical structure of the ontologies should apply equally well.

Our ultimate goal is to devise a process, whereby coders are aided by Machine Learning and Machine Learning improves with new coding by humans. For that we will include uncoded sentences as a special category. The goodness of the prediction of the uncoded vs. coded distinction should reflect how explicit expert panels are about uncertainty. We expect predictions to improve over time.

## 6. References


Aly, M. (2005). Survey on Multiclass Classification Methods. *Neural Networks*, pp.1-9.

Flach, P. (2012). Machine Learning. *The Art and Science of Algorithms that Make Sense of Data*. Cambridge University Press.

Hammitt, J. K, J. B. Wiener, B. Swedlow, D. Kall and Z. Zhou. 2005. Precautionary Regulation in Europe and the United States: A Quantitative Comparison, *Risk Analysis*, 25, 1215-1228

Japkowicz, N. and Shah, M. (2011). *Evaluating Learning Algorithms. A Classification Perspective*. Cambridge University Press.

Levin, R., Hansson, S.O. and Rudén, C. (2004). Indicators of uncertainty in chemical risk assessments. *Regulatory Toxicology and Pharmacology*, 39, 33-43.

Lynch, D., Vogel, D. 2001. *The Regulation of GMOs in Europe and the United States: A Case-Study of Contemporary European Regulatory Politics*. New York : Council on Foreign Relations, Inc

Myers, Greg (1989): "The pragmatics of politeness in scientific articles". *Applied Linguistics*, 10: 1-35.

Perkins, J. (2010). *Python Text Processing with NLTK 2.0 Cookbook*. Packt Publishing, 2010.

Porter, M.F., 1980, An algorithm for suffix stripping, *Program*, 14(3) :130-137

Vázquez, Ignacio and Diana Giner (2009): "Writing with Conviction: The Use of Boosters in Modelling Persuasion in Academic Discourses". *Revista Alicantina de Estudios Ingleses* 22 (2009): 219-237

Walker, W., Harremoës, P., Rotmans, J., Van der Sluijs, J., Van Asselt, M.B.A, Jansen, P., Krayer von Krauss M.P. (2003). Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support. *J Integr Assess*, 4, 1, 5-17.

Conway, Erik and Naomi Oreskes, 2010. *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*. USA: Bloomsbury

Shackley, S. and Wynne, B., 1996. Representing Uncertainty in Global Climate Change Science and Policy: Boundary-Ordering Devices and Authority. *Science, Technology, & Human Values*, 21 (3), pp. 275-302.

Proctor, R. and L. Schiebinger eds. 2008. *Agnotology: The Making and Unmaking of Ignorance*. Stanford: Stanford University Press.

Popper, Karl R. "The Logic of the Social Sciences," in *The Positivist Dispute in German Sociology*, translated by Glyn Adey and David Frisby (London: Heinemann, 1976), pp. 87-104.

Lakoff, George. 1972. "Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts." *Journal of Philosophical Logic,* 2: 458-508.

Hyland, Ken. 1996. "Writing without Conviction? Hedging in Science Research Articles." *Applies Linguistics*, 17/4:433-454

Crompton, Peter. 1997. Hedging in Academic Writing: Some Theoretical Problems. *English for Specific Purposes*, Vol. 16, No. 4, pp. 271-287

Hattis, Dale and David E. Burmaster. 1994. Assessment of Variability and Uncertainty Distributions in Practical Risk Analyses. *Risk Analysis*, 14/5, 713-730

Morgan, M.G. and M. Henrion. 1990. *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge, UK: Cambridge University Press

Pate-Cornell, M. E. 1996. Uncertainties in Risk Analysis: Six Levels of Treatment. *Reliability Engineering and System Safety*, 54, 95-111

van Asslet, Majrolein B. and Jan Rotmans. 2002. "Uncertainty in Integrated Assessment Modelling. From Positivism to Pluralism". *Climatic Change*, 54: 75-105.

van der Sluijs, Jeroen P., Matthieu Craye, Silvio Funtowicz, Penny Kloprogge, Jerry Ravetz, and James Risbey. 2005, "Combining Quantitative and Qualitative Measures of Uncertainty in Model-Based Environmental Assessment: The NUSAP System." *Risk Analysis*, Vol. 25, No. 2, pp. 481-492.

Grandjean, P. and E. Buz-Jorgensen E. (2007). "Total Imprecision of Exposure Biomarkers: Implications for calculating Exposure Limits." *American Journal of Industrial Medicine*, 50, p. 512-519.

Kang S., R.L. Kodell, and J.J. Chen. (2000). „Incorporating model uncertainties along with data uncertainties in microbial risk assessment, *Regulatory Toxicology and Pharmacology,* 32, p. 68–72.

Nautta, Maarten J. 2000. "Separation of Uncertainty and Variability in Quantitative Microbial Risk Assessment Models." *International Journal of Food Microbiology,* 57, 9-18

Dorne J. L. and A.G. Renwick. 2005. "The Refinement of Uncertainty/Safety Factors in Risk Assessment by the Incorporation of Data on Toxicokinetic Variability in Humans. *Toxicological Sciences,* 86, p. 20–26.

Kroes, R., ,D.Muller, J. Lambe, M.R.H. Lowik, J. van Klaveren, J. Kleiner, R. Massey, S. Mayer, I. Urieta, P. Verger, A. Visconti. 2002. "Assessment of intake from the diet." *Food and Chemical Toxicology* 40 (2002) 327–385

EFSA. 2006. Guidance of the Scientific Committee on a request from EFSA related to Uncertainties in Dietary Exposure Assessment, *The EFSA Journal*, 438, p. 1-54.

WHO/IPCS (World Health Organization/International Program on Chemical Safety). 2007. *Guidance Document on Characterizing and Communicating Uncertainty in Exposure Assessment*.