

Bounding analysis applied to lung cancer risk

Minh Ha-Duong* Elizabeth Casman M. Granger Morgan

February 7, 2003

Abstract

For cancers with more than one risk factor, the sum of estimated numbers of cancers attributed to the individual factors may exceed the total number of cases observed. In this study we bound the fraction of lung cancer occurrences not attributed to specific well-studied causes, in order to keep estimates of the less well delimited risks consistent with those of known risks. Available data and expert judgment are used to attribute portions of the observed lung cancer incidence to known causes such as smoking, residential radon and asbestos exposure, to describe the uncertainty surrounding these estimates, and quantify the interaction between pollutants. An upper bound on the residual risk is inferred using a coherence constraint on the total number of deaths and the principle of maximum uncertainty, using imprecise probabilities.

1 Introduction

Usually, the risk of exposure to environmental contaminants is calculated in a front-to-back mode, which involves estimates of toxic releases, environmental and physiological transformations, exposure models and dose-response functions, see for example Committee on Health Risks of Exposure to Radon (BEIR VI), Board on Radiation Effects Research, Commission on Life Sciences, National Research Council [1999]. That methodology works well when the relevant science is well developed, but the uncertainties in the individual studies make it difficult to draw a consistent picture of the whole health issue.

This paper takes the opposite, back-to-front, approach and presents a method of risk analysis applicable when only part of the problem is adequately characterised. In this study we bound the fraction of lung cancer occurrences not attributed to specific well-studied causes, in order to keep estimates of the less well delimited risks consistent with known risks. Some of the major environmental risk factors for lung cancer are shown table 1. "Well characterized" here means that population-wide longitudinal attributional studies exist.

*Chargé de Recherches au CNRS, visiting researcher at the Center for Integrated Study of the Human Dimensions of Global Change. Mail to minh.ha.duong@cmu.edu, or at Engineering and Public Policy Department, Carnegie-Mellon University, 5000 Forbes Avenue, Pittsburgh PA 15213.

Well characterized factors	Less well characterized factors
Cigarette smoking	Occupational exposures:
Passive smoking	Asbestos
Indoor radon	Arsenic
	Chromates
	Chloromethyl ethers
	Diesel exhaust
	Nickel
	Polycyclic aromatic hydrocarbons (PAHs)
	Ambient air pollution

Table 1: Environmental risk factors for lung cancer

Morgan [2001] argued that methods of bounding analysis could be used for environmental risk analysis. For health risks with multiple external causes, the available knowledge constrains the magnitude of the poorly characterized risks. If most risks were known with precision, this would be a simple subtraction problem. However disease risks from environmental causes are often estimated from models or inferred from studies involving limited numbers of subjects and inconsistent notions of controls or have other methodological problems that contribute to the uncertainty of the results. It is common to see the central tendencies of such risk estimates expressed as ranges, especially when there are competing plausible models. Sometimes the sum of the individual risks exceeds the total risk. How to quantify and bound the residual “unclaimed” risk is the subject of this paper.

The case study is lung cancer in the United States. This group of cancers has multiple demonstrated environmental causes as well as several suspected or poorly described causes. Expert judgment is used to attribute a portion of the observed cancers to known causes such as smoking, radon and asbestos. Information about the risks from unspecified causes are inferred using a coherence constraint on the total number of deaths, and the principle of maximum uncertainty.

Our proposed method builds upon the seminal work of [Walley, 1991, chapter 4]. Mathematically, this is an application of [Smets, 2000] Transferable Belief Model, which elaborates on the Shafer [1976] theory of evidence. We elicit information about a finite set of variables (risk factors for cancer) and, represent this information as constraints on a linear programming problem involving a convex family of probabilities. We invoke the maximum unspecificity criterion in order to estimate the upper bound for the less well-studied members of the set. Ours is not the first combination of linear programming, expert elicitation, and imprecise probabilities. Lins and de Souza [2001] combined these elements to elicit prior probabilities for a single continuous parameter.

This text is organized as follows. Section 2 presents the conceptual model, which is an application of the mathematical Transferable Beliefs Model to risk assessment. Based on this, Section 3 discusses our method to elicit and validate expert opinion using a maximum unspecificity criterion. Expert elicitation in Section 4 is a preliminary numerical application.

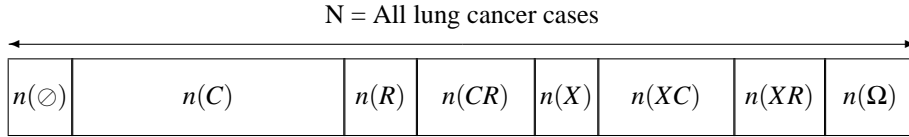


Figure 1: The basic statistic n . Breakdown of the total number of lung-cancer deaths by causative factors, including synergistic causes. Cause A not shown for picture clarity.

2 Model

2.1 Multiple pollutants may cause lung cancer

Denote N the health end-point, that is the total annual number of lung cancer deaths. Denote Ω the set of the possible causes of lung cancer deaths. For example, $\Omega = \{C, R, A, X\}$ where C means tobacco smoke primarily from cigarettes, R means indoor exposure to radon, A means asbestos and X is the group of poorly understood environmental factors of interest.

The model assumes that N is readily observable and therefore known with precision, see Archer and Lyon [2000] for a discussion. It is also assumed, classically, that exposure is an all-or-nothing affair. With these two assumptions, each death can be linked to zero or more possible causes in Ω . Most deaths are caused by smoking alone, but there are synergistic cases where more than one cause is involved, such as smoking and radon.

Figure 1 shows one way to subdivide N by causes that includes synergistic effects. We denote $n(s)$ the number of deaths linked to causes s , where s denotes any subset of Ω . Since we consider four possible causes in Ω , there could be sixteen ($= 2^4$) possible s , but to simplify the analysis and to be consistent with the cancer literature, we will consider only the two-factor interactions involving cigarette smoke.

To adopt a more precise and cautious definition, $n(s)$ is the number of cases not exposed to pollutants not in s . This implies that causes not in s are known to be non-contributing to that lung cancer. For deaths in $n(s)$, any cause in s may have caused the lung cancer, but which one is uncertain and there may have been synergies. The two extreme subsets need more explanation.

The number of lung cancer deaths where all causes of Ω have been positively excluded counts toward $n(\emptyset)$ shown to the left of the bar Figure 1. Cases that could not be linked to any pollutant in Ω are considered spontaneous lung cancer.

On the other end, the full set $n(\Omega)$ (a short notation for $n(CRAX)$) corresponds to the situation when no cause of lung cancer has been excluded because either all risk factors have been observed to be present, or there is no information about risk factors. In this way, $n(\Omega)$ also denotes the missing data points, that is deaths about which no information has been recorded.

Direct measurement of this basic statistic n is impossible, since exposure to a pollutant does not necessarily result in a cancer fatality and because retrospectively, lifetime exposures to the various carcinogens can only be roughly estimated. In Section 3 we

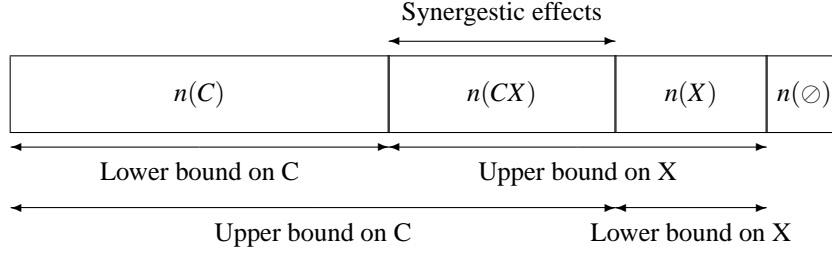


Figure 2: Upper and lower bounds on the number of cases attributable to C and X

determine n using expert elicitation. For now, let us assume n is known.

2.2 Bounding the risk attributable to single and joint pollutants

Assume for now that the background rate r_0 is negligible. The basic statistic n can be used to bound the number of cases attributable to, for example, smoke C as follows:

- The lower bound is the number of cases attributed only to smoking (we lump both passive and active smoking together). This is $n(C)$.
- The upper bound is the number of cases exposed to smoke and possibly other factors. That is $\sum_{E, C \in E} n(E)$.

Denote $\bar{n}(C)$ and $\underline{n}(C)$ the upper and lower bounds, respectively. For example, the upper bound on cigarette causality is: $\bar{n}(C) = n(C) + n(XC) + n(CR) + n(XCR) + n(CA) + n(XCA) + n(CRA) + n(XCRA)$.

Figure 2 illustrates this definition of the upper and lower bounds of the number of lung cancer deaths attributable to X and C . For clarity the figure is drawn showing only two causes, with $\Omega = \{C, X\}$.

In epidemiologists' terms, the *attributable fraction* of pollutant C is the proportion of all cases that could be avoided if this pollutant were eliminated. Denote $af(C)$ the attributable fraction. The model suggests the following bounds for smoking attributable fraction:

$$\frac{\underline{n}(C)}{N} (1 - r_0) \leq af(C) \leq \frac{\bar{n}(C)}{N} \quad (1)$$

The lower bound account for the $1 - r_0$ share of spontaneous lung cancer cases in those cases exposed to cigarette. The upper bound attributes all cigarette-exposed deaths to this factor.

For this paper we assume that the background rate of lung cancer in the U.S., r_0 , is 3 cases out of 100.000 people. This background rate can be defined as the ratio of cases over the unexposed population. Denoting P the total population, p_C , p_R and p_A the exposure probabilities to C, R, A respectively, and assuming independence:

$$r_0 = \frac{n(\emptyset)}{(1-p_C)(1-p_R)(1-p_A)P} \quad (2)$$

Consider now the bounds on deaths attributed to multiple causes. Denote these causes s , a subset of Ω , for example $s = CR$. For the lower bound on the number of deaths attributable to these causes acting jointly, we continue to adopt the number of cases exposed only to these causes, that is:

$$\underline{n}(s) = n(s) \quad (3)$$

And as the upper bound, we continue to adopt the number of cases exposed to s and possibly other factors, that is:

$$\bar{n}(s) = \sum_{E, s \subseteq E} n(E) \quad (4)$$

That \bar{n} corresponds to the commonality function in the Transferable Belief Model. Bounds on the attributable fraction can be computed as in equation 1.

2.3 Unspecificity, a measure of uncertainty

Structurally, the only uncertainty in this bounding analysis model comes from the synergistic causes, because it is not possible to attribute the cancer to any one of these causes. Consider these two (of the three) extreme cases:

- If each death were attributed to exactly one cause, then there would be no uncertainty, and all lower bounds would coincide with their upper counterpart. We would have $n(C) + n(R) + n(A) + n(X) = N - n(\emptyset)$. Note that since n is a positive function that sums up to N , this implies that $n(s) = 0$ for all other subsets.
- If no information were available, each death would be attributed to the synergy of all factors. We would have all the lower bounds at 0 and all upper bounds at N . Mathematically, this is $n(\Omega) = N$. Note that this constitutes a proper uninformative distribution: it is not the uniform probability distribution on Ω .

Unspecificity is an numeric indicator that equals one in the first case, and in the second case equals the number of elements of Ω . It is the expected value of the number of elements of s with respect to the probability distribution $m(s) = \frac{n(s)}{N}$, that is:

$$U = \frac{n(C) + n(R) + n(A) + n(X) + 2(n(CR) + n(RA) + \dots) + 3 \dots + 4n(\Omega)}{N} \quad (5)$$

In this paper unspecificity is a kind of generalized cardinality, that specifies the number of alternatives. The reason for using this word is that when a death is attributed to the synergy of k factors, it can be said that the unspecificity of this information is k . See Rocha [1997] for an extensive discussion of this concept.

A lower unspecificity measure corresponds to better information, so the third extreme case needs discussion: unspecificity is zero when and only when $n(\emptyset) = N$. This is the case when for all deaths, all non-spontaneous causes of Ω have been positively excluded. It means that all the substances in Ω are actually safe (with respect to lung cancer). This is the highest level of information achievable, to the point that it makes Ω irrelevant.

Note how this interpretation hinges on the idea of counting missing data with $n(\Omega)$. This is an application of a general principle of maximum uncertainty, also known as Laplace’s principle of “raison insuffisante”. The principle states that one should select the statistic that is the most unspecific, compatible with existing information. This is the principle that we use next section to estimate the bounds on the unknown cause, given information about all others.

3 Expert elicitation

3.1 Procedure

The goal of the procedure is to determine a $n(s)$ vector representing an expert judgment. This judgment is elicited by asking questions on the epidemiological measures of risk. The answers are interpreted as linear constraints on n . These constraints determine a set \mathcal{B} of basic statistics. The most unspecific n in \mathcal{B} is chosen to represent the expert judgment, according to the maximum uncertainty principle. This amounts to solving a linear program in a space with $2^{|\Omega|}$ dimensions.

In detail, here are sample constraints derived from a hypothetical expert elicitation that define \mathcal{B} . It is understood that all $n(s)$ are non-negative, summing up to N .

- It is important to underline that $n(\emptyset)$ does not have the same status as $n(X)$, which will be deduced as a residual. The number of spontaneous cancers is a parameter that needs to be elicited directly by asking about the background rate r_0 (see equation 2). With our assumptions on exposure probabilities, we found $n(\emptyset) = 0.013N$.
- Three-way interactions and higher are not allowed. That is, $n(s) = 0$ if s has 3 or more elements.
- The smoking attributable fraction of causes other than smoking is at most 90 percent. For example, $\frac{\bar{n}(R)}{N - n(0)} \leq 0.9$.
- Smoking alone causes 95 percent of smoking-related cases. That is $n(C) = 0.95\bar{n}(C)$

Other ways of translating judgments into constraints are possible, for example using relative risk, but are not used in this introductory paper. Note that both quantitative and comparative judgments are possible, which may ultimately be important because some of the pollutants have been well studied, but we are interested in the less well-known pollutants.

3.2 Validation with compatible probability distributions

The direct way to validate the basic statistic n (in the sense to see if the expert agrees with the mathematical reinterpretation of his knowledge) would be to ask him to evaluate it. However, evaluating $2^{|\Omega|}$ numbers could be cognitively overloading. An alternative is based on the notion of a compatible probabilistic distribution [Dempster, 1967]. A probabilistic attribution P is a breakdown of N by causes disregarding any synergistic effect.

We say that a probabilistic attribution P is compatible with the basic statistic n if and only if it does not contradict the bounds determined by n , that is:

$$\forall s \subset \Omega, \sum_{\emptyset \neq x \in s} P(x) \leq \sum_{E, s \cap E \neq \emptyset} n(E) \quad (6)$$

The right hand side of Equation 6 corresponds to the belief function in the Transferable Belief Model. This function of s is interpreted in the present model as the upper bound on the number of deaths related to the causes in s acting either jointly *or separately*.

From here on, P denotes a probability attribution compatible with the basic statistic n , and \mathcal{P}_n denote the family of all such probabilistic attributions compatible with n . This \mathcal{P}_n carries all the information about n .

It gives a geometric interpretation to the basic statistic: \mathcal{P}_n is a convex polyhedron in a space with $|\Omega|-1$ dimensions. Figure 3 represents this convex set corresponding to the basic statistic previously shown in figure 1.

The fundamental criteria to assess the validity of an expert's implicit n is that all compatible probabilistic attributions should appear equally likely, no compatible probabilistic attribution should appear unrealistic. This translates to the idea that synergistic effects can not be attributed to any single cause.

Of course, \mathcal{P}_n is an infinite family and it is not possible to check perfectly these two criteria. Still it is possible to present a few compatible probability attributions to the expert, and ask if he agrees that they are all equally likely, and that none of them is unrealistic. This could involve vertexes of \mathcal{P}_n , or some point in the middle.

3.3 Other validations: risk-ranking, relative risk

Another way to validate n is to examine the risk-ranking it implies, using the natural partial order defined by a basic statistic n . It will be said that according to n , risk Y is no bigger than risk X when it causes less deaths according to all realistic probabilistic attributions of causes (assuming that all distributions in \mathcal{P}_n are realistic is the other validation criterion). Mathematically:

$$Y \preceq X \Leftrightarrow \forall P \in \mathcal{P}_n, P(Y) \leq P(X) \quad (7)$$

It is possible to work with the full partial order, since there is at most $|\Omega|(|\Omega|-1)/2$ comparisons. Assuming $|\Omega| = 7$ for example, there are no more than 21 information items, which can be presented naturally in the diagonal half of a table. Moreover, practically there will be fewer than 21 items, since not all risks can be compared. It is

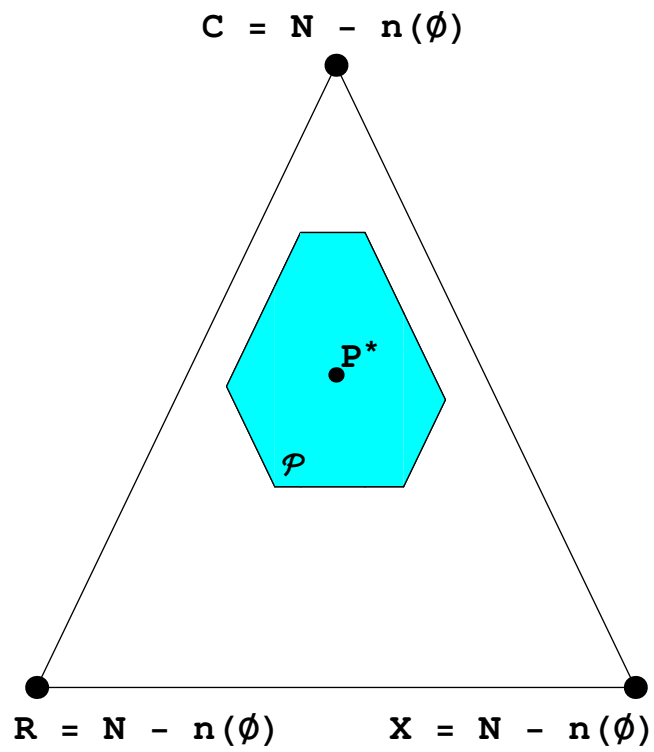


Figure 3: The set \mathcal{P} of all probabilistic attributions compatible with the basic statistic shown in figure 1.

to be expected, for example, that some experts may not wish to compare some of the less-known risks because of missing information.

We will ask experts to rank risks during the elicitation process. Results will be validated by comparing the partial order derived from n with the expert's *a priori* risk ranking.

A third way to validate the findings is to consider if the relative risks and the interactions between pollutants make sense. The definition of relative risk for smoking cigarettes $rr(C)$, for example, is the lung cancer rate associated with exposure to tobacco smoke divided by the background lung cancer rate. Given exposure probabilities in the general population, we will assess the bounds on the relative risk for the various pollutants using the formula in [Committee on Health Risks of Exposure to Radon (BEIR VI), Board on Radiation Effects Research, Commission on Life Sciences, National Research Council, 1999, appendix C p. 229]. Some experts may have greater familiarity with this statistic, and would be able to calibrate $n(s)$ by its effects on $rr(s)$.

4 Application

This paper's numerical simulations were performed using a *Mathematica* notebook¹. The code directly implements matrix calculus for belief functions as outlined in Smets [2001]. This is the most straightforward method given that Ω remains small, but it would not scale well to tens of pollutants, since it involves square matrices with $2^{2|\Omega|}$ elements. For example, 10 pollutants implies storing in memory arrays with 1M numbers.

In this section Ω the set of possible causes of lung cancer is:

- C Smoking
- R Radon
- A Asbestos, glass wool, ceramic fibers
- X All other environmental risk factors

The following preliminary data were used to illustrate the methodology. We supply a fictitious elicitation result showing responses given as point estimates, ranges, and inequalities:

	$ar(s)$	Smoking attributable fraction
C	0.75–0.85	
R	0.10–0.15	≤ 0.90
A	0.01–0.05	≤ 0.90
X	?	≤ 0.90
\emptyset	0.013	

which translates to the following constraints:

$$\begin{aligned}
 0.75 &\leq af(C) \leq 0.85 \\
 0.10 &\leq af(R) \leq 0.15 & \text{and} & \bar{n}(\{C, R\}) \leq 0.90\bar{n}(\{R\}) \\
 0.01 &\leq af(A) \leq 0.05 & \text{and} & \bar{n}(\{C, A\}) \leq 0.90\bar{n}(\{A\}) \\
 af(X) &=? & \text{and} & \bar{n}(\{C, X\}) \leq 0.90\bar{n}(\{X\})
 \end{aligned}$$

¹Available on the web or upon request, under the GNU General Public License.

The next table shows the implications of the most unspecific imprecise probability distribution compatible with these constraints. Bounds on the attributable fraction af is reported as percentage, while the relative risk rr is by definition relative to the background rate. The exposure probabilities needed to compute rr are exogenous and preliminary. The effect of this calculation on the bounds of rr would serve as a calibration/validation reference for the expert who may be more familiar with small sample studies than population effects, and might adjust his initial responses in light of seeing their mathematical implications.

Bounds	C	R	A	X
\overline{af}	85	15	5	6
\underline{af}	75	10	1	0
<i>Exposure probability</i>	50	50	5	5
\overline{rr}	12.3	1.4	2.1	2.4
\underline{rr}	7.	1.2	1.2	1.

This result attributes up to six percent of lung cancer deaths to X , the group of unknown environmental pollutants. Where causes are suspected but population exposure models are yet unavailable, (for example the case of occupational exposures to diesel exhaust), the risk analyst knows that the upper bound of its effect is six percent of total lung cancer cases, no more than ninety percent of which would be attributed to smoking.

5 Concluding remarks

5.1 Discussion

With less than ten pollutants, computing time is not a problem. Expert elicitation could be done interactively, solving for n after each expert's reply. This would allow the interviewer to point out and resolve inconsistency when there is no solution. But assuming that experts were willing to form judgments on a wider range of pollutants, the curse of dimensionality can be addressed along the following lines. Rather than using matrix calculus, it is possible to use faster algorithms (namely the Fast Möbius transform) for belief function computations. If this is not enough, further simplifications can be made if additional assumptions on n , for example disallowing 3-way interactions or more, are accepted.

The proposed method takes all information items provided by the expert with equal force. A potential advance of this research could be to ask experts to rank the reliability of each information item, or even to give an estimate of confidence for them.

Further research could deal with inter-expert validation, a question linked with the unresolved issue of judgment fusion. The Transferable Belief Model underlying this work offers a measure of contradiction between different sources of information: it reinterprets $n(\emptyset)$, the number of spontaneous lung cancer deaths found when one combines the opinion of all experts. The problem is how to combine the experts.

Each expert's judgment determines a set \mathcal{B} of coherent basic statistics. Consider the intersection of all these sets: it is represented simply by putting all the information items from all experts together. If the intersection of all these sets is non-empty, then experts agree on this intersection. The principle of maximum unspecificity can be used to form a group judgment.

If the intersection is empty, the experts contradict each other. Studying which information items cause the contradiction (which constraints make the LP infeasible) can lead to the substantive sources of disagreement, and in that way inform both future research priorities as well as the decision-making process. How (or if) to fuse the judgments and quantify the degree of contradiction is still an active research question, see ISI [2000] for example.

5.2 Conclusion

This manuscript proposed an application of the [Smets, 2000] Transferable Belief Model to estimate an upper bound on the number of lung cancers caused annually by the group of causes for which comprehensive longitudinal studies are lacking. Such a result is interesting from a risk management perspective, as it gives an indication of the level of effort control of these pollutants deserves.

This was done by attributing a portion of the observed cancers to known causes such as smoking, radon and asbestos, and then deducing information about the residual using maximum unspecificity. This procedure deals with the following critical aspects:

1. Uncertainty in the known causes is explicitly stated, using statements on upper and lower bounds.
2. Synergistic effects in the known causes are part of the framework.
3. The consistency between known causes and poorly understood agents are key to the analysis. As Figure 2 illustrates, it is the lower bound on smoking that mostly constrains the upper bound on the residual.

This paper presents the methodology. The epidemiology revealed by future expert elicitation will be the subject of another paper.

References

- V. E. Archer and J. L. Lyon. Errors and biases in the diagnosis of cancer of the lung and their influence on risk estimates. *Medical hypotheses*, 54(3):400–407, 2000.
- Committee on Health Risks of Exposure to Radon (BEIR VI), Board on Radiation Effects Research, Commission on Life Sciences, National Research Council. *Health effects of exposure to radon*. National Academy Press, 1999. ISBN 0-309-05645-4.
- A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.

- Fusion 2000 Conference*, Paris, 10–13 July 2000. ISIF (International Society on Information Fusion). URL <http://www.onera.fr/fusion2000/>.
- Gertrudes Coelho Nadler Lins and Fernando Menezes Campello de Souza. A protocol for the elicitation of prior distributions. New York, USA, 26–29 June 2001. Cornell University. URL <http://ippserv.rug.ac.be/~isipta01/>.
- Granger Morgan. The neglected art of bounding analysis. *Environmental Science and Technology*, apr 1:162A–164A, 2001.
- Luis M. Rocha. Relative uncertainty and evidence sets: a constructivist framework. *International Journal of General Systems*, 26(1–2):35–61, 1997.
- Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton (NJ), 1976. ISBN 0-691-10042-X (hardback).
- Philippe Smets. Belief functions and the transferable belief model, 2000. URL <http://ippserv.rug.ac.be>.
- Philippe Smets. Matrix calculus for belief functions, 2001. URL <http://iridia.ulb.ac.be/~psmets/MatrixRepresentation.pdf>.
- Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991. ISBN 0-412-28660-2.