Evidence synthesis through a degradation model applied to myocardial infarction.

Daniel Commenges, Boris P. Hejblum

Lifetime Data Analysis manuscript No.
(will be inserted by the editor)

Evidence synthesis through a degradation model applied to myocardial infarction

Received: date / Accepted: date

Abstract We propose an evidence synthesis approach through a degradation model to estimate causal influences of physiological factors on myocardial infarction (MI) and coronary heart disease (CHD). For instance several studies give incidences of MI and CHD for different age strata, other studies give relative or absolute risks for strata of main risk factors of MI or CHD. Evidence synthesis of several studies allows incorporating these disparate pieces of information into a single model. For doing this we need to develop a sufficiently general dynamical model; we also need to estimate the distribution of explanatory factors in the population. We develop a degradation model for both MI and CHD using a Brownian motion with drift, and the drift is modeled as a function of indicators of obesity, lipid profile, inflammation and blood pressure. Conditionally on these factors the times to MI or CHD have inverse Gaussian ($\mathcal{IG}$) distributions. The results we want to fit are generally not conditional on all the factors and thus we need marginal distributions of the time of occurrence of MI and CHD; this leads us to manipulate the inverse Gaussian normal distribution ($\mathcal{IGN}$) (an $\mathcal{IG}$ whose drift parameter has a normal distribution). Another possible model arises if a factor modifies the threshold. This led us to define an extension of $\mathcal{IGN}$ obtained when both drift and threshold parameters have normal distributions. We applied the model to results published in five important studies of MI and CHD

and their risk factors. The fit of the model using the evidence synthesis approach was satisfactory and the effects of the four risk factors were highly significant.

Keywords Causality · Causal inference · Coronary heart disease · Degradation model · Epidemiology · Evidence synthesis · Inverse Gaussian distribution · Myocardial infarction · Stochastic processes

1 Introduction

Most often in epidemiology, risks of events are modeled using a proportional hazard assumption. Degradation models may be closer to physiological mechanisms in many cases; indeed it is often the case that an event occurs when a degradation process reaches a certain threshold (Doksum and Normand, 1995; Aalen and Gjessing, 2001; Hashemi et al, 2003; Lee and Whitmore, 2006; Aalen et al, 2008). A good example is myocardial infarction (MI). It has been shown that MI is most commonly due to occlusion (blockage) of a coronary artery following the rupture of an atherosclerotic plaque (Hansson, 2005; Nicholls, 2009). A Brownian motion with positive drift seems well adapted to describe the progressive growth of atheroma, with MI occurring when this process reaches a certain threshold. One advantage of this approach is that we can link MI with broader coronary heart disease (CHD) events which happen before MI when occlusion is not complete but the heart already suffers from hypoxia. These CHD events may occur when the atheromatous process reaches a threshold below that required for MI. One of the reason to use such a model is to express the effect of risk factors. Several risk factors are already known for MI. Most analyses focus on one particular risk factor rather than presenting a global model. Few works have attempted to develop more global dynamic analysis: Wilson et al (1998), using data from the Framingham study, developed prediction scores using indicators of lipid profile, diabetes, obesity, blood pressure and tobacco consumption as explanatory variables in a conventional Cox model; Gamborg et al (2011) used the dynamic path analysis of Fosen et al (2006) to take into account the possible evolution of obesity and blood pressure, but they did not take into account other factors.

However, for calibrating complex dynamic models, cohorts having recorded all the relevant risk factors may be lacking; in particular, a factor like C-reactive pro-

tein (CRP), a marker of inflammation, has attracted interest only recently (Ridker et al, 1997) so that this marker is not available in historical studies. So, there is an interest in developing a method for synthesizing evidence from different studies with the following potential advantages: (i) incorporating recently discovered factors, (ii) investigating to which extent the results are consistent across studies, (iii) gaining power. Synthesis studies are different from meta-analysis (Van Houwelingen et al, 2004): while meta-analyses aim at giving a global conclusion based on several studies which have estimated the same parameter (the effect of treatment for instance), synthesis analyses aim at incorporating information of different types for estimating the parameters of a global model (parameters which may not appear in any single study). Evidence synthesis (other than meta-analyses) is chiefly represented by "Bayesian synthesis" (Raftery et al, 1995), although a likelihood approach was suggested by Schweder and Hjort (1996); this has also a link with so-called indirect methods (Jiang and Turnbull, 2004). The literature on evidence synthesis based on dynamical models is rather scarce; an example in the field of HIV epidemiology is Presanis et al (2011) who used a Bayesian approach.

The aim of the paper is to present an approach based on a degradation model for synthesis of information of different types coming from different studies and to apply it to the epidemiology of MI and CHD. The paper is organized as follows. In section 2 we present a frequentist approach to evidence synthesis. Section 3 proposes a degradation model for MI and (CHD); on our way we come upon a new distribution, the inverse Gaussian normal-normal ($\mathcal{IGNN}$) distribution. Section 4 presents the application of the approach to several studies of MI and CHD. Section 5 concludes.

2 An approach to evidence synthesis through a dynamical model

Consider a dynamical statistical model describing the causal influences of different processes. It is expressed in terms of a possibly multivariate stochastic process, the possible laws of which are indexed by a parameter vector θ, $\theta \in \Theta$. The parameter θ could be estimated from observations of one study. It may happen that no study contains information on all the processes involved in the model. In that case it may be necessary to synthesize the information of several studies. This has the

advantage of giving a more robust result (not relying on a single study) and also gives the opportunity of examining whether the different studies have consistent results. In a synthesis analysis we assume that study k gives an estimate $\tilde{Q}_k$ of a quantity Q_k, accompanied by a standard error σ_k that is assumed known for sake of simplicity. For instance Q_k may be the incidence of a certain condition in one study, the incidence of another condition in another study, a relative risk in yet another study. So let us admit that we can define Q_k and express it as a function of the parameters θ: we have $Q_k(\theta)$.

We propose to consider that $\tilde{Q}_k$ is like an observation of $Q_k(\theta^*)$, where θ^* would be the true parameter value if the model was well-specified. Typical $\tilde{Q}_k$ are proportions or maximum likelihood estimates of relative risks. In both cases they are asymptotically normal. Thus we can write a contribution to a pseudo-loglikelihood:

$$L_k(\theta) = -\frac{(\tilde{Q}_k - Q_k(\theta))^2}{2\sigma_k^2}. \tag{1}$$

If a study gives several contributions, $\tilde{Q}_k$ is a vector which is assumed to have a multinormal distribution with variance matrix Ω_k. The contribution to the pseudo-likelihood is then:

$$L_k(\theta) = -\frac{1}{2}(\tilde{Q}_k - Q_k(\theta))^T \Omega_k^{-1} (\tilde{Q}_k - Q_k(\theta)). \tag{2}$$

Summing over k we get the total pseudo-loglikelihood: $L(\theta) = \sum_{k=1}^{K} L_k(\theta)$. The estimate $\hat{\theta}$ maximizes $L(\theta)$. If $L(\theta)$ has a maximum the problem is identifiable. If it does not, we can either add new studies, or add *a priori* knowledge on the parameters or other quantities Q_k in a Bayesian spirit.

It is good here to give an example. We have developed a degradation model (described in section 4) which gives the joint distribution of the time of occurrence of MI and of risk factors; these distributions are indexed by a parameter vector θ which includes in particular the effects of the risk factors on the drift, the threshold parameter(s), and the correlations between risk factors (see Table 9). The Atherosclerosis Risk In Communities (ARIC) surveillance study (National Heart Lung and Blood Institute, 2006) provided an estimate of the incidence of MI among American men. For instance, the estimated incidence in the 55-64 age stratum based on the observation of 32,572 person-years was $\tilde{Q}_k = 6.26‰$. Here

we have:

$$Q_k(\theta) = \frac{1}{10}\frac{F_\theta(65) - F_\theta(55)}{1 - F_\theta(55)}, \tag{3}$$

where F_θ is the marginal cumulative distribution function of time of occurrence of MI in our model given the value θ of the parameters. The variance of $\tilde{Q}_k$ in the person-years method is classically estimated by $\sigma_k^2 = \frac{\tilde{Q}_k}{32\,572}$; this is justified by assuming that the observed number of cases has a Poisson distribution (Clayton and Schifflers, 1987). In fact the ARIC study yielded estimates of both MI and CHD, so we had to use formula (2).

We do not expect that the model for the $\tilde{Q}_k$ is well specified. We assume that the $\tilde{Q}_k$ are independent. Let θ_0 the value which maximizes the expectation of the pseudo-likelihood. Under mild regularity assumptions, an extended theory of M-estimators (Van der Vaart, 2000; Freedman, 2006) ensures the consistency of $\hat{\theta}$ for θ_0. The variance of $\hat{\theta}$ cannot be estimated directly by the inverse Hessian of the pseudo-loglikelihood (H^{-1}), but by the sandwich estimators:

$$Var\left(\widehat{\theta}\right) = H^{-1}\left[\sum_{k=1}^{K} U_k U_k^T\right] H^{-1}, \tag{4}$$

where $U_k = \frac{\partial L_k}{\partial \theta}_{|\hat{\theta}}$ is the score for study k.

When a study gives several contributions we should evaluate the covariance matrix Ω_k and use formula (2). In many cases however the correlations between contributions are very small so that they can be treated as independent: this is the case of incidence estimates for different age strata. In other cases the correlations can be difficult to estimate. We can then treat them as independent when writing the pseudo-likelihood but not in the sandwich estimator, an approach similar to the GEE (Liang and Zeger, 1986). In principle we could use this approach for all the studies but we would need a large number of studies in order to get reliable estimates of the variances. In this paper we will examine whether the contributions of the same study can be considered as independent and the $L_k, k = 1, \ldots, K$ will be the independent contributions to the pseudo-loglikelihood.

3 A degradation model for myocardial infarction and CHD

3.1 The degradation model; modeling the drift

The atheromatous process $A(t)$ can be modeled as a Brownian motion with drift defined by the stochastic differential equation: $\mathrm{d}A(t) = \lambda \mathrm{d}t + \mathrm{d}B_{_A}(t)$, where B_A is a Brownian motion and λ is the drift. The time parameter was taken as age (in years) minus 20 and we take $A(0) = 0$. A basic degradation model is that MI happens when the atheromatous process reaches a certain threshold η, so that the time T at which MI occurs is the first hitting time of $A(t)$; for fixed η, T has an inverse Gaussian ($\mathcal{I}G$) distribution with parameters $(\eta/\lambda, \eta^2)$; its density is:

$$f(t)_{(\lambda,\eta)} = \left[\frac{\eta^2}{2\pi t^3}\right]^{1/2} \exp\left(\frac{-\lambda^2\left(t-\frac{\eta}{\lambda}\right)^2}{2t}\right) \quad \mathbb{1}_{\{t\geq 0\}}.$$

What is interesting from an epidemiological point of view is to model the drift as a function of physiological conditions suspected to play a role in the atheromatous process. Here we will take into account four of them: obesity, represented by body mass index (BMI), lipid profile represented by low density lipid concentration (LDL), inflammation represented by C-reactive protein concentration (CRP) and blood pressure represented by systolic blood pressure (SBP); in this paper the risk factors are assumed constant. We assume a linear model for λ:

$$\lambda = \lambda_0 + \beta_{\mathrm{BMI}}\mathrm{BMI} + \beta_{\mathrm{LDL}}\mathrm{LDL} + \beta_{\mathrm{CRP}}\mathrm{CRP} + \beta_{\mathrm{SBP}}\mathrm{SBP} + \varepsilon_\lambda, \tag{5}$$

where λ_0 is a baseline drift and ε_λ has a normal distribution with zero expectation and may represent unmeasured risk factors. We will in fact use transforms of BMI, LDL, CRP and SBP in order to get an approximately multinormal distribution (see section 4.2). Conditional on the values of these factors and the random error term ε_λ, the time of occurrence of MI has an $\mathcal{I}G$ distribution.

However for our synthesis we wish to use results of studies which have not recorded all of these factors. Thus, we potentially need all the distributions of T obtained by conditioning or not on these factors, and for computing them, all the corresponding distributions of λ. We assume that the indicators (BMI, LDL, CRP, SBP) have a multinormal distribution; equation (5) gives us the conditional distribution of λ given (BMI, LDL, CRP, SBP). From this we can compute analytically

the marginal distribution of λ (which is normal), as well as the distribution of λ conditional on BMI for instance (but marginal on LDL, CRP and SBP). Then the marginal distributions of T are inverse Gaussian Normal ($\mathcal{IGN}$) (Whitmore, 1986). If the drift parameter λ has a $\mathcal{N}(m_\lambda, s_\lambda^2)$ distribution, then the hitting time T has the distribution $\mathcal{IGN}(m_\lambda, s_\lambda^2, \eta)$, with density:

$$f(t)_{(m_\lambda, s_\lambda^2, \eta)} = \left[\frac{\eta^2}{2\pi t^3 (1 + s_\lambda^2 t)}\right]^{1/2} \exp\left(\frac{-(\eta - m_\lambda t)^2}{2t(1 + s_\lambda^2 t)}\right) \quad \mathbb{1}_{\{t \geq 0\}} \qquad (6)$$

This is an improper distribution in that $P(T = \infty) > 0$; this is not a problem in our model since not everybody develops a MI.

The model can be extended for defining two hitting times. It often happens that the progression of the atheromatous process first produces symptoms related to hypoxia (like angina pectoris) before the completion of MI; CHD includes these symptoms as well as MI. Thus two thresholds that we will denote η_{CHD} and η_{MI} can be defined and determine the distribution of the time of occurrence of CHD, T_{CHD}, and of MI, T_{MI}. This is illustrated in Figure 1.

A graph of the causal influences between processes as suggested in Commenges and Gégout-Petit (2009) is represented in Figure 2.

3.2 Modeling drift and threshold

Another model arises if the threshold rather than the drift varies with the value of an explanatory variable. Several authors (Aalen et al, 2008; Sæbø et al, 2005; Pennell et al, 2010) have used a degradation model where both drift and starting point may depend on covariates; these authors have also introduced random effects for these parameters and Whitmore (1986) considered random drift and variance. All these authors parameterized the model in term of starting point with fixed threshold, while we assume that the starting point is 0 and our parameter is the threshold. This is just two different parameterizations but with different interpretations. Parameterizing in term of starting point is adapted if the subjects start at different levels of the pathological process. In our application it is natural to consider that all subjects start with a very low level of atherosclerosis, thus $A(0) = 0$, and to parameterize in term of threshold. For instance it may be asked whether SBP has a cumulative effect represented by an effect on the drift, or an

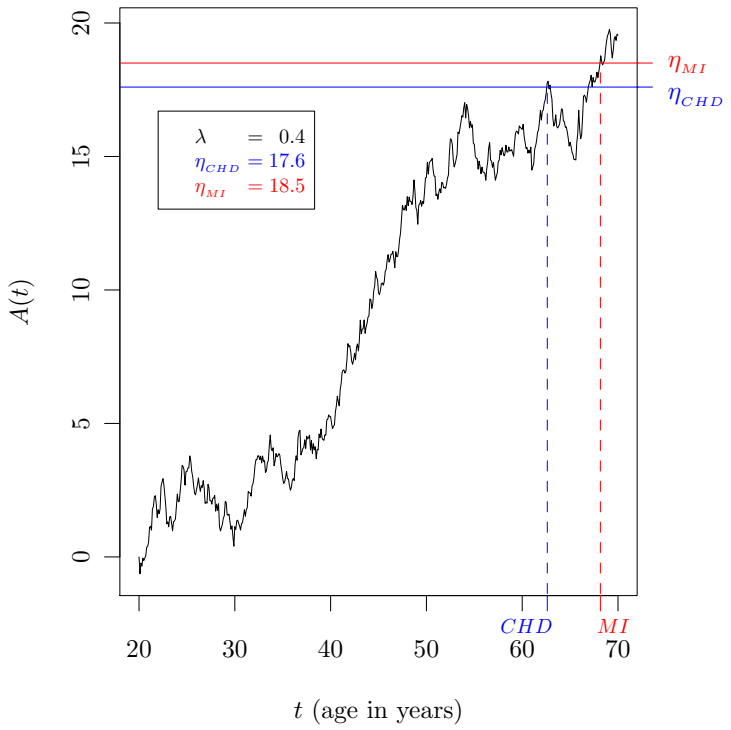

Fig. 1 Example of the trajectory of the atheromatous process $A(t)$. η_{CHD} and η_{MI} are the two thresholds; here, the subject has CHD symptoms at 62 and develops an infarctus at 68.

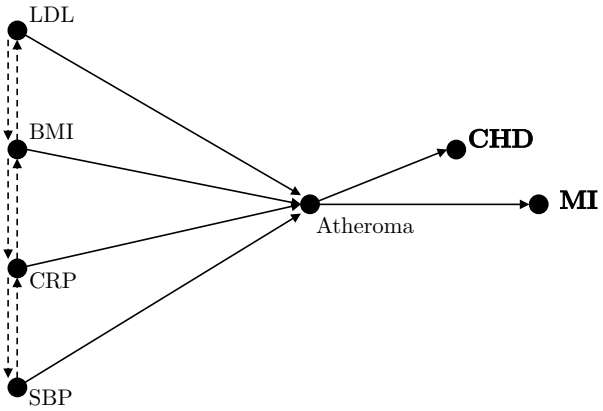

Fig. 2 Causal graph linking physiological conditions, atheromatous process, CHD and MI. Plain arrows mean purported causal influences and dashed arrows mean possible influences which have not been modeled in a mechanistic way and which result in correlations; we have not shown all the dashed arrows between LDL, BMI, CRP and SBP.

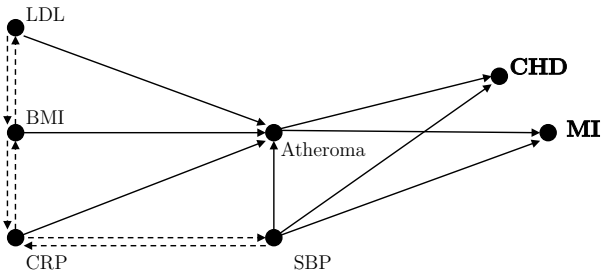

Fig. 3 Causal graph linking physiological conditions, atheromatous process, CHD and MI. Here SBP influences both the atheromatous process (effect on the drift) and directly CHD and MI (effect on the threshold). Plain arrows mean purported causal influences and dashed arrows mean possible influences which have not been modeled in a mechanistic way and which result in correlations; we have not shown all the dashed arrows between LDL, BMI, CRP and SBP.

instantaneous effect represented by an effect on the threshold, or both. SBP could favor the occurrence of MI by increasing the drift of the atheromatous process and also for a given state of the atheromatous process by lowering the threshold. The model for the threshold could be: $\eta = \eta_0 + \beta'_{\text{SBP}}\text{SBP} + \varepsilon_\eta$. The graph of causal influences would then be as in Figure 3. Conditional on the explanatory variables and ε_λ and ε_η, the distribution of T is still $\mathcal{I}G$. As before, for the synthesis analysis we need marginal distributions for which both λ and η are normal and may be correlated. We call the resulting distribution inverse Gaussian normal-normal ($\mathcal{IGNN}$), which to the best of our knowledge, has not been previously described. It happens that the density of this distribution has an analytic form, given in Appendix A. This distribution may also be useful in Bayesian computations if we put normal priors on baseline drift and thresholds.

4 Application to five studies

4.1 Studies giving information on CHD and MI risks

We have used five large studies. The National Health and Nutrition Examination Survey (NHANES) has been used to estimate the joint distribution of the physiological indicators in the population (see section 4.2). Four studies give information

Table 1 The studies used for the synthesis: NHANES: National Health and Nutrition Examination Survey; ARIC: Atherosclerosis Risk In Communities; PHS: Physicians Health Study; HPFS: Health Professionals Follow-up Study; FHS: Framingham Heart Study; # Contributions gives the number of independent contributions to the pseudo-loglikelihood, and in parenthesis the number of elementary contributions.

Study	**Nature of $\widetilde{Q}_k$**	**Event**	**Risk factors**	**# Contributions**
NHANES	Correlations	–	BMI, LDL, CRP, SBP	1 (6)
ARIC	Incidences	MI and CHD	–	4 (8)
PHS	Relative risks	MI	CRP	3 (3)
HPFS	Absolute risks	CHD	BMI	5 (5)
FHS	Absolute risks	CHD	BMI, LDL, SBP	1 (10)

on CHD and MI risks. Table 1 gives the list of the studies used together with the type of information they bring.

The ARIC (Atherosclerosis Risk In Communities) surveillance study gives estimates of the incidence of both MI and CHD for four age strata; see Table 2, taken from Tables 4.9 and 4.10 of National Heart Lung and Blood Institute (2006). The estimates of the MI incidences of different age strata can be considered as approximately independent; this is also the case for CHD incidences. We also make this approximation for incidence of MI and CHD for different age strata. This approximation however is not tenable for incidence estimates of MI and CHD for the same age stratum because the CHD cases observed include the MI cases. For stratum k we have $Y_{k,\mathrm{CHD}} = Y_{k,\mathrm{MI}} + Y_{k,\overline{\mathrm{MI}}}$, where $Y_{k,\mathrm{CHD}}$, $Y_{k,\mathrm{MI}}$ and $Y_{k,\overline{\mathrm{MI}}}$ are the observed numbers of CHD, MI and non-MI cases respectively. The incidences are estimated by $\tilde{Q}_{k,1} = \frac{Y_{k,\mathrm{MI}}}{a_{k1}}$ and $\tilde{Q}_{k,2} = \frac{Y_{k,\mathrm{CHD}}}{a_{k2}}$ respectively, where a_{k1} and a_{k2} are the respective numbers of person-years. Assuming independence between $Y_{k,\mathrm{MI}}$ and $Y_{k,\overline{\mathrm{MI}}}$ we have $\mathrm{cov}(Y_{k,\mathrm{CHD}}, Y_{k,\mathrm{MI}}) = \mathrm{var}(Y_{k,\mathrm{MI}})$. Assuming Poisson distributions for these numbers, simple computations show that we can estimate the covariance between $\tilde{Q}_{k,1}$ and $\tilde{Q}_{k,2}$ by $\frac{\tilde{Q}_{k,1}}{a_{k2}}$; the variances are estimated by $\frac{\tilde{Q}_{k,j}}{a_{kj}}, j = 1, 2$. Thus the ARIC study brings four independent contributions to the pseudo-loglikelihood (based on estimates of MI and CHD incidences for the four age strata) which are computed by formula (2): $L_k(\theta) = -\frac{1}{2}(\tilde{Q}_k - Q_k(\theta))^T \Omega_k^{-1}(\tilde{Q}_k - Q_k(\theta)), k = 1, 2, 3, 4$. The

Table 2 Incidence of CHD and MI in men in USA, ARIC Surveillance, 1987-2001.

Age	CHD sample	CHD Incidence	MI sample	MI Incidence
35-44	56,457	1.19‰	61,554	1.05‰
45-54	42,257	3.55‰	45,831	3.22‰
55-64	29,606	7.32‰	32,572	6.26‰
65-74	20,796	12.20‰	23,049	9.64‰

$Q_k(\theta) = (Q_{k,1}(\theta), Q_{k,2}(\theta))^T$ were computed by equation (3), using the $\mathcal{IGN}$ (or $\mathcal{IGNN}$) distributions of T_{CHD} and T_{MI}. For instance in formula (3) we have to compute $F_\theta(65)$, the value at 65 of the marginal cumulative distribution of the time of occurrence of MI.

The Physicians Health Study (PHS) is a randomized study of the prevention of cardiovascular diseases based on 22,071 subjects followed-up between 1982 and 1995. We use a case-control study nested in the PHS. Among these subjects, 543 men who developed a MI during the follow-up and 543 controls were chosen for studying the effect of CRP (Ridker et al, 1997). The study gives estimates of the relative risks for CRP strata with respect to a reference stratum (Table 3); it is not clear whether these are relative risks or odds-ratios but the two computations yield nearly the same result. The only information about age is the mean, which is 59. For simplifying the computations we attribute this age to all the subjects. The follow-up period was 13 years, thus we computed the risk for the age-period $(59 - 72)$. The relative risk of the CRP stratum]0.055;0.115[with respect to the stratum]0;0.055] was computed as :

$$RR^1_{\text{CRP}} = \frac{\left(F_{\theta|\text{CRP}\in]0.055;0.115[}(72) - F_{\theta|\text{CRP}\in]0.055;0.115[}(59)\right)\left(1 - F_{\theta|\text{CRP}\in[0;0.055]}(59)\right)}{\left(F_{\theta|\text{CRP}\in[0;0.055]}(72) - F_{\theta|\text{CRP}\in[0;0.055]}(59)\right)\left(1 - F_{\theta|\text{CRP}\in]0.055;0.115[}(59)\right)},$$

where $F_{\theta|\text{CRP}\in]a;b[}(.)$ is the cumulative distribution function of the T_{MI} conditional on CRP $\in]a;b[$. To avoid the numerical integration required for the computation of $F_{\theta|\text{CRP}\in[a;b]}(t)$, we approximated it by $F_{\theta|\text{CRP}=c}(t)$ where c is the median value of CRP in the sub-sample $[a,b]$ of the NHANES study. The standard deviations σ_k were computed from the confidence intervals given in the publication. In Ridker et al (1997) the confidence intervals have been derived from a normal approximation of the distribution of the regression coefficient. Thus the $\tilde{Q}_k$ were the estimated values of this coefficients and the standard deviations σ_k were deduced from the

Table 3 Relative risk of MI in men according to the quartile of CRP concentration, PHS, 1982-1995.

	Quartile of *CRP* concentration ($mg \cdot dL^{-1}$)			
	$\leq$ **0.055**	**0.056 − 0.115**	**0.116 − 0.210**	$\geq$ **0.211**
Relative risk	1.0	1.7	2.6	2.9
95% CI	–	1.1-2.9	1.6-4.3	1.8-4.6

Table 4 CHD risk in men in USA according to BMI strata, HPFS, 1988-2004.

	BMI				
	$<$ **23**	**23 − 24.9**	**25 − 26.9**	**27 − 29.9**	$\geq$ **30**
Number of events	349	625	821	643	333
Number of subjects at risk	7,669	12,104	11 466	7,712	3,400

confidence intervals of the relative risks as $\sigma_k = (log(\hat{R}R^k_{\text{CRP}}) - log(RR^k_{inf}))/1.96$ where RR^k_{inf} is the lower bound of the confidence interval for RR^k_{CRP} given in Ridker et al (1997). Approximate computation of the correlations between the estimates of the three relative risks gives values lower than 0.01 and thus we consider that the PHS study gives three independent contributions $L_5(\theta), L_6(\theta), L_7(\theta)$ of the form given by formula (1).

The Health Professionals Follow-up Study (HPFS) is a prospective cohort study bearing on 42,351 subjects. Flint et al (2010) have estimated the risk of CHD on a 12-year period as a function of the BMI at inclusion; see Table 4. The absolute risk for the 12 year follow-up period for subjects presenting a BMI lower than 23 was computed assuming that all subjects were 53 years old at inclusion (the mean age at inclusion) as:

$$Q_8(\theta) = \frac{F_{\theta|\text{BMI}\in[0;23[}(65) - F_{\theta|\text{BMI}\in[0;23[}(53)}{1 - F_{\theta|\text{BMI}\in[0;23[}(53)}. \tag{7}$$

Here, $F_{\theta|\text{BMI}\in[a;b[}(.)$ is the cumulative distribution function of T_{CHD} given $\text{BMI} \in [a;b[$. The variance was estimated by $\frac{\tilde{Q}_k}{a_k}$. Similar formulas were applied for the other BMI strata. The estimates for the five BMI strata are approximately independent, so the HPFS study yielded five independent contributions $L_k(\theta), k = 8, 12$.

The Framingham Heart Study (FHS) is a cohort study especially designed for cardiovascular epidemiology. Kannel et al (2002) have estimated the effect of BMI

Table 5 Risk of CHD in US men according to BMI strata, FHS, 1971-1987.

	BMI		
	$\mathbf{< 25}$	$\mathbf{25 - 29.9}$	$\mathbf{\geq 30}$
Number of events	99	188	72
Number of subjects at risk	772	1,306	375

Table 6 CHD risk in US men according to LDL strata, FHS, 1971-1983.

	LDL $(mg \cdot dL^{-1})$		
	$\mathbf{< 130}$	$\mathbf{130 - 159}$	$\mathbf{\geq 160}$
Number of events	104	124	155
Number of subjects at risk	929	866	719

Table 7 CHD risk in US men according to SBP strata, FHS, 1971-1983.

	SBP $(mm\ Hg)$			
	$\mathbf{< 130}$	$\mathbf{130 - 139}$	$\mathbf{140 - 159}$	$\mathbf{\geq 160}$
Number of events	110	77	115	81
Number of subjects at risk	1127	526	556	303

on CHD risk and give the number of CHD events over a 16 years period and the number of persons at risk. This allowed us to build Table 5. From the results of Wilson et al (1998) we can construct the same kind of tables for LDL and SBP: see Tables 6 and 7. The estimators of absolute risks for the different strata of BMI, LDL and SBP were computed by the same type of formula as (7) presented for the PHS study. These estimators however can not be considered as independent because they rely on observations of the same subjects; thus we can treat them in the GEE spirit, considering that for the computation of the sandwich estimator there is only one independent contribution $L_{13}(\theta)$ (the sum of the loglikelihood contributions for the 10 observations).

4.2 Construction of normal physiological indicators

The NHANES study is a repeated transversal surveys study recording a large number of physiological and behavioral factors on large representative samples of the

American population. We used the 2007-2008 survey giving observations of BMI, CRP, LDL and SBP on $n = 2871$ men aged 20 or older. To make the computation of the marginal distributions of the event times feasible we need an underlying normal distribution for the physiological factors. Inspection of the empirical distribution of these variables reveals that they do not have normal distributions. In order to construct indicators with a distribution close to the normal we used Box-Cox transformations (Box and Cox, 1964) defined as $\frac{X^\kappa - 1}{\kappa}$ $\kappa \neq 0$ and $\ln(X)$ if $\kappa = 0$. The parameter κ can be estimated by maximum likelihood (Velilla, 1993; Yeo and Johnson, 2000). This is implemented in the R package CAR (Fox and Weisberg, 2010). We found the optimal values of κ to be -0.58, 0.41, -0.04, -1.27 for BMI, LDL, CRP and SPB respectively.

As shown in Figure 4 the optimal Box-Cox transformations yield new variables with a distribution much closer to the normal than the original ones. Of course marginal normal distributions do not mean that we have a multinormal distribution but we expect to get closer to it (Kowalski, 1970). We applied a further transformation to standardize the indicators (zero mean and unit variance), so as to compare more easily the effects of these physiological factors.

From this study we could estimate the covariance (or correlation) matrix of our transformed variables:

$$corr(\text{LDL}, \text{BMI}, \text{CRP}, \text{SBP}) = \begin{pmatrix} 1.00 & 0.07 & 0.09 & -0.05 \\ 0.07 & 1.00 & 0.34 & 0.10 \\ 0.09 & 0.34 & 1.00 & 0.11 \\ -0.05 & 0.10 & 0.11 & 1.00 \end{pmatrix}$$

We consider the empirical correlation as observations. To get approximate normal distribution we used Fisher transformation. For the correlation between LDL and BMI for instance, we considered $\tilde{Q}_{14,1} = \text{arctanh}[\text{corr}(\text{LDL}, \text{BMI})]$ as normal with mean $\text{arctanh}(\rho_{\text{LDL,BMI}}(\theta))$ (where $\rho_{\text{LDL,BMI}}(\theta)$ is computed from the model) and variance $\frac{1}{n-3}$. However the $\tilde{Q}_{14,j}, j = 1, \ldots, 6$ may be correlated in a way which is difficult to analyze. Thus we consider that the NHANES study gives just one independent contribution to the pseudo-loglikelihood $L_{14} = \sum_{j=1}^{6} \frac{n-3}{2}(\tilde{Q}_{14,j} - Q_{14,j}(\theta))^2$.

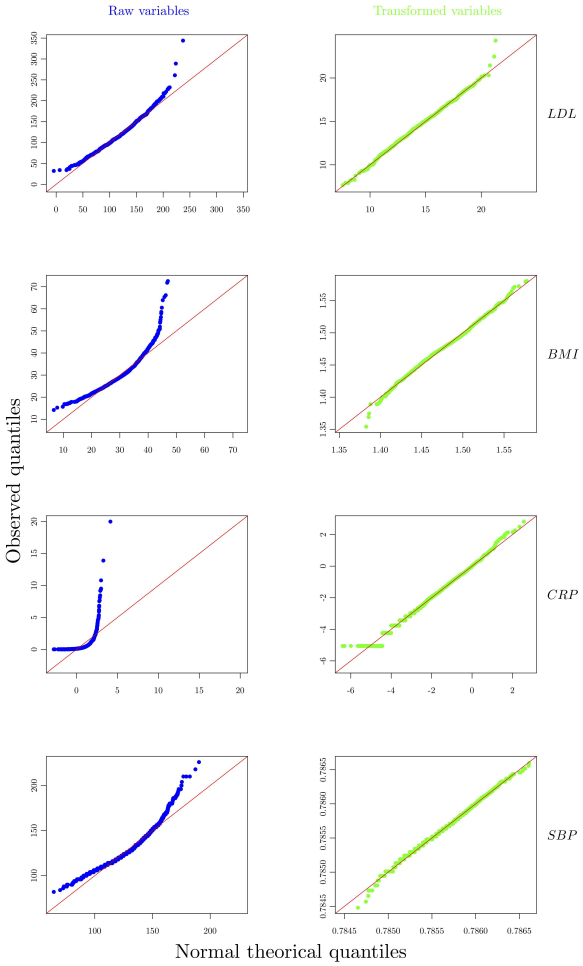

Fig. 4 Quantile-quantile plot of raw and transformed physiological variables. Left: original indicators; right: after Box-Cox transformation.

4.3 Results

Finally, the global pseudo-loglikelihood was the sum of 14 independent contributions : $Ł(\theta) = \sum_{k=1}^{14} L_k(\theta)$. The pseudo-loglikelihood was computed using a Fortran program and was maximized using a Marquardt algorithm (Marquardt, 1963) (also programmed in Fortran). The algorithm converged in 29 iterations (it was verified that different starting points led to the same convergence point). Table 8 gives the fit of the model for the information of the studies about MI and

CHD. We do not show the fits for the correlations of the physiological factors in the NHANES study: they were very good and the contribution was $L_{14} = -0.003$. For the other quantities, we consider that the fit is rather satisfactory given the heterogeneity of the studies. For instance the risk observed in the HPFS study for the 27–30 BMI stratum is around 8%; it is to be compared with the risk computed with the model of developing CHD between age 53 and 65 for this BMI stratum, which is around 10%: we get the good order of magnitude. Of course this is not a good fit in view of the small standard deviation attributed to the observed risk; if the model was well specified $-L_k$ would have a chi-squared distribution with one degree of freedom: the value of 17.5 is much too large. We know from the beginnning that our model cannot be well specified and this is why we used the term "pseudo-loglikelihood" rather than "loglikelihood".

Table 9 gives the parameter estimates together with sandwich estimates of their standard deviations; the estimate of the standard deviation of ε_λ was very close to zero and is not shown in the table. We note that the four regression parameters are positive, meaning that increasing the level of these factors increases the risk, which is consistent. Since the values are greater than 2 times their standard deviations, they are significant. There is a positive baseline drift and, as expected $\eta_{\text{CHD}} < \eta_{\text{MI}}$. From the sandwich covariance matrix we can compute the standard deviation of $\hat{\eta}_{\text{CHD}} - \hat{\eta}_{\text{MI}}$. We found 0.051 so that $\hat{\eta}_{\text{CHD}} - \hat{\eta}_{\text{MI}}$ is significantly different from zero. Figure 5 displays the cumulative distribution of the times of occurrence of MI (probability to have MI before age t) for three profiles: (i) mean level for all risk factors: $\lambda = \lambda_0 = 0.99$, (ii) above the mean level by one standard deviation for all factors: $\lambda = \lambda_0 + \beta_{\text{LDL}} + \beta_{\text{BMI}} + \beta_{\text{CRP}} + \beta_{\text{SBP}} = 0.354$, (iii) below the mean level by one standard deviation for all factors: $\lambda = \lambda_0 - \beta_{\text{LDL}} - \beta_{\text{BMI}} - \beta_{\text{CRP}} - \beta_{\text{SBP}} = -0.155$. We see that the latter profile has very low risk to develop MI during lifespan while profile (i) has high risk, for instance a probability of 0.80 to develop MI before 80.

Finally we have fitted the model described in section 3.2, where SBP could modify both thresholds (we took the same regression coefficient for both). The pseudo-loglikelihood for this model was -212.56. This is not significantly better than that of the basic model (without influence on the thresholds) which was -212.79.

Table 8 Fit of the information from the four studies. RR^1_{CRP}, RR^2_{CRP}, RR^3_{CRP} are the relative risks for the CRP strata 0.56-1.14, 1.15-2.10, ≥ 2.11 respectively, versus the reference stratum CRP ≤ 0.55, and AR stands for absolute risk. Last column: $-L_k$, where $L_k, k = 1, \ldots 13$ are the first 13 independent contributions to the pseudo-loglikelihood. For ARIC the contributions comes from the bivariate observations of incidences of CHD and MI in age strata. For FHS the global contribution is given. The 14th contribution brought by the correlations is not shown.

Study	**Event**	**Nature of $\widetilde{Q}$**	$Q(\widehat{\theta})$	$\widetilde{Q}$	σ	$-L_k$
ARIC	CHD	Incidence 35-44	0.0013	0.0012	0.00015	
ARIC	MI	Incidence 35-44	0.0011	0.0010	0.00013	3.5
ARIC	CHD	Incidence 45-54	0.0045	0.0035	0.00029	
ARIC	MI	Incidence 45-54	0.0041	0.0032	0.00026	6.2
ARIC	CHD	Incidence 55-64	0.0070	0.0073	0.00050	
ARIC	MI	Incidence 55-64	0.0063	0.0063	0.00044	14.9
ARIC	CHD	Incidence 65-74	0.0080	0.0122	0.00077	
ARIC	MI	Incidence 65-74	0.0077	0.0096	0.00065	32.6
PHS	MI	$log\big(RR^1_{\mathrm{CRP}}\big)$	0.47	0.53	0.27	0.0
PHS	MI	$log\big(RR^2_{\mathrm{CRP}}\big)$	0.77	0.96	0.26	0.3
PHS	MI	$log\big(RR^3_{\mathrm{CRP}}\big)$	1.17	1.06	0.24	0.1
HPFS	CHD	$AR_{\mathrm{BMI}}(< 23)$	0.034	0.046	0.0024	10.6
HPFS	CHD	AR_{BMI}(23-25)	0.056	0.052	0.0021	2.6
HPFS	CHD	AR_{BMI}(25-27)	0.077	0.072	0.0025	2.2
HPFS	CHD	AR_{BMI}(27-30)	0.103	0.083	0.0033	17.5
HPFS	CHD	$AR_{\mathrm{BMI}}(\geq 30)$	0.159	0.189	0.0074	7.9
FHS	CHD	$AR_{\mathrm{BMI}}(< 25)$	0.042	0.128	0.013	
FHS	CHD	AR_{BMI}(25-30)	0.080	0.144	0.010	
FHS	CHD	$AR_{\mathrm{BMI}}(\geq 30)$	0.136	0.192	0.023	
FHS	CHD	$AR_{\mathrm{LDL}}(< 130)$	0.037	0.112	0.011	
FHS	CHD	AR_{LDL}(130-159)	0.100	0.143	0.013	
FHS	CHD	$AR_{\mathrm{LDL}}(\geq 160)$	0.174	0.216	0.017	
FHS	CHD	$AR_{\mathrm{SBP}}(< 130)$	0.035	0.098	0.0093	
FHS	CHD	AR_{SBP}(130-139)	0.087	0.146	0.0167	
FHS	CHD	AR_{SBP}(140-160)	0.128	0.207	0.0172	
FHS	CHD	$AR_{\mathrm{SBP}}(\geq 160)$	0.212	0.267	0.0297	114.4

Table 9 Parameter estimates: $\rho_{X,Y}$ is the correlation between X and Y, the β's are the regression coefficients of the standardized transformed physiological factors, λ_0 is the baseline drift, the η's are the thresholds.

	Estimate	**Sandwich SD**
$\boldsymbol{\rho}_{\text{BMI,LDL}}$	0.070	0.0015
$\boldsymbol{\rho}_{\text{BMI,CRP}}$	0.343	0.0003
$\boldsymbol{\rho}_{\text{BMI,SBP}}$	0.096	0.0014
$\boldsymbol{\rho}_{\text{LDL,CRP}}$	0.093	0.0007
$\boldsymbol{\rho}_{\text{LDL,SBP}}$	-0.045	0.0024
$\boldsymbol{\rho}_{\text{CRP,SBP}}$	0.107	0.0006
$\boldsymbol{\beta}_{\text{BMI}}$	0.053	0.014
$\boldsymbol{\beta}_{\text{LDL}}$	0.092	0.015
$\boldsymbol{\beta}_{\text{CRP}}$	0.024	0.006
$\boldsymbol{\beta}_{\text{SBP}}$	0.087	0.015
$\boldsymbol{\lambda_0}$	0.099	0.008
$\boldsymbol{\eta}_{\text{MI}}$	17.50	0.81
$\boldsymbol{\eta}_{\text{CHD}}$	17.06	0.79

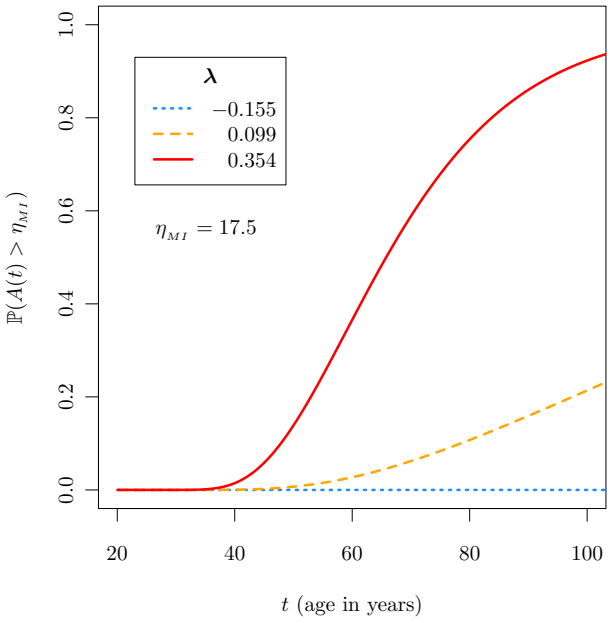

Fig. 5 Probability of developing MI as a function of age for three risk factors profiles: $\lambda = 0.099$: all risk factors at their mean value; $\lambda = -0.155$: all risk factors at one standard deviation below the mean; $\lambda = 0.354$: all risk factors at one standard deviation above the mean.

5 Conclusion

We have shown that a complex degradation model could be used for evidence synthesis of disparate results coming from different studies. The results are consistent in that we found a positive baseline drift, positive effects of the risk factors, positive values of the thresholds with higher threshold for MI than for CHD. We could have restricted the parameter space to impose the constraint $\eta_{\text{MI}} > \eta_{\text{CHD}}$. However at the maximum of the likelihood, either we would find $\hat{\eta}_{\text{MI}} > \hat{\eta}_{\text{CHD}}$ or $\hat{\eta}_{\text{MI}} = \hat{\eta}_{\text{CHD}}$. In the former case the constraint is inactive so that the result is the same whether we impose the constraint or not; in the latter case the constraint is active and this would throw suspicion on the model adequacy because without the constraint we would have $\hat{\eta}_{\text{MI}} < \hat{\eta}_{\text{CHD}}$.

We are limited in the epidemiological and clinical interpretation of these results because of the small number of studies that we have included. We have made the assumption of a linear relationship between λ and the risk factors. This is necessary to keep the computations relatively simple. A non-linear relationship would entail a non-normal joint distribution and we would lose the analytical marginal or partially conditional distributions. This is feasible but at the price of heavier computations. Also we have not included tobacco consumption; this factor raises a methodological problem due to the difficulty of finding a normalizing distribution. As smokers and non-smokers form a dichotomy, the situation calls for a two-component mixture of IG distributions rather than a normal mixture (IGN) over smoking status. Another potential problem is that we implicitly assume that death acts as an independent censoring. There may be selection effects due to deaths by diseases which share some common physiological process: for instance oxidative stress may play a role in both CHD and cancer.

The synthesis analysis can be viewed as an extension of meta-analysis. Indeed fixed-effect meta-analysis is a particular case when the $Q_k(\theta)$ are the same parameter for all k. Often meta-analyses put a random effect per study. This would also be possible in the synthesis approach and would certainly improve the fit, at the price of more computations.

This is a "proof-of-concept" paper, which shows that such an approach is very promising and could be developed in several directions. It could be applied to a

large number of studies of MI or CHD for producing reliable results, allowing to understand and to link the results of the literature on the subject. The evidence synthesis approach would become even more important if we moved toward life-course epidemiology for which information about the evolution of physiological factors would be incorporated, as suggested in (Commenges, 2012). This approach could also be applied to other topics as well.

Appendix A: The $\mathcal{IGNN}$ distribution

The inverse Gaussian normal-normal ($\mathcal{IGNN}$) distribution is an $\mathcal{IG}$ where parameters λ and η have a bivariate normal distribution with marginals $\mathcal{N}(m_\lambda, s_\lambda^2)$ and $\mathcal{N}(m_\eta, s_\eta^2)$ and correlation coefficient ρ. The probability density function of this distribution, although more complicated than that of $\mathcal{IGN}$, has an analytic form that we have derived with the help of Maple$^{\mathrm{TM}}$:

$$
\begin{aligned}
f(t)_{(m_\lambda, m_\eta, s_\lambda^2, \sigma_\eta^2, \rho)} = & -\frac{s_\eta \sqrt{1 + t s_\lambda^2 (1-\rho^2)}}{\pi \sqrt{t}\left(t^2 s_\lambda^2 - t(2 s_\lambda \rho s_\eta - 1) + s_\eta^2\right)} \; \mathrm{e}^{-\frac{\left(\mathrm{s}_\lambda^2 \mathrm{m}_\eta^2 + \mathrm{s}_\eta^2 \mathrm{m}_\lambda^2 - 2\mathrm{s}_\lambda \mathrm{m}_\eta \rho \mathrm{s}_\eta \mathrm{m}_\lambda\right) \mathrm{t} + \mathrm{m}_\eta^2}{2\left(1 + \mathrm{t}\mathrm{s}_\lambda^2 (1-\rho^2)\right) \mathrm{s}_\eta^2}} \\
& + \frac{\left((s_\lambda \rho s_\eta m_\lambda - s_\lambda^2 m_\eta) t + s_\lambda \rho s_\eta m_\eta - s_\eta^2 m_\lambda - m_\eta\right) \sqrt{2t^2 s_\lambda^2 + (2 - 4 s_\lambda \rho s_\eta) t + 2 s_\eta^2}}{2\sqrt{\pi}\left(t^2 s_\lambda^2 - t(2 s_\lambda \rho s_\eta - 1) + s_\eta^2\right)^2} \\
& \mathrm{e}^{-\frac{(\mathrm{t}\mathrm{m}_\lambda - \mathrm{m}_\eta)^2}{2\left(\mathrm{t}^2 \mathrm{s}_\lambda^2 - \mathrm{t}(2\mathrm{s}_\lambda \rho \mathrm{s}_\eta - 1) + \mathrm{s}_\eta^2\right)}} \; \mathrm{erf}\left(\frac{\sqrt{\mathrm{t}}\left((\mathrm{s}_\lambda \rho \mathrm{s}_\eta \mathrm{m}_\lambda - \mathrm{s}_\lambda^2 \mathrm{m}_\eta)\mathrm{t} + \mathrm{s}_\lambda \rho \mathrm{s}_\eta \mathrm{m}_\eta - \mathrm{s}_\eta^2 \mathrm{m}_\lambda - \mathrm{m}_\eta\right)}{\mathrm{s}_\eta \sqrt{1 + \mathrm{t}\mathrm{s}_\lambda^2 (1-\rho^2)} \sqrt{2\mathrm{t}^2 \mathrm{s}_\lambda^2 + (2 - 4\mathrm{s}_\lambda \rho \mathrm{s}_\eta)\mathrm{t} + 2\mathrm{s}_\eta^2}}\right)
\end{aligned}
$$

where erf is the error function ($\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x t^{-2} \mathrm{d}t$).

References

Aalen O, Borgan Ø, Gjessing H (2008) Survival and event history analysis: a process point of view. Springer Verlag

Aalen OO, Gjessing HkK (2001) Understanding the shape of the hazard rate: a process point of view. Statistical Science 16(1):1–22

Box GEP, Cox DR (1964) An Analysis of Transformations. Journal of the Royal Statistical Society: Series B (Methodological) 26(2):211–252

Clayton D, Schifflers E (1987) Models for temporal variation in cancer rates. i: age–period and age–cohort models. Statistics in medicine 6(4):449–467

Commenges D (2012) The stochastic system approach to causality with a view toward lifecourse epidemiology. Arxiv preprint arXiv:12035728

Commenges D, Gégout-Petit A (2009) A general dynamical statistical model with causal interpretation. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71(3):719–736

Doksum K, Normand S (1995) Gaussian models for degradation processes-part i: Methods for the analysis of biomarker data. Lifetime Data Analysis 1(2):131–144

Flint AJ, Hu FB, Glynn RJ, Caspard H, Manson JE, Willett WC, Rimm EB (2010) Excess weight and the risk of incident coronary heart disease among men and women. Obesity 18(2):377–383

Fosen J, Ferkingstad E, Borgan Ø, Aalen O (2006) Dynamic path analysis – a new approach to analyzing time-dependent covariates. Lifetime data analysis 12(2):143–167

Fox J, Weisberg S (2010) An R Companion to Applied Regression, 2nd edn. Sage Publications, Inc

Freedman D (2006) On the so-called huber sandwich estimator and robust standard errors. The American Statistician 60(4):299–302

Gamborg M, Jensen G, Sørensen T, Andersen P (2011) Dynamic path analysis in life-course epidemiology. American Journal of Epidemiology 173(10):1131

Hansson G (2005) Inflammation, atherosclerosis, and coronary artery disease. New England Journal of Medicine 352(16):1685–1695

Hashemi R, Jacqmin-Gadda H, Commenges D (2003) A latent process model for joint modeling of events and marker. Lifetime Data Analysis 9(4):331–343

Jiang W, Turnbull B (2004) The indirect method: inference based on intermediate statistics – a synthesis and examples. Statistical Science 19(2):239–263

Kannel WB, Wilson PWF, Nam BH, D'Agostino RB (2002) Risk stratification of obesity as a coronary risk factor. American Journal of Cardiology 90(7):697–701

Kowalski C (1970) The performance of some rough tests for bivariate normality before and after coordinate transformations to normality. Technometrics pp 517–544

Lee MLT, Whitmore GA (2006) Threshold Regression for Survival Analysis: Modeling Event Times by a Stochastic Process Reaching a Boundary. Statistical Science 21(4):501–513

Liang K, Zeger S (1986) Longitudinal data analysis using generalized linear models. Biometrika 73(1):13

Marquardt DW (1963) An algorithm for least-squares estimation of nonlinear parameters. Journal of the Society for Industrial and Applied Mathematics 11(2):431–441

National Heart Lung and Blood Institute (2006) Incidence and Prevalence : 2006 Chart Book on Cardiovascular and Lung Diseases. Tech. rep.

Nicholls SJ (2009) Relationship between LDL, HDL, blood pressure and atheroma progression in the coronaries. Current Opinion in Lipidology 20(6):491–496

Pennell M, Whitmore G, Lee M (2010) Bayesian random-effects threshold regression with application to survival data with nonproportional hazards. Biostatistics 11(1):111–126

Presanis AM, De Angelis D, Goubar A, Gill ON, Ades AE (2011) Bayesian evidence synthesis for a transmission dynamic model for HIV among men who have sex with men. Biostatistics 12(4):666–681

Raftery A, Givens G, Zeh J (1995) Inference from a deterministic population dynamics model for bowhead whales. Journal of the American Statistical Association pp 402–416

Ridker PM, Cushman M, Stampfer MJ, Tracy RP, Hennekens CH (1997) Inflammation, aspirin, and the risk of cardiovascular disease in apparently healthy men. New England Journal of Medicine 336(14):973–979

Sæbø S, Almøy T, Aastveit A (2005) Disease resistance modelled as first-passage times of genetically dependent stochastic processes. Journal of the Royal Statistical Society: Series C (Applied Statistics) 54(1):273–285

Schweder T, Hjort N (1996) Bayesian synthesis or likelihood synthesis–what does borel's paradox say? Report-Internationl Whaling Commission 46:475–480

Van der Vaart A (2000) Asymptotic statistics. 3, Cambridge Univ Pr

Van Houwelingen H, Arends L, Stijnen T (2004) Advanced methods in meta-analysis: multivariate approach and meta-regression. Tutorials in Biostatistics: Statistical modelling of complex medical data 2:289

Velilla S (1993) A note on the multivariate Box-Cox transformation to normality. Statistics & Probability Letters 17(4):259–263

Whitmore GA (1986) Normal-gamma mixtures of inverse Gaussian distributions. Scandinavian Journal of Statistics 13(3):211–220

Wilson PWF, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB (1998) Prediction of coronary heart disease using risk factor categories. Circulation 97(18):1837–1847

Yeo IK, Johnson RA (2000) A new family of power transformations to improve normality or symmetry. Biometrika 87(4):954–959