UNIVERSITY OF OSLO

HEALTH ECONOMICS
RESEARCH PROGRAMME

Absenteeism, Health Insurance, and Business Cycles

Morten Nordberg

Knut Røed
The Ragnar Frisch Centre for Economic Research and HERO

Working Paper 2003: 17

Absenteeism, Health Insurance, and Business Cycles

by Morten Nordberg and Knut Røed *

May 2003

Health Economics Research programme at the University of Oslo
HERO 2003

Keywords: Absenteeism, Dependent risks

JEL Classification: J22, C41

* This research is financed by the Norwegian Ministry of Labour and Government Administration and the Health Economics Research Program at the University of Oslo. We wish to thank Arne Kolstad and Snorre Kverndokk for valuable comments. Correspondence to: Knut Røed, the Ragnar Frisch Centre for Economic Research, Gaustadalléen 21, 0349 Oslo, Norway. E-mail: knut.roed@frisch.uio.no.

Health Economics Research programme at the University of Oslo
Financial support from The Research Council of Norway is acknowledged.
ISSN 1501-9071, ISBN 82-7756-123-7

Abstract

We use a dependent competing risks hazard rate model to investigate individual sickness absence behaviour in Norway, on the basis of register data covering more than 2 million absence spells. Our findings are: i) that business cycle improvements yield lower work-resumption rates for persons who are absent, and higher relapse rates for persons who have already resumed work; ii) that absence sometimes represents a health investment, in the sense that longer absence ‘now’ reduces the subsequent relapse propensity; and iii) that the work-resumption rate increases when sickness benefits are exhausted, but that work-resumptions at this point tend to be short-lived.

1 Introduction

The purpose of this paper is to evaluate how the economic environment affects worker absenteeism, with a focus on the design of the health insurance system and state of the business cycle (labour market tightness). We take advantage of a unique Norwegian database (The Frisch Centre Database) containing a complete account of absence spells due to sickness (lasting at least two weeks) in Norway during the 1992-1999 period. In total, the data contain more than two million absence spells. These data make it possible to establish duration specific work-resumption rates and subsequent relapse rates, and to investigate the extent to which these rates are related to labour market developments and to institutional aspects of the health insurance system.

Our methodological approach is based on a hazard rate framework. We set up a comprehensive (grouped) dependent competing risks hazard rate model, in which we not only seek to explain the first transition into either work-resumption or (more lasting) insurance schemes such as rehabilitation and disability; we also account for the subsequent relapse hazard and the potentiality of later (delayed) transitions into rehabilitation/disability. There is stochastic dependence between the various hazard rates, arising from the presence of unobserved heterogeneity; e.g. in the form that persons with unobserved characteristics that yield a low work-resumption hazard also tend to have unobserved characteristics that yield a high relapse hazard once work is actually resumed. We model the multivariate distribution of unobserved characteristics in a non-parametric fashion, in order to ensure a maximum degree of data-based (as opposed to assumption-based) inference.

Norway ranks among the countries of the world with the most generous sick-pay scheme. The fact that it also ranks among the countries with the highest fraction of working days lost due to illness (CESifo, 2002) illustrates a possible moral hazard dilemma. In the face of high replacement ratios, some workers may be tempted to register as sick even though they are able to work[1]. This kind of behaviour may take different forms, from outright shirking, to just a slight bias in the assessment of one's own health condition. There is an existing literature that evaluates the relationship between the level of sickness benefits and individual absenteeism. This literature demonstrates convincingly that sickness absence is not always a clear-cut response to a medical condition, but that the level of compensation does play an important role (see Barmby *et al*., 2002, for a recent overview). Given the rarity of exogenous variation in sickness benefits, the most recent contributions to the literature have tried to take advantage of institutional reforms that have imposed some kind of changes in the incentive structures. For example, Meyer *et al*. (1995) compare individuals injured before and after increases in the maximum benefit amount in two American states (with persons below the initial maximum serving as controls), and find that the elasticity of 'time out of work' with respect to the benefit level was around 0.3-0.4 in these cases. Johansson and Palme (2002) take advantage of the sickness insurance reform that was implemented in Sweden in 1991, which implied a dramatic decrease in replacement ratios, and also find significant disincentive effects.

In our own data, there is no independent variation in the sizes of sickness benefits, since virtually all workers receive a full replacement of their normal income. However, after 12 months of absence, this benefit is no longer available, and alternative benefits (such as rehabilitation or disability) provide much lower replacement; hence in cases of long-term absence, it is possible to investigate the consequences of

[1] The lack of sickness insurance may also be inefficient, since each work may not take into account the damage they can inflict on other workers (or customers) if they show up on the work-place with infectious diseases, or diseases that seriously reduce the quality of their work (Chatterji and Tilley, 2002).

quite substantial changes in economic incentives. Moreover, there are other more subtle sources of variation in the cost of being absent from work that are clearly exogenous from the viewpoint of individuals. In particular, we expect the cost of being absent to be inversely related to business cycle developments. The reason for this is that the cost of being absent (and in particular the cost of being caught 'shirking') is higher the more insecure a job is considered to be and the more difficult it is to get a new one[2] (Barmby *et al*., 1994). As a result, we expect sickness propensity to decline in bad times. The problem with confronting this idea directly with data is that observed absence behaviour may also be affected by labour market tightness through two other channels, which are not directly linked to economic incentives. First, labour market tightness may affect the degree of strain/stress felt by each worker, and thereby their actual medical condition. And second, labour market tightness may affect the selection of workers into employment, since the healthiest workers are likely to have the most stable attachment to the labour force. The latter of these effects is a pure selection mechanism, which does not require that individual absence behaviour is affected by business cycles at all.

Attempts to disentangle causality from selection in the observed pro-cyclical absence pattern in Norway have been based on two different scientific approaches. First, Askildsen *et al*. (2002) focus on the absence behaviour of individuals that are employed throughout the business cycle. They find that the positive correlation between sickness absence and business cycles remains, and conclude that the correlation

---

[2] A more direct approach to investigate the role of job insecurity is to compare absence behaviour for persons with secure (permanent) and insecure (temporary) jobs. The problem with this approach is of course that the selection into secure and insecure jobs is not random. Based on Swedish data, Arai and Thoursie (2001) try to solve this endogeneity problem by focusing on industry/region absence time-series aggregates, and regress these on the fraction of temporary employment. Again, the existence of a causal effect is confirmed.

pattern therefore cannot be entirely driven by selection. Second, Dyrstad and Ose (2002) separate causality and selection on the basis of the identifying assumption (for which they provide some theoretical justification) that selection effects are relatively more important when the rate of unemployment is low, while discipline effects are relatively more important when the rate of unemployment is high. They find that discipline effects do have a role to play. More surprisingly, they also find that these effects are particularly important for longer absence spells. This may suggest that persons with poor health, and hence a potential for having long periods of absence from work, are particularly afraid of losing their jobs in bad times.

In the present paper, we isolate the causal effects of business cycle developments on work-resumption prospects for ongoing absence spells, by conditioning on the state of the business cycle at the moment of entry into sickness absence. Conversely, by looking at the 'effects' on work-resumption prospects of business cycles at the moment of entry, conditioned on the business cycle at the time of potential work-resumption, we also (indirectly) shed light on the business cycle selection process. Our main finding is that there are substantial *causal* effects of business cycle developments on work-resumption rates. But, although we are able to eliminate all selection effects, we cannot without further assumptions decompose the causal relationship into discipline and 'work-pressure' effects. However, as we explain later on, we interpret the totality of our findings as indicative of all the mechanisms discussed above being present, selection effects, discipline effects and work-pressure effects.

The next section describes the data, and the economic environment in which the data were generated. Section 3 sets up the statistical model and discusses identification issues. Section 4 presents the results, and Section 5 summarises the main conclusions.

2 Data, Institutions and Business Cycles

The Norwegian sickness insurance system is generous. The replacement ratio is normally 100 per cent from the first day of absence[3]. But spells lasting more than three days (i.e. all spells analysed in the present paper) require a medical certificate. The maximum duration of sickness benefits is one year, and at least six months of work are required in order to be eligible for a new one-year-period. After that, a 100 per cent replacement ratio is no longer an option. The person must then either resume work, enrol into some kind of rehabilitation program, or apply for disability benefits. The latter option usually requires that rehabilitation has been tried first. Rehabilitation- and disability benefits typically provide replacement ratios between 50 and 60 per cent.

One possible effect of duration-limited sickness benefits is that the duration limit becomes a sort of focal point, both for the patient and the doctor, in the sense that the patient is 'allowed' to consider himself sick until the time of benefit exhaustion. In that case, we would observe an increase in the work resumption rate around the time of benefit exhaustion. If absence spells were unaffected by the insurance system, and only driven by medical factors, there would be no reason to expect such an increase.

The data we use in the present analysis are collected from Social Security registers, and contain monthly observations of all absence spells in Norway (lasting at least two weeks) during the 1992-1999 period for persons below 60 years. We trace each spell until it ends in either work resumption or in some alternative (and more

---

[3] This is paid by the employers the first two weeks, after which Social Security covers the cost up to a ceiling of around 40,000 Euro per year, while employers (in most cases) adds the amount necessary to ensure a 100 per cent replacement ratio.

lasting) form of Social Security benefit (rehabilitation or disability[4]). In the case of work-resumption, we trace the person for the next 12 months as well, in order to identify relapses into absence spells or subsequent (delayed) transitions to rehabilitation/disability. If a relapse occurs within the first six months after an absence spell was completed, the patient is not entitled to a new benefit period; hence from an administrative point of view, the new spell is a continuation of the previous spell. In the terminology used in this paper, such continuation spells belong to the same '*spell sequence*', together with the intervening work-resumption spells. Hence, a new spell sequence can only be started after at least six months of work without any recorded sickness. A spell sequence is terminated (censored) when one of the following six events occur: i) work has been resumed for an uninterrupted period of one year; ii) a new sickness spell is recorded after 6-12 months of work resumption; iii) a transition to rehabilitation or disability is recorded; iv) the person become 60 years of age; v) the person dies; or vi) the observation period has come to an end.

Table 1
Descriptive Statistics

	Men	Women
Number of individuals	439,509	567,739
Absence spells:		
Total number	865,016	1,246,347
Out of which ended in (per cent):		
Work resumption	90.66	91.73
Rehabilitation, disability or other benefits	5.52	4.58
Censored at the end of the observation period	3.81	3.68
Average duration (months, including censored spells)	3.00	3.05
Work resumption spells:		
Total number	784,245	1,138,898
Out of which ended in (per cent)		
Relapse into absence within 6 months	16.75	18.43
Relapse into absence after 6-12 months	12.62	13.18
Rehabilitation/disability within 6 months	2.06	2.06

[4] The reason why we do not distinguish between rehabilitation and disability is that most disability transitions are preceded by rehabilitation attempts. Indeed, Norwegian Social Security regulations imply that rehabilitation is to be considered before a disability pension can be provided. Since this may sometimes be a lengthy process, we are not able to identify the outcome of all the rehabilitation spells that we observe.

Table 1
Descriptive Statistics

Rehabilitation/disability after 6-12 months	0.33	0.33
Censored within 12 months	12.40	12.35
Resumption spell lasting at least 12 months	55.84	53.65
Total number of spell sequences	805,861	1,153,848
Fraction of individuals with multiple spell sequences	45.02	53.23

Table 1 provides some essential statistics regarding the data. A little more than 1 million individuals experienced at least one recorded absence spell during the whole eight-year period. Around 90 per cent of these spells ended in work-resumption[5]. But more than 30 per cent of these work-resumptions were short-lived, in the sense that a new absence spell started within the first year. The observation period was characterised by relatively large changes in labour market tightness. This is illustrated in Figure 1, where we have plotted two business cycle indicators; the deviation from the estimated trend in the quarterly GDP level (collected from Statistics Norway, see Johansen and Eika, 2000), and the estimated human-capital-adjusted monthly transition rate from unemployment to employment (collected from Gaure and Røed, 2003). Both series reveal a deep recession during 1992, after which a steady recovery took over. From the autumn of 1996 and two years onwards, Norway's economy boomed, after which a new slow-down emerged.

[5] Note that we do not actually observe work-resumption; rather we infer that work-resumption has occurred when sickness benefits are terminated without being replaced by other kinds of benefits. This induces an element of measurement error in our work-resumption indicator. We may erroneously infer that work-resumption has occurred in cases were workers pull out of the labour market without rehabilitation –or disability benefits. This is the main why we have restricted the analysis to persons below 60 years of age (and hence well below the age-floor in the most common early retirement schemes).

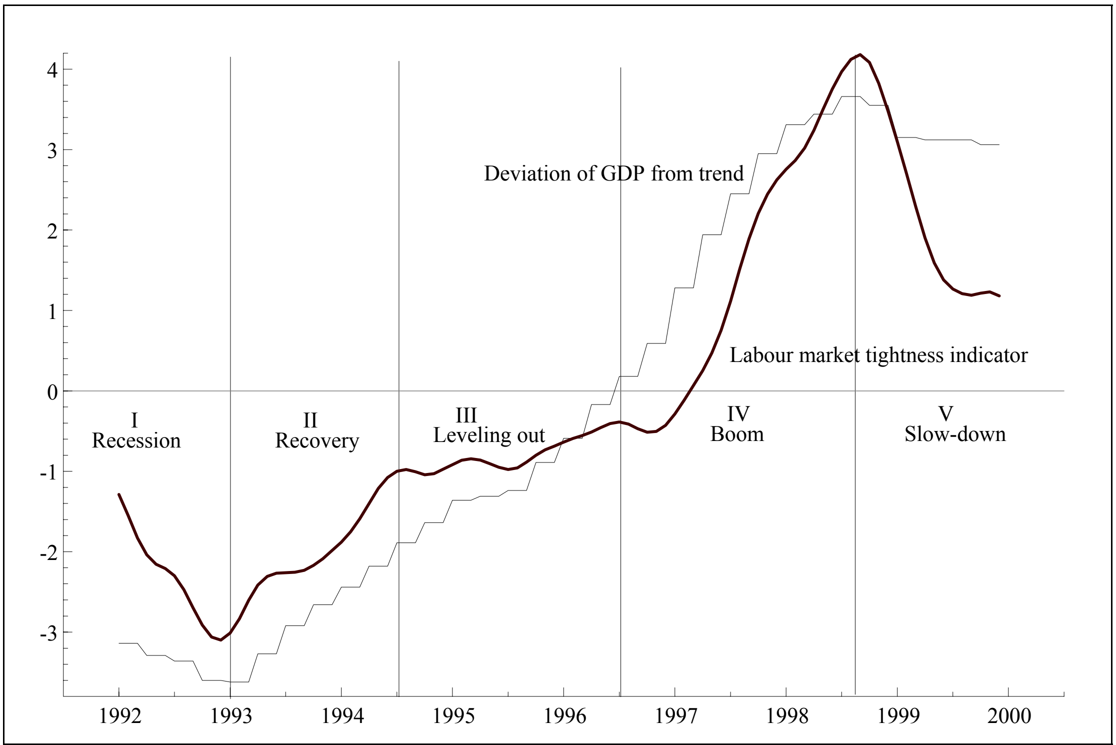

Figure 1. Business cycle developments in Norway 1992.1-1999.12

3 Identification and Estimation of Causal Effects

The statistical model we use is a dependent competing risks hazard rate model with two alternative origin states – absence (a) and work resumption (w) - and three alternative destination states - absence (a), work resumption (w) and rehabilitation/disability (r). The way we construct the dataset ensures that the first origin state in each spell sequence is sickness absence. Let i be the subscript over individuals, j the subscript over origin states, and k the subscript over destination states. Let d denote process time (i.e. spell duration) and let t denote calendar time. The four hazard rates are then defined as

$$\theta_{ijk}(t,d) = \lim_{\Delta d \to 0} \frac{P(d \leq D \leq d + \Delta d, K = k \mid D \geq d, i, t)}{\Delta d}, \tag{1}$$
$$j = a, w; \quad k = a, w, r; \quad j \neq k.$$

The hazard rates can approximately be interpreted as the probabilities of making a particular transition in a very short time interval starting at duration d and calendar time t, given that no transitions occurred before that point in time. Since we observe state occupation on a discrete time basis, we set up the statistical model in terms of discrete (grouped) hazard rates. Let $\overline{t}_{ij}$ be the calendar time at which individual i entered into the current state j. The grouped composite hazard, i.e. the probability of exiting to one of the two other states during duration period d, given that no exit occurred before that, is given as

(2) $$h_{ijdt} = 1 - \exp\left(-\sum_{k \neq j} \int_{d-1}^{d} \theta_{ijk}(\overline{t}_{ij} + u, u) du \right).$$

For absence spells, we use *months* as the unit of measurement, while for work-resumption spells we use *quarters*. We assume for simplicity that the hazard rates are constant within observed time intervals. We also assume that the hazard rates belong to the Mixed Proportional Hazards (MPH) family, i.e. they are proportional in factors depending on observed covariates (x_{it}), calendar time (t), spell duration (d), destination specific unobserved scalar variables (v_{ikt}) and (in the case of work resumption spells) the duration of the last absence spell ($\overline{d}_{ia}$). We can then write the integrated hazard functions $\left(\varphi_{ijktd} = \int_{d-1}^{d} \theta_{ijk}(\overline{t}_i + u, u) du \right)$ as

(3) $$\begin{aligned} \varphi_{iaktd} &= \exp(\beta_{ak}(x_{it}) + \sigma_{ak}(t) + \lambda_{ak}(d) + v_{ikt}),\ k = w, r, \\ \varphi_{iwktd} &= \exp(\beta_{wk}(x_{it}) + \sigma_{wk}(t) + \lambda_{wk}(d) + \gamma_{wk}(\overline{d}_{ia}) + v_{ikt}),\ k = a, r, \end{aligned}$$

where $\beta_{jk}(.)$, $\sigma_{jk}(.)$, $\lambda_{jk}(.)$ and $\gamma_{wk}(.)$ are functions which parameters are to be recovered from the data. In (3), the effects of unobserved heterogeneity vary freely over time for each individual. Without further restrictions it is of course not possible to identify the model. We assume that the unobserved heterogeneity terms can be decomposed into factors that depend on time only, and factors that depend on the identity of the individual (and remain fixed over time for each individual). More precisely, we assume that

(4) $$v_{ikt} = \delta_k(\bar{t}_{ia}) + v_{ik},$$

where $\bar{t}_{ia}$ is the calendar time of entry into the current spell sequence and $\delta_k(.)$ is a function. The rational behind letting the distribution of unobserved heterogeneity depend on the calendar time at entry into absence consists of three parts (conf. the introduction): First, business cycle developments may affect the extent to which a given medical condition is viewed by the individual as justifying an absence spell. Second, the business cycle may directly affect the medical condition of some people through its impact on the working pressure (strain and stress). And third, the business cycle may affect the composition of the labour force.

In order to estimate the model, we specify the functions of interest in the following way:

(5) $$\begin{aligned}
&\beta_{jk}(x_{it}) = x_{it}{'}\beta_{jk}, \\
&\sigma_{jk}(t) = \sigma_{1jk}bc_t + \sigma_{2jk}\Delta bc_t + \psi_m, \;\; m = Jan, Feb, .., Dec, \\
&\lambda_{jk}(d) = \lambda_{jkd}, \; d = 1,2,..,12, \\
&\gamma_{wk}(\bar{d}_{ia}) = \gamma_{wk\bar{d}_{ia}}, \; \bar{d}_{ia} = 1,2,..,12, \\
&\delta_{jk}(\bar{t}_{ia}) = \delta_{1jk}bc_{\bar{t}_{ia}} + \delta_{2jk}\Delta bc_{\bar{t}_{ia}},
\end{aligned}$$

where c_t is the value of the labour market tightness indicator[6] (see Figure 1) in period t and ψ_m is a season-parameter. Furthermore, we assume that individual heterogeneity and business cycles have the same proportional effects on the rehabilitation/disability hazard regardless of the initial state, i.e.

(6) $$\beta_{ar} = \beta_{wr}, \; \sigma_{1ar} = \sigma_{1wr}, \; \sigma_{2ar} = \sigma_{2wr}, \; \delta_{1ar} = \delta_{1wr}, \; \delta_{2ar} = \delta_{2wr}.$$

[6] We have also estimated the model with the GDP-based business cycle indicator depicted in Figure 1, but this did not lead to any substantial changes in the results.

According to this specification, there are no restrictions at all regarding the spell duration effects (or the effects of past sickness duration on the subsequent work resumption spell), since a separate parameter is attributed to each possible ongoing (or completed) duration. The calendar time effects, however, are restricted to depend linearly on the level- and difference of a business cycle indicator and on the season. The motivation for including both the level and the difference of the business cycle indicator as time varying covariates in the hazard rates(in the σ_{jk} terms) is that both these variables convey important information about the expected costs associated with remaining absent or relapsing into absence. Our prior expectation is that both variables have negative effects on the work-resumption hazard and positive effects on the relapse hazard. The reason for including the same two variables' values at the time of entry into absence in the selection function (in the δ_{jk} terms) is of a similar kind. If a low absence cost is associated with less serious illnesses being recorded as sickness, we expect that the two business cycle variables contribute to a positive selection effect. However, the *level* of the business cycle will also capture possible negative labour force composition effects, since the health status of the labour force may be poorer as more marginal workers are pulled into the labour market. Hence, we expect that the change in the business cycle contribute to a positive selection into the state of absence (in terms of higher work-resumption hazards and lower relapse hazards), but that the level of the cycle may have a positive or a negative effect, depending on the importance of labour force selection.

A common problem in duration analyses is to disentangle the effects of spell duration (duration dependence) from the influences of unobserved heterogeneity. The source of this difficulty is that persons that are, for example, only mildly hit by a disease will tend to resume work quickly, leaving behind persons with more serious diseases. And since the seriousness of the disease cannot be observed, this implies that – conditional on everything that can be observed – there tends to be a negative correlation between spell duration and work-resumption prospects even when there is no *causal* relationship at all between duration and work-resumption prospects. It has been proved, however, that the MPH model of the type we use is non-parametrically identified under some regularity assumptions, given that there is at least some cross-sectional variation in observed explanatory variables that affect the hazard rates in question (Elbers and Ridder, 1982; Heckman and Singer, 1984a). The intuition is that the 'selection consequences' associated with a given value of an unobserved covariate depends on the corresponding values of the observed covariates in a manner that can be traced out from the proportionality assumption. In our data, there are also two sources of information that can be used to disentangle true duration dependence from unobserved heterogeneity without reliance on proportionality. The first is the existence of *repeat spell sequences* for the same individuals. Since we assume the unobserved covariates v_{ik} to be constant for each individual, this introduces a powerful fixed-effect type source of model identification (Van den Berg, 2001). The second source of identification is the existence of *time varying* covariates (McCall, 1994; Brinch, 2000). The basic idea is that persons who according to observed covariates have experienced, say, a high probability of work-resumption earlier in their absence spell without actually resuming work, have revealed a low expected value of their unobserved work-resumption propensity, ceteris paribus. An important reservation regarding these sources of identification of unobserved heterogeneity is that they only identify heterogeneity that has a constant effect at the individual level on the various hazard rates. We cannot identify unobserved heterogeneity in the degree of duration dependence itself, i.e. in the way e.g. recovery prospects change over spell duration.

Given that different types of health problems are associated, not only with different recovery prospects as such, but also with different recovery profiles over time, we suspect that the influences of unobserved heterogeneity cannot be entirely removed from our estimates of spell duration effects.

We now turn to the estimation of the model, based on observations of explanatory variables and transitions. We use a non-parametric approach to account for unobserved heterogeneity also, to make sure that the results are really driven by the data and not by unjustified restrictions. In practice, this implies that the unobserved variables v_{ik} are discretely distributed (Lindsay, 1983) with the number of mass-points chosen by adding points until it is no longer possible to increase the likelihood function (Heckman and Singer, 1984b). Let K_{it} be the set of feasible transition states for individual i at time t and let y_{itk} be an outcome indicator variable which is equal to 1 if the corresponding observation period ended in a transition to state k, and zero otherwise. Furthermore, let N_i be the set of observations available for individual i. The contribution to the likelihood function formed by a particular individual, conditional on the vector of unobserved variables $v_i = (v_{iw}, v_{ir}, v_{ia})$ can then be formulated as

$$(7) \qquad L_i(v_i) = \prod_{i \in N_i} \left[\begin{array}{l} \prod_{k \in K_{it}} \left[\left[\left(1 - \exp\left(-\sum_{k \in K_{itd}} \varphi_{ijktd} \right) \right) \dfrac{\varphi_{ijktd}}{\sum_{k \in K_{itd}} \varphi_{ijktd}} \right]^{y_{itk}} \right] \\ \times \left[\exp\left(-\sum_{k \in K_{it}} \varphi_{ijktd} \right) \right]^{\left(1 - \sum_{k \in K_{id}} y_{itk} \right)} \end{array} \right]$$

where $K_{it} = \{w, r\}$ when $j = a$ and $K_{it} = \{a, r\}$ when $j = w$, except in cases where work is resumed ($j = w$) after full exhaustion of sickness benefits ($\overline{d}_{ia} = 12$), in which case $K_{it} = \{r\}$ during the subsequent two quarters.

We assume that the unobserved heterogeneity vectors are discretely distributed with W points of support, and estimate these mass points together with their associated probabilities. In terms of observed variables, the likelihood is then given as

$$L = \prod_{i=1}^{N} \sum_{l=1}^{W} q_l \prod_{b=1}^{B_i} L_{ib}(v_l), \quad \sum_{l=1}^{W} q_l = 1 \tag{8}$$

where q_l is the probability of the particular combination of unobserved variables v_l.

Maximisation of (8) constitutes a huge computational task. Since each spell sequence is split into several observation periods (in order to avoid restrictions on the way spell duration and business cycles affects the hazard rates), each with a separate outcome indicator, the total number of observations is extremely large (12,418,185). Moreover, the likelihood function is not globally concave, and as more support points are added to the distribution of unobserved heterogeneity, it is well known that it becomes increasingly difficult to locate the global maximum. In order to solve these problems, we have utilised the 'implicit dummy approach', which takes into account the prior knowledge that many (in practice almost all) calculations involving the dummy variables are superfluous (since most of them are equal to zero most of the time); see Gaure and Røed (2003). The model was estimated on a supercomputer at the High Performance Computing Centre, University of Oslo. A complete list of the explanatory variables used in the model is provided in Appendix.

4 Results

We estimated separate models for men (5,061,300 observations) and women (7,356,885 observations). For men, we found that eight mass-points were required in the heterogeneity distribution in order to maximise the likelihood function, while for women it sufficed with six points. The total number of parameters to be estimated was 219 for men and 211 for women. Given this large number of parameters, we do not present complete estimation results, but focus on the parameters of interest. We first describe the main pattern of transition probabilities, as they evolve over duration of absence –and work-resumption spells (Section 4.1). We then turn to the causal (4.2) and the composition (4.3) effects of business cycles. Finally, we present results regarding the effects of observed and unobserved heterogeneity (4.4).

4.1 Overall Transition Rates and the Spell Duration Pattern

Figure 2 describes the expected progression of absence spells in terms of transition rates to work-resumption and rehabilitation/disability. The two upper panels plots the estimated duration parameters $(\hat{\lambda}_{jkd})$, together with their 95 per cent confidence intervals. The two confidence interval boundaries are not easy to spot, which highlights the point that statistical uncertainty is not a major issue in this analysis. Based on the point estimates, the two lower panels depict the implied monthly transition probabilities (or grouped hazard rates) for representative entrants into sickness absence[7]. The probability of resuming work during the course of the first month is 51 per cent for men and 47 per cent for women. It then declines monotonically until it reaches a trough after 9 months at around 12 per cent. Then, in the last month of the eligibility period, it again rises again to 30-35 per cent.

[7] Representative entrants are constructed by scaling the first month transition probabilities to the average observed transition rates in the first duration month.

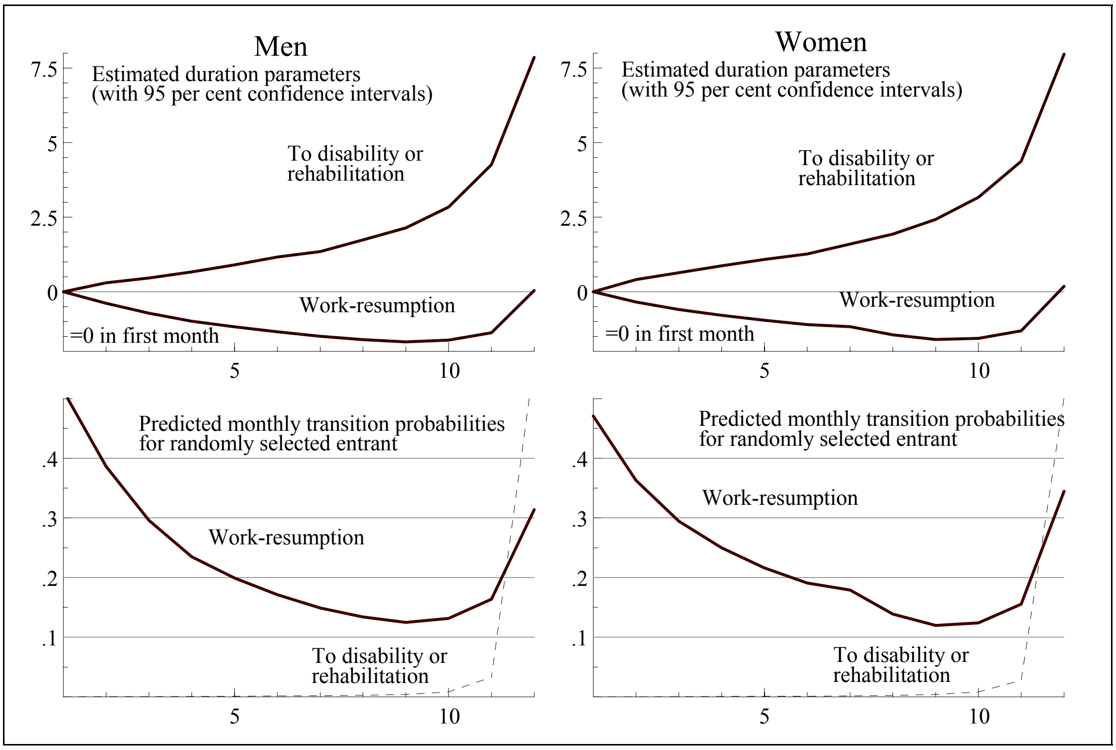

Figure 2. Estimated spell duration parameters $(\hat{\lambda}_{jkd})$ with 95 per cent point-wise confidence intervals (upper panels) and predicted monthly transition probabilities for representative entrant into absence (lower panels).

Note: There are confidence intervals around all the profiles in the upper panels, even they are not visible due to the thickness of the lines.

The work-resumption peak identified in the last month of the eligibility period coincides with the point of benefit exhaustion. It is hard to believe that this has nothing to do with the termination of sickness benefits. Yet, the result must be interpreted with care, since, at this point, benefit exhaustion forces some kind of transition. It may be the case that attempts to resume work are not really made, or that they fail, such that the start of rehabilitation or the take up of disability pension is only delayed for a short time period. We return to this issue below.

The probability of making a transition to disability or rehabilitation is (unsurprisingly) negligible during most of the absence spell. However, having reached the end of the eligibility period entails a very high probability (above 40 per cent for our representative individual) of making that transition. This reflects that persons who at this point are still not able to resume work for health reasons, are entitled to rehabilitation or disability transfers[8].

Figure 3 describes the progression of the subsequent work-resumption spells, conditional on the duration of the completed absence spell. Note that the time units in Figure 3 – as well as the predicted probabilities – are measured on a quarterly basis. For persons with short initial absence spells, the probability of making a subsequent transition to rehabilitation or disability is close to zero. The probability rises substantially if the sickness benefits were exhausted (or nearly exhausted). In particular, the probability of entering into rehabilitation or disability during the first quarter is around 35 per cent for persons who exhausted their entire sickness benefit entitlement. The probability of a relapse into sickness typically lies between 5 and 10 per cent for each quarter. And this probability is higher the shorter was the initial spell. In particular, it seems that very short spells (only one month) is associated with a relatively high relapse probability. For example, if the initial absence spell lasted three months, rather than just one, the relapse hazard is reduced by 6.6% (0.8%) for men and as much as 32.4% (0.7%) for women (standard errors in parentheses). The latter finding suggests that absence spells to some extent may be viewed as a health investment, in the sense that having a longer absence spell 'now' reduces the probability of a relapse in the future.

[8] Note that there is not a 100 per cent probability of escaping sickness benefits in the presumed final month of the benefit period. This is related to the fact that we do not measure absence duration accurately; there are probably some cases in which benefits are exhausted in the 11th or the 13th months. In the analysis, we interpret the few absence spells lasting more than 12 months as continuing with the spell duration dummy variables 'frozen' at 12 months.

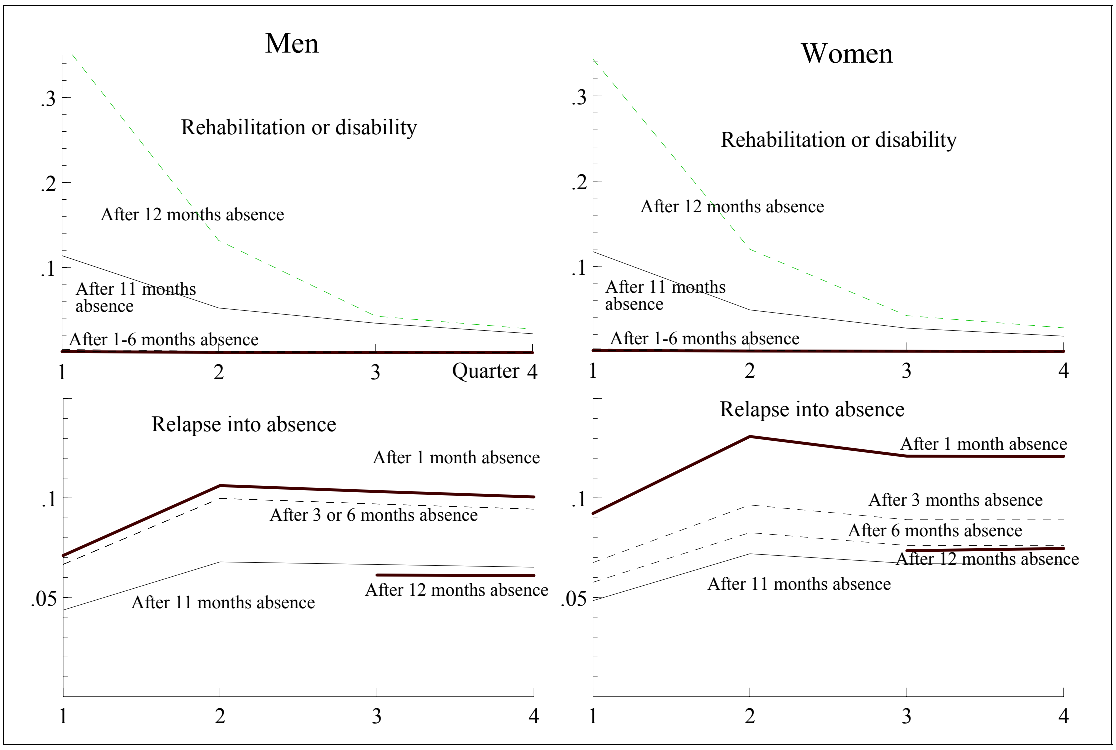

Figure 3. Predicted transition probabilities to rehabilitation/disability and relapse into absence during first four quarters of work-resumption, for different durations of the initial absence spell (calculated for a randomly selected entrant into work resumption spell)

As we pointed out above, transition rates out of absence may give an incomplete picture of how the institutional characteristics of the benefit system affects absence behaviour, since some persons may return to absence soon after an 'apparent' work-resumption transition or simply move on to rehabilitation or disability. Figure 4 summarizes the pattern of work-resumption, relapse and disability transitions, combining the processes depicted in Figures 2 and 3 in the form of a *joint probability* of work-resumption and *not* returning to absence or transit to a rehabilitation/disability state within the subsequent 12 months; i.e. the probability of 'lasting' work-resumption. We see that the probability of making this kind of 'lasting' transition out of the Social Insurance system is substantially lower than the 'short-term' probability of resuming work. Moreover, the spike in the transition rate around the time of benefit exhaustion becomes much less pronounced, implying that a substantial part of the increase in the work-resumption rate around the time of benefit exhaustion is short-lived. This probably reflects that the work behaviour of persons who have experienced a one-year sickness spell is to a large extent determined by purely medical conditions.

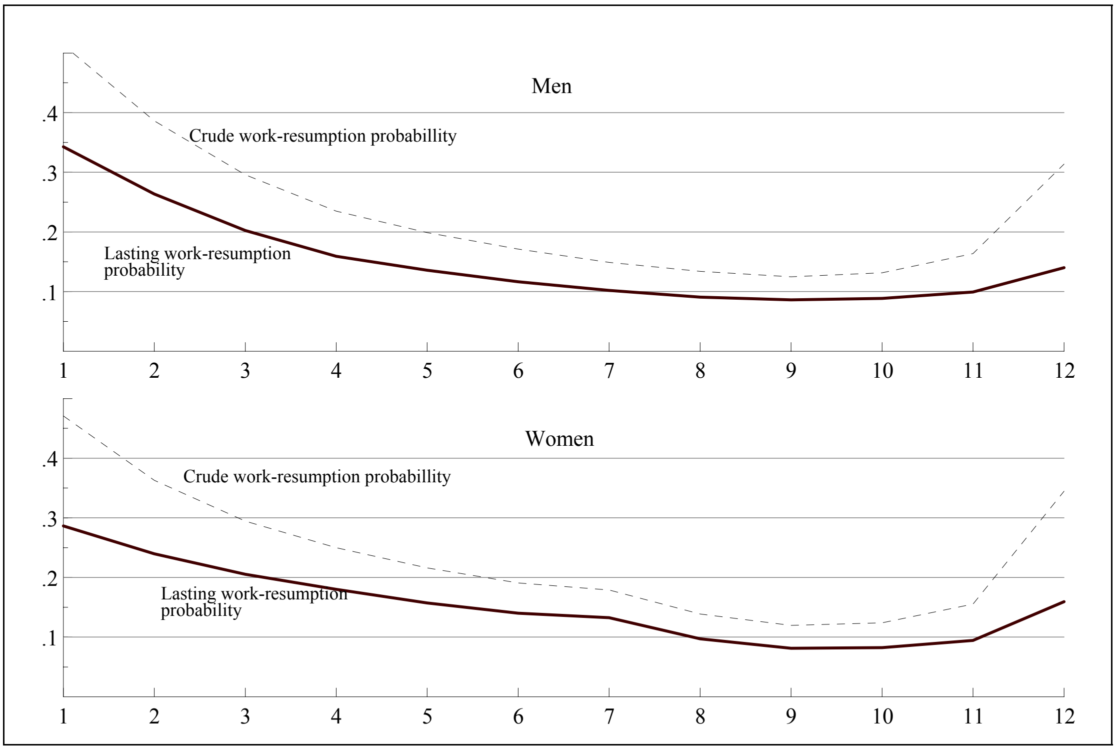

Figure 4. Predicted monthly probabilities of resuming work with and without conditioning on no relapse or rehabilitation/disability transition in the subsequent 12 months.

One may still question whether the remaining 12th month spike represents a genuine increase in lasting work-resumption hazards or whether it reflects measurement error in the sense of complete withdrawals from the labour force (without rehabilitation –or disability benefits). We have had the opportunity to check that out for some cohorts in the dataset. In particular, we selected persons who completed a full sickness period of 12 months during the last three months of 1996 and who did not take up any form of rehabilitation –or disability benefits during 1997, and investigated the distribution of recorded wage income (collected from administrative registers) for this group of persons in 1997. As it turned out, only 8.6 per cent of these persons did not have any wage income at all in 1997. An additional fraction of 11.2 per cent earned less than 50,000 NOK (6,500 Euro). Comparable numbers for persons who completed sickness spells of only six months duration in the same period was 5.0 per cent (with zero income) and 17.0 per cent (with positive earnings below 50,000 NOK). Based on these statistics, we conclude that most of the apparent rise in the last-ing work-resumption hazard that occurs at the end of the sickness benefit period is *genuine*.

4.2 The Causal Effects of Business Cycles on Work-Resumption

Table 2 presents the estimated business cycle effects. The two business cycle indica-tors are both normalised such that they have unit range, i.e. a positive change in the variable of one unit corresponds to a change from the worst state (or largest decline) of labour market tightness that was observed during the estimation period to the best state (or largest increase). This implies that (*exp(parameter estimate)-1*), which for small values of the parameter estimate is close to the parameter estimate itself, can be interpreted as the relative change in the corresponding hazard rate that is predicted to occur when the business cycle (or the change in the business cycle) changes from its lowest to its highest value. The estimated business cycle effects are in line with our prior expectations for both men and women, although they seem to be stronger for women. Improvements in the business cycle tend to reduce the work-resumption rates, and increase the relapse rates, substantially – both through a level and a differ-ence effect - suggesting that there is a disciplinary effect associated with poor and de-teriorating business cycle conditions. The effects of business cycles are both statisti-cally and substantively significant. For men, we find that when the labour market is both tight and tightening, the work-resumption hazard is 15-20 per cent lower – and the relapse hazard 15-20 per cent higher - than when the labour market is both slack and slackening. For women, the work-resumption hazard is as much as 25-30 per cent higher – and the relapse hazard 25-30 per cent higher – in a tight and tightening labour market (compared to a slack and slackening). As pointed out in Section 1, there is also another interpretation of this relationship than that of disciplinary effects, namely that the higher work-pressure associated with favourable business cycle conditions makes it more difficult for persons with a fragile health to resume work. Note, however, that the selection effects associated with less healthy workers being pulled into the labour market cannot explain the parameters presented in Table 2, since we condition on the business cycle conditions at the moment of entry into sickness (see next subsection).

Table 2
Estimated effects of the business cycle at the time of potential transition
(Standard errors in parentheses)

	Men			Women		
Destination state	Work-resump.	Rehab. Disab.	Relapse	Work-resump.	Rehab. Disab.	Relapse
Level of business cycle at the moment of potential transition	-0.168 (0.018)	-0.414 (0.031)	0.111 (0.020)	-0.259 (0.014)	-0.320 (0.026)	0.198 (0.016)
Change in business cycle at the moment of potential transition	-0.046 (0.009)	0.121 (0.028)	0.073 (0.013)	-0.053 (0.007)	0.145 (0.023)	0.084 (0.010)

The probability of entering into disability or rehabilitation declines quite substantially in the level of the business cycle, but increases in its difference. One may speculate that strain and stress at the workplaces are related to large *increases* in economic activity more than to high activity itself. Hence, it may be more difficult for employees with poor health to resume work during times of increasing demand, and, as a result, more workers are pushed into rehabilitation or disability.

4.3 The Effects of Business Cycles on the Composition of Entrants into Absence

Table 3 present the estimated effects of business cycles on the composition of entrants into sickness absence. Again, we find that the estimates are in line with prior expectations. Improvements in the business cycles are clearly associated with more healthy entrants into absence spells, with higher work-resumption hazards and lower relapse hazards, ceteris paribus. Our results at this point therefore fits well into the discipline hypothesis; i.e. that it is more tempting to consider a given medical condition as justifying an absence spell when business cycles are improving than when business cycles are deteriorating. However, the level of the business cycle seems to have exactly the opposite effect. Persons who become sick during favourable, but stable, business cycle conditions tend to have poor work-resumption prospects and high relapse propensities, ceteris paribus. We interpret this as a confirmation of the 'selection hypothesis', i.e. that the average health condition among employed workers is negatively related to the employment rate.

Table 3
Estimated selection effects of the business cycle at the time of entry into absence
(Standard errors in parentheses)

	Men			Women		
Destination state	Work-resump.	Rehab. Disab.	Relapse	Work-resump.	Rehab. Disab.	Relapse
Level of business cycle at the moment of entry into absence	-0.182 (0.018)	0.272 (0.029)	0.170 (0.019)	-0.097 (0.014)	0.314 (0.024)	0.096 (0.016)
Change in business cycle at the moment of entry into absence	0.174 (0.009)	-0.062 (0.034)	-0.068 (0.013)	0.140 (0.007)	-0.043 (0.029)	-0.083 (0.010)

4.4 Observed and Unobserved Heterogeneity

Unsurprisingly, one of the most important characteristics to explain individual hazard rates is age. This is illustrated in Figure 5. Work resumption prospects decline monotonously with age, particularly for men. The rehabilitation/disability and relapse hazards increase with age. Again, the age effect is stronger for men than for women. It is natural to interpret the differences across age groups as reflecting differences in average health conditions. The apparently larger effect for men, particularly at high ages, is probably related to the fact that older men on average have much longer work-careers than women.

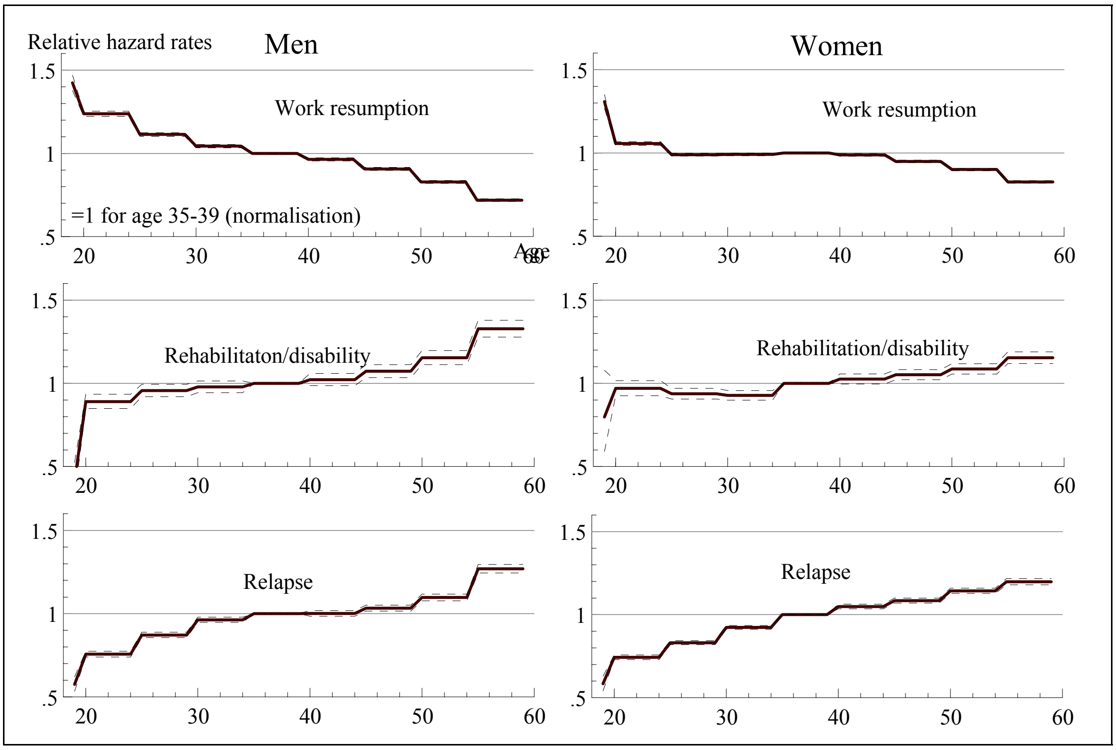

Figure 5. Estimated relative hazard rates for different age groups (with 95 per cent point-wise confidence intervals).

Educational attainment also has a substantial impact on absence behaviour. This is illustrated in Figure 6. The higher is the education, the higher is the work-resumption probability and the lower is the relapse probability. The latter effect is much stronger than the former. It is natural to interpret the relationship between absence behaviour and education as reflecting two different mechanisms. First, there are unobserved differences in health conditions across groups with different educational attainment; see e.g. Grossman (2000) for a recent summary of empirical evidence regarding this relationship. And, second, there are unobserved differences in the type of work that is performed by the different education groups, with poorly educated workers more subject to manual work and long work-careers.

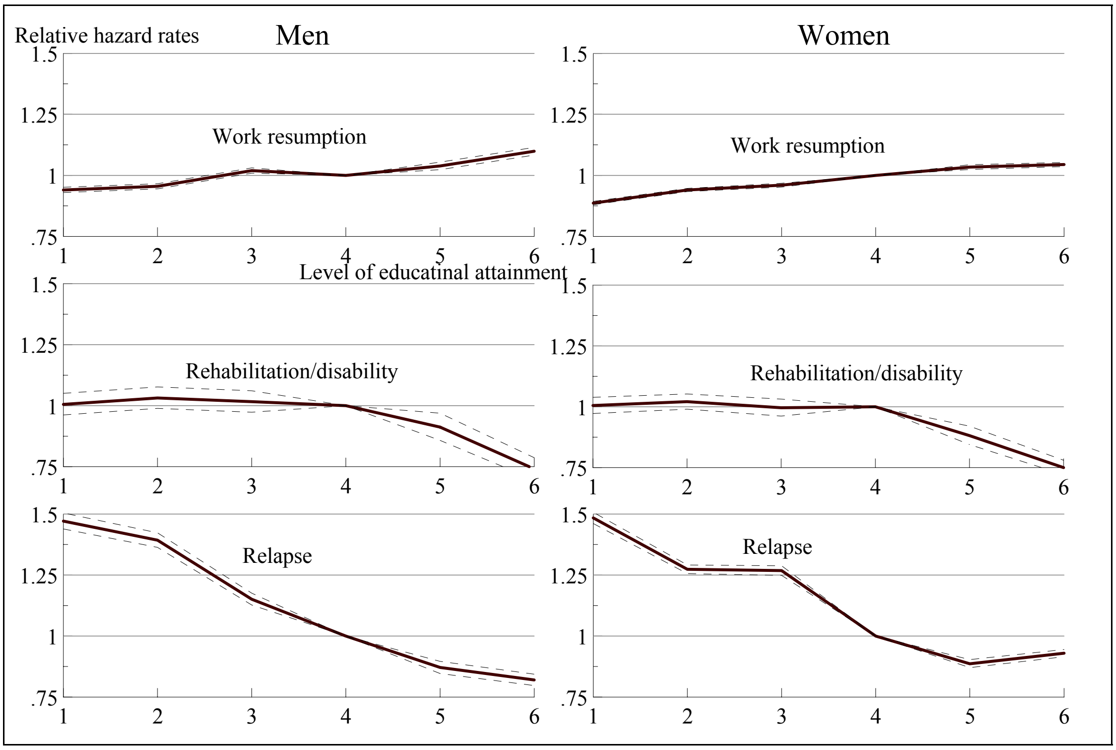

Figure 6. Estimated relative hazard rates for different educational groups (with 95 per cent point-wise confidence intervals).
Note: Educational attainment is grouped as follows: 1=only compulsory education, 2=lower secondary school, 3=higher secondary school, vocational subjects, 4=higher secondary school, general subjects, 5=college, 6=university degree.

We now turn to some results regarding the effects of observed heterogeneity that we do not spell out in detail (complete results are available upon request). For both men and women, income just prior to the absence spell (represented in the model by seven dummy variables) seems to affect the work-resumption propensity in a u-shaped pattern, implying that the work resumption hazard is lowest for persons with medium incomes. This is mirrored by a hump-shaped pattern in the effect of income on the relapse hazard. For men, high income unambiguously and quite strongly reduces the probability of entering into rehabilitation or disability. There is a similar, but much weaker effect for women. And for women, the effect is not entirely monotonous; at very high incomes, there is a tendency for even higher incomes to increase the rehabilitation/disability hazard. Marital status has a substantive impact on absence behaviour, particularly for men. Being married implies a higher work-resumption hazard, a (slightly) lower rehabilitation/disability hazard and a (substantially) lower relapse hazard. Immigrants from developing countries have a lower work-resumption hazard and higher relapse hazard than natives. The latter of these effects is quite strong, i.e., compared to natives, the relapse hazard is as much as 60 per cent higher for immigrant men and 47 per cent higher for immigrant women. The rehabilitation/disability hazard is, on the other hand, around 25 per cent lower for immigrants than for natives.

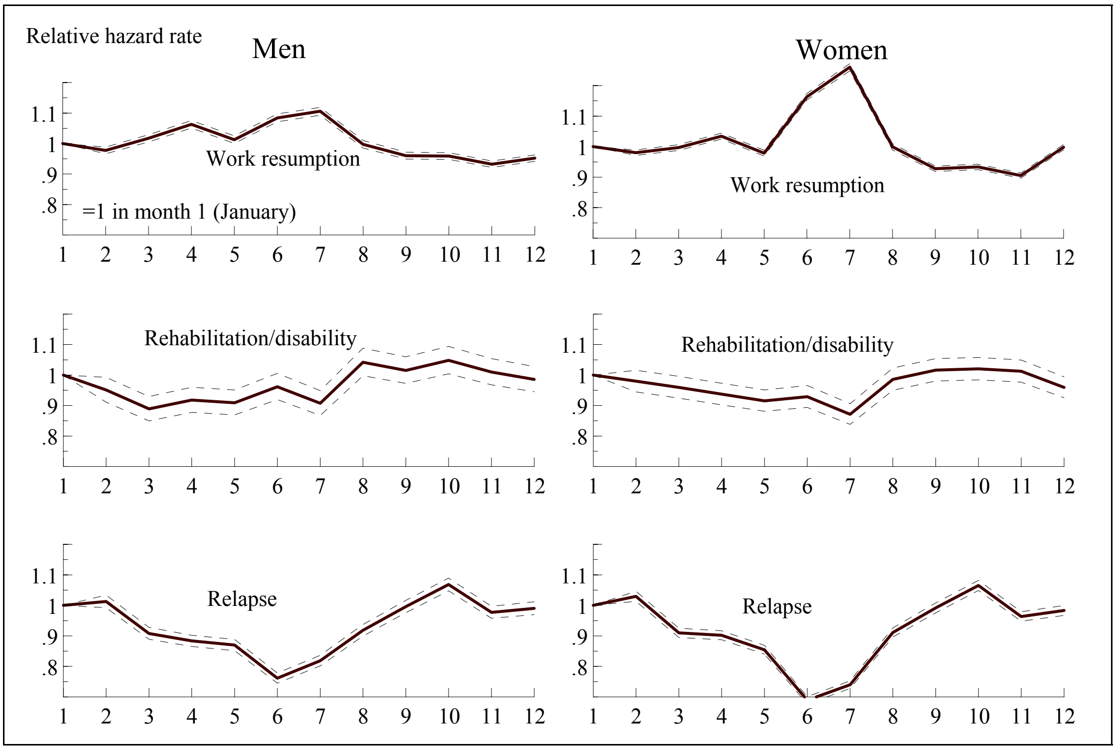

Figure 7. The estimated seasonal pattern in relative hazard rates (with 95 per cent confidence intervals).

Notes: Months are numbered from January (1) to December (12). For the quarterly transition probabilities, we use the first calendar month in each (overlapping) quarter, e.g. if the quarter in question is February, March and April, we attribute the February-parameter to this observation.

We also estimated seasonal effects in the various transition rates. The result is illustrated in Figure 7. There is a strong seasonal pattern in work resumption and relapse hazards, particularly for women. While the work-resumption hazards are highest during the summer, the relapse hazards are highest during the winter.

The estimated mass-point distribution of unobserved heterogeneity is not directly interpretable, since different combinations of mass-points and probabilities sometimes yield (almost) equivalent results. Our experience suggests, however, that the first and second order moments of the heterogeneity distribution are relatively robust. Hence, in Table 4, we report the imputed correlation coefficients. They reveal essentially the same correlation structure for men and women, and suggest that there is a negative correlation between unobservables that affect the work-resumption propensity, on the one hand, and the rehabilitation/disability and relapse propensities, on the other. There is a positive correlation between unobservables affecting rehabilitation/disability propensity and relapse propensity.

Table 4
Estimated correlation coefficients for unobserved characteristics, based on 8 mass-points in the male heterogeneity distribution and six mass-points in the female heterogeneity distribution

$$corr(\exp(v_j), \exp(v_k)),\ k \neq j$$

	Men		Women	
	Rehabilitation Disability	Relapse	Rehabilitation Disability	Relapse
Work-resumption	-0.204	-0.365	-0.263	-0.600
Rehabilitation/Disability		0.243		0.298

5 Concluding Remarks

In this paper, we have investigated the transition rate pattern out of sickness absence spells, on the basis of Norwegian register data containing more than two million absence spells distributed over a period of eight years. A main objective has been to take advantage of the huge dataset to avoid unjustified restrictions on the way observed, as well as unobserved, heterogeneity affect the transition rates. For this purpose, we have set up and estimated a dependent competing risks transition rate model for direct transitions into work-resumption and rehabilitation/disability, as well as for delayed transitions into rehabilitation/disability and relapses into sickness absence.

Our main conclusions are as follows:

i) Business cycle conditions have substantial effects on work-resumption behaviour. Favourable business cycle developments during the progression of absence spells imply lower work-resumption rates and higher relapse rates (even when the business cycle conditions at the moment of entry are controlled for). These effects are both statistically and substantively significant, and stronger for women than for men.

ii) Institutional features of the health-insurance system also matter for work-resumption behaviour. In particular, we find that the work-resumption propensity increases around the time of full benefit exhaustion after one year of absence. A large part of this increase is subsequently diluted, however, through higher relapse and disability rates. Hence, economic factors seem to have a significant, but limited effect on work behaviour at this late stage of absence spells.

iii) There is substantial degree of heterogeneity in individuals' work-resumption prospects. The most important explanatory variables are age and educational attainment. Higher age implies, ceteris paribus, lower work-resumption probabilities, higher relapse probabilities, and higher rehabilitation/disability probabilities. Higher education implies higher work-resumption rates and (in particular) lower relapse rates.

These observations can hardly be translated directly into policy-recommendations regarding the design of public sickness insurance schemes. One important lesson, however, is that sickness absence is not a pure medical phenomenon that can be treated as completely independent of the workers' preferences, work ethics and incentives. And the fact that workers seem to adjust their absence behaviour according to the economic incentives they face implies that governments have to cope with a moral hazard dilemma. Our results leave little doubt that a thriftier health insurance scheme could contribute to a reduction of sickness absence in Norway. However, the potential gain in terms of reduced 'unnecessary' sickness absence has to be traded off against the possible promotion of sickness 'presenteeism' and the cost of increased uncertainty and inequality (associated with less than full insurance). Moreover, we have found that there is a cost associated with taking up work too quickly, in the sense that this increases the relapse probability later on. Hence, a longer spell of absence may in some cases be considered a health investment. Liquidity constrained individuals may not be able to make this investment on their own.

References

Arai, M. and Thoursie, P. S. (2001) Incentives and Selection in Cyclical Absenteeism. FIEF Working Paper Series, No. 167.

Askildsen, J. E., Bratberg, E. and Nilsen, Ø. A. (2002) Unemployment, Labour Force Composition and Sickness Absence. A Panel Data Study. IZA Discussion Paper No 466, 2002.

Barmby, T. A., Sessions, J. G. and Treble, J. G (1994) Absenteeism, Efficiency Wages and Shirking. *Scandinavian Journal of Economics*, Vol. 96, 561-566.

Barmby, T. A., Ercolani, M. G and Treble, J. G. (2002) Sickness Absence: An International Comparison. *The Economic Journal*, Vol. 112, F315-F331.

Brinch, C. (2000). 'Identification of structural duration dependence and unobserved heterogeneity with time-varying covariates'. Memorandum No. 20/2000, Department of Economics, University of Oslo.

CESifo (2002) CESifo Forum, Vol. 3, No. 4, Munich, Germany.

Chatterji, M. and Tilley, C. J. (2002) Sickness, Absenteeism, Presenteeism, and Sick Pay. *Oxford Economic Papers*, Vol. 54, 669-687.

Dyrstad, J. M. and Ose, S. O. (2002) Non-Linear Unemployment Effects in Sickness Absence: Discipline or Composition Effects? Working Paper No. 25/2002, Department of Economics, Norwegian University of Science and Technology.

Elbers, C. and Ridder, G. (1982) True and Spurious Duration Dependence: The Identifiability of the Proportional Hazard Model. *Review of Economic Studies*, Vol. 49, 403-410.

Gaure, S. and Røed, K. (2003) How Tight is the Labour Market? A Micro-Based Macro Indicator. Memorandum No. 9/2003, Department of Economics, University of Oslo (www.oekonomi.uio.no/memo/memopdf/memo0903.pdf)

Grossman, M. (2000) The Human Capital Model. In A. J. Culyer and J. P. Newhouse (eds.) Handbook of Health Economics. Elsevier Science B. V., Amsterdam, 347-408.

Heckman, J. and Singer, B. (1984a) The Identifiability of the Proportional Hazard Model. *Review of Economic Studies*, Vol. 51, 231-241.

Heckman, J. and Singer, B. (1984b) A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data. *Econometrica*, Vol. 52, 271-320.

Johansen, P. R. and Eika, T. (2000) Driving Forces Behind Business Cycles in the 1990's. *Economic Surveys* No. 3/2000, Statistics Norway (http://www.ssb.no/english/subjects/08/05/10/es/200003/johansen.pdf).

Johansson, P. and Palme, M. (2002) Assessing the Effect of Public Policy on Worker Absenteeism. *Journal of Human Resources*, Vol. 37, No. 2, 381-409.

Lindsay, B. G. (1983) The Geometry of Mixture Likelihoods: A General Theory. *The Annals of Statistics*, Vol. 11, 86-94.

McCall, B. P. (1994) Identifying State Dependence in Duration Models. American Statistical Association 1994, Proceedings of the Business and Economics Section, 14-17.

Meyer, B. D., Viscusi, W. K. and Durbin, D. L. (1995) Workers' Compensation and Injury Duration: Evidence from a Natural Experiment. *American Economic Review*, Vol. 85, No. 3, 322-340.

Van den Berg, Gerard J. (2001) Duration Models: Specification, Identification, and Multiple Durations. *Handbook of Econometrics*, Vol. 5, Chapter 55, Elsevier.

Appendix: List of Explanatory Variables

The following variables were included in estimation of the destination specific hazard rates:

<u>Transitions to work resumption</u>

Age: 9 dummy variables (<20, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59)
Income: 7 dummy variables
Region: 6 dummy variables
Duration of *ongoing* absence spell: 12 dummy variables
Season of *ongoing* absence spell: 12 dummy variables
Sickness degree: 4 dummy variables (<50%, 50-70%, 71-99%, 100%)

Educational attainment: 7 dummy variables
Business cycle indicator in potential transition month: 1 scalar variable
Change in business cycle indicator from the last to the current potential transition month: 1 scalar variable
Business cycle indicator in the month of entry into the spell sequence: 1 scalar variable
Change in business cycle indicator from the month before entry to the month of entry into the spell sequence: 1 scalar variable
Marital status: 1 dummy variable
Non-Western immigrant: 1 dummy variable

Transitions to rehabilitation/disability

Age: 9 dummy variables (<20, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59)
Income: 7 dummy variables
Region: 6 dummy variables
Indicator for completed absence spell (work resumed): 1 dummy variable
Duration of *ongoing* absence spell for still active absence spells: 12 dummy variables
Duration of *completed* absence spell for completed absence spells: 12 dummy variables
Duration of *ongoing* work-resumption spell for active work-resumption spells: 4 dummy variables
Indicator variable for having completed an absence spell lasting 12 months (i.e. exhausted benefits) and being in the second quarter of ongoing work-resumption spell
Indicator variable for having completed an absence spell lasting 12 months (i.e. exhausted benefits) and being in the third or fourth quarter of ongoing work-resumption spell
Season of ongoing absence or work-resumption spell: 12 dummy variables
Sickness degree: 4 dummy variables (<50%, 50-70%, 71-99%, 100%)
Educational attainment: 7 dummy variables
Business cycle indicator in potential transition month or in the first month of the potential transition quarter (for work-resumption spells): 1 scalar variable
Change in business cycle indicator from the last to the current potential transition month or to the first month of the potential transition quarter (for work-resumption spells): 1 scalar variable
Business cycle indicator in the month of entry into the spell sequence: 1 scalar variable
Change in business cycle indicator from the month before entry to the month of entry into the spell sequence: 1 scalar variable
Marital status: 1 dummy variable
Non-Western immigrant: 1 dummy variable

Transitions to relapse (a new sickness spell after work is resumed)

Age: 9 dummy variables (<20, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59)
Income: 7 dummy variables
Region: 6 dummy variables
Duration of *completed* absence spell: 12 dummy variables
Duration of *ongoing* work-resumption spell: 4 dummy variables
Season of ongoing absence or work-resumption spell: 12 dummy variables
Sickness degree: 4 dummy variables (<50%, 50-70%, 71-99%, 100%)
Educational attainment: 7 dummy variables
Business cycle indicator in the first month of the potential transition quarter: 1 scalar variable
Change in business cycle indicator from the last to the first month of the potential transition quarter: 1 scalar variable
Business cycle indicator in the month of entry into the spell sequence: 1 scalar variable
Change in business cycle indicator from the month before entry to the month of entry into the spell sequence: 1 scalar variable
Marital status: 1 dummy variable
Non-Western immigrant: 1 dummy variable