HEALTH ECONOMICS
Health Econ. **11**: 221–231 (2002)
Published online 29 November 2001 in Wiley InterScience (www.interscience.wiley.com). **DOI:**10.1002/hec.648

A general model of the impact of absenteeism on employers and employees

Mark V. Pauly[a,*], Sean Nicholson[a], Judy Xu[a], Dan Polsky[b], Patricia M. Danzon[a], James F. Murray[c] and Marc L. Berger[c]
[a] *Health Care Systems Department, The Wharton School, 3641 Locust Walk, Philadelphia, PA 19104-6218, USA*
[b] *Division of General Internal Medicine, University of Pennsylvania, PA, USA*
[c] *Outcomes Research & Management, Merck & Co., Inc., USA*

Summary

Most studies on the indirect costs of an illness and the cost effectiveness of a medical intervention or employer-sponsored wellness program assume that the value of reducing the number of days employees miss from work due to illness is the wage rate. This paper presents a general model to examine the magnitude and incidence of costs associated with absenteeism under alternative assumptions regarding the size of the firm, the production function, the nature of the firm's product, and the competitiveness of the labor market. We conclude that the cost of lost work time can be substantially higher than the wage when perfect substitutes are not available to replace absent workers and there is team production or a penalty associated with not meeting an output target. In the long run, workers are likely to bear much of the incidence of the costs associated with absenteeism, and therefore be the likely beneficiaries of any reduction in absenteeism. Copyright © 2001 John Wiley & Sons, Ltd.

Keywords absenteeism; workloss; productivity; indirect costs

Introduction

There is a growing interest in quantifying the benefits to employers of programs that reduce the number of days their employees miss from work because of illness [1,2]. If these benefits are sufficiently large, employers may be able to justify offering better health insurance policies to their employees and implementing wellness programs. At present, almost all cost of illness studies and cost effectiveness analyses of medical interventions use the wage rate to estimate the benefit of reduced absenteeism [3–6]. This approach is often based on the so-called 'human capital' approach to valuing health, in which the loss of a healthy day represents the loss of production whose value in competitive labor markets equals the money wage. In this paper, we argue that such approaches are based on assumptions that are not generally appropriate for modeling employer decisions on employment or impacts on firm output, and we examine the cost of work loss under alternative assumptions. On the one hand, we show that in a general long-run model with full employment – so that employer and societal perspectives on cost coincide – wages represent a lower bound for losses from a day of missed work that could be much larger than the wage per day. On the other hand, in some circumstances the long-run benefit from reduced work loss is captured, not by the

*Correspondence to: Health Care Systems Department, The Wharton School, 3641 Locust Walk, Philadelphia, PA 19104-6218, USA. Tel.: +1-215-898-2838; fax: +1-215-573-2157; e-mail: pauly@wharton.upenn.edu

Received 3 July 2000
Accepted 11 June 2001

employer, but by the employee. However, even in this case, it will be rational for employers to implement any cost-effective program that reduces work loss. Thus, traditional measurement methods are likely to misestimate the true gain to employers and to society from implementing policies that would reduce absenteeism. If there is less than full employment, then these conclusions will be modified, but it will still be true that there are costs of lost work time that are positive and exceed friction costs, that vary across production processes and labor markets, and that ought to be taken into account in the analysis from either an employer or societal perspective.

Our objective in this paper is, therefore, to specify a more general model of the incidence and effects of reductions in the lost work time. We identify characteristics of firms and markets that determine whether the gross benefits will be large or small, and how they will be distributed between employer and employee. We provide a categorization of combinations of characteristics that can help to predict how large the gain will be for a given employer or a set of employers. We also consider how employer and societal perspectives will differ when there is less than full employment.

Cost of illness studies estimate the sum of the direct and indirect costs of an illness [3,5,7]. Direct costs include the medical costs of treating individuals who have the illness; indirect costs include the value of the work loss that results when these individuals are absent from work or experience diminished productivity at work while recovering from the illness. This paper focuses on the *marginal* cost of an absence due to illness and the marginal benefit of an intervention that reduces absenteeism, rather than the *average* cost as in cost of illness studies. Focusing on the marginal cost of absenteeism and the incidence of these costs will help identify parties that have an incentive to implement programs to improve health.

The basic model: a single homogeneous input

For our initial discussion, we present a model in which workers are hired in competitive labor markets under full employment; the money wage is determined as the equilibrium price in that market. Firms and workers are assumed to be identical *a priori*, although workers and workforces will experience variation in actual work loss days. Workers obtain the same utility whether they are well or ill on a workday; the additional leisure associated with time off is exactly offset by pain and discomfort. The perspective, therefore, is that of firm owners and their workers; there are no (non-transitionally) unemployed people whose costs and benefits need to be considered. We begin by analyzing a firm with a simple production function that can store its output at zero cost and does not offer sick leave benefits. We then consider alternate scenarios that may alter the appropriate measure of the cost of absenteeism.

In the benchmark model, firms hire homogeneous workers (L) to produce a product (Y) that can be stored at zero cost. With low inventory costs, there are no penalties if a firm's output varies from day to day due to variations in the number of absent workers. If all the workers scheduled to work actually show up, the relatively high output for that day increases inventory. Conversely, if an unusually large number of workers are ill on a particular day, output is relatively low and inventory is depleted. The costs of absenteeism will be smaller in this production scenario than in any other because firms will not need to make potentially expensive adjustments to minimize the short-term effects of absenteeism due to illness. Data processing, billing, and telemarketing firms may be characterized by such a production function. If a data processor is ill and misses work, the work can usually be completed by the absent worker at a later date without any loss in the firm's revenue.

A firm with the production function described above would hire workers until the marginal product is equal to the market wage of w per day, which is assumed to be determined by worker productivity or demand elsewhere in the local economy:

$$Y = f(L)$$
$$\text{MRP} = pf' = w$$

where p is the price of the product, f' is the worker's marginal product, and MRP is the marginal revenue product of labor. In the benchmark model, there is no sick leave policy; employees receive a wage w for each day they actually work and zero if they are absent due to an illness.

When an employee is absent, the firm's revenue and labor costs are both lower by the amount w, the daily wage. Employees bear the costs associated with absenteeism and would reap all the gains associated with reduced work loss. If the firm institutes a program that improves the health of its workers and reduces absenteeism by one day per worker, then an employee's annual income would increase by approximately w. (We ignore the possibility that mild illness reduces the productivity of those who do show up for work, the so-called 'presenteeism' effect; we also ignore any effects on the market equilibrium wage per day from changes in the aggregate supply of (healthy) working days.)

We now adapt the model to account for sick day benefits. Most firms provide sick day benefits to their employees by allowing workers to be paid for a certain number of days when they are absent due to an illness. Paid wages are fixed for a given time period (e.g. a year) and are not affected by the number of days a worker is actually absent. We assume that sick days are not predictable for any individual worker, although the average number of sick days can be predicted for a large number of similar workers.

The existence and persistence of this fringe benefit is not due to employer altruism. Instead, it can be explained by a rational desire of risk averse workers to have the (large) firm pool their individual risks and insure against the loss of wage income due to the unpredictable nature of illness. Such workers are willing to accept lower wages per day actually worked in return for a guarantee of payment when illness strikes. Consider firms that want to set a wage rate for a year that consists of 250 work days. All firms expect workers to miss an average of m days per year due to illness, and their expectations are correct. In a competitive labor market, the annual wage rate will be set equal to the marginal revenue product of $(250 - m)$ days, or $w(250 - m)$, where w is both the marginal revenue product of labor and the wage per day actually worked, as determined by the market. Spreading the worker's marginal revenue product over the 250 work days in the year yields an average wage per day paid of $w^* = w(250 - m)/250$. Thus, employers do not 'give' sick days; workers pay for them in the form of lower wages per worked day, but workers prefer this arrangement to one of positive payment per day worked and zero payment per day missed because this arrangement averages out the variation across workers in the numbers of days missed due to illness.

The cost to the firm when a worker is absent due to illness is the worker's marginal revenue product, which is equal to the wage per day actually worked (w). A program that decreases m by 1 day will increase the value of the firm's output by w, which is moderately larger than the wage rate per day paid (w^*), if m is small. Even at this very simple level, the wage per day paid is an underestimation of the value to the firm of reducing m, but the wage per day worked is an accurate measure.

Who receives the gains when a firm that offers such a sick leave policy institutes a program that reduces work loss due to illness? A key issue in answering this question is whether other employers can identify or determine that the illness probability of a set of workers has fallen. Let us first consider the most transparent (though, we will argue, not always the most realistic) case in which the treated workers' improved health is observable by and permanently 'transferable' to any new employer. In Gary Becker's classic sense, this health capital is general rather than firm-specific [8]. Examples of this kind of treatment might be smoking cessation or weight loss programs. We ignore for the moment laws that might forbid discrimination in wages, and instead assume that employers are free to pay any mutually agreed-upon wage; employers are permitted to pay higher wages to workers with better health habits. If treatment is successful, then treated workers will be more productive for a long period of time; they can be identified as such. In this case, it is clear that these workers will receive higher wage offers from competing employers (to reflect their improved health), which should increase the wage they receive from their current employer. In equilibrium, the benefits from improved health will be almost entirely captured by the healthier employees (in the form of an increase of $w/250$ in the wage per day paid) and *not* by their employers.[a] Of course, in the short run, before the wage contracts are renegotiated, employers would be likely to capture the benefits of reduced work loss.

This example also demonstrates the necessary conditions for employers to translate health improvements directly into higher employer profits rather than higher employee wages. The health intervention must be one that is secret or difficult to detect, and/or impossible for other employers to reproduce. Hidden, unique health improvement programs are the ones for which the traditional

method of measuring the gain to employers from cost-effective work loss reduction programs are most appropriate, but easily observable and easily replicable programs will largely benefit workers in the form of higher wages.[b]

For example, suppose that one cause of missed work is substance abuse. If an employer can discover a low cost and effective treatment program that is difficult for other employers to copy (unique) and whose beneficial effects are hard for other employers to observe (hidden), the employer needs to pay no more than the prevailing wage to the now-more-productive treated workers, and yet can benefit from their increased productivity. If, in contrast, the low-cost program can be easily copied, all other employers will be expected to do so; average worker productivity will rise and with it the wage. Or, alternatively, if the innovating firm's more productive workers can be identified in the labor market and the treatment program has a permanent effect on the worker's health, their wages will rise.

The potential failure of employers to capture in productivity all the long-run benefits of cost-effective interventions to reduce work loss should not, however, eliminate their incentive to institute such interventions. The reason why employers provide employee benefits in competitive labor markets is presumably because doing so is the best way to attract and retain employees; indeed, in the fully competitive model any employer who does not provide the benefit will be driven out of business by higher labor costs than those of his competitors who do provide benefits. Concretely, if an intervention provides improved productivity benefits, then employers will provide that intervention as long as these benefits exceed the program's cost. As usual in competitive markets, there are strong incentives for firms to take actions that temporarily reward the firm that takes the initiative but ultimately benefit workers or consumers.

This view also suggests that employers ought to value providing coverage for care that will be effective for health in the future even for workers with observable health improvements who might move to another firm. Since such benefits will make workers more productive at any new firm, workers should be willing to accept wages lowered by the cost of the benefit in the current period at the firm paying for the care, to reflect their higher future earnings either at that firm or at some alternative future employer. It is well known that workers do accept lower wages when employers offer valued health benefits [9]. The ability of the employers to observe the potential health improvement is obviously key to this conclusion, but it does mean that effective care with a long payoff period nevertheless, ought, to be valued by high turnover firms and their employees.

We now consider whether the cost to the firm of work loss and the incidence of this cost is different when firms pay workers an annual salary and require them to make up during what would otherwise be their leisure time for any work missed due to illness. In a competitive labor market, the annual salary would need to be sufficiently high to generate the same utility for the worker as the compensation package in other firms, where workers are paid when they are absent due to illness but do not have to make up for the work. In effect, the first firm must pay more (compared to the other firms) in salary because it requires more work, some of it at inconvenient times. Such a policy transforms unpredictable illness into unpredictable work on weekends or at night. If workers derive utility from income and leisure and the number of sick days are exogenous, workers would require an annual salary s such that

$$U(w(250 - m),\ L) = U(s,\ L - m)$$

At the optimal number of days worked per year, assumed here to be 250, the utility of a day of leisure would equal the foregone wage (w). If the utility function is additively separable, then the required salary (s) to attract employees will be at least $250w$. The cost to salaried workers of an absence is the value of the leisure time they use to make up for the work they miss. A program that reduces m by one day has no (long run) effect on the firm's output, but it does increase the utility of the salaried worker by increasing leisure by one day. As before, a salaried worker bears the incidence of the cost of an absence and reaps the benefits of a reduction in absenteeism.

These conclusions need not hold in a short-run situation in which absence rates change but wages have yet to adjust. For instance, if the absence rate increases by one day in a firm that requires workers to make up for lost work but wages have yet to adjust to a higher chance of inconvenient work, then the cost to the employer would be less than the daily wage per worker. However, once wages adjust, the cost will always be equal to or greater than the wage.

The main conclusions from the benchmark model are:

(1) Even with very simple production functions, the value to the firm and/or to the worker of avoiding a work loss day will equal the wage per day worked.
(2) In competitive labor markets, some of the benefits from reduced work loss could fall to workers in the form of higher wages and/or more convenient work times – depending on the nature of the health improvement, worker preferences and labor supply elasticities, and the knowledge structure in the labor market.
(3) In a competitive labor market, firms will have incentives to institute programs that reduce absenteeism even if their employees reap the benefits of the lower work loss.

Team production

Thus far in our analysis, the daily wage has been assumed to be an accurate indicator of the relative cost of work loss (across jobs and firms). Now we consider alternative forms of the production process and the firm's reaction to work loss in which the value of avoiding lost work time can be much greater than the daily wage. There are two necessary general conditions for the consequences to the firm that are greater than the worker's wage. One is that the effect of the worker's absence on firm revenues and/or output is greater than the value of the worker's daily output. The other is that it is not costless to find a perfect substitute for an absent worker.

Even if output, once produced, can be stored, unexpected absences can have serious consequences if production of output requires that a set of team members show up at the same time. When a member of a team is absent, he affects the marginal product of the entire team, rather than just his own marginal product. Consider an extreme form of team production where a firm employs capital (K) and two labor inputs (L_1 and L_2) in a Leontief production function to produce output Y:

$$Y = \min\{aL_1, bL_2, dK\}$$

The terms a, b, and c are productivity parameters. The cost, C, of producing output Y is

$$C = Y(w_1/a + w_2/b + v/d)$$

where v is the rental price per unit of capital, and w_1 and w_2 are the wages of the two distinct labor inputs. Consider a case where $a = 1$, $b = 2$, $d = 3$, and output sells for a price of \$10 per unit. If the firm wants to produce six units of output, then it needs to assemble a team consisting of six type-one workers, three type-two workers, and two units of capital. If $w_1 = \$4$, $w_2 = \$6$, and $v = \$9$, then the team will be paid its marginal revenue product of \$60 and the firm will have economic rents of zero.

Consider first the extreme case in which there is what might be called 'team-specific human capital'; no worker new to the team can be effectively substituted for an absent team member. If one of the three type-two workers scheduled to work does not show up, the remaining team would only be able to produce four rather than six units of output. (We assume that the production function is such that the other workers cannot compensate for the lost team members; they may not have their skills, or full production may require two extra pairs of hands.) The cost of the absence would be \$20, or the value of the lost output. Alternatively, one can think of the cost of the absence as the sum of the marginal product of the absent worker ($1 \times \$6$), the marginal product of the idle type-one workers ($2 \times \$4$), and the marginal product of the idle capital ($\frac{2}{3} \times \$9$) caused by the absence of the type-two worker. With team production, the cost of an absence (\$20 in this example) can clearly exceed the wage of the absent factor (\$6 in this example). When setting annual compensation, the firm should reduce each team member's compensation by the expected cost of his absences. If a type-two worker is expected to miss 5 days per year, then their annual pay should be \$100, or ($5 \times \$20$), less than if he were present every day.

The value of the entire team's lost production, \$20 in the above example, represents an upper bound to the cost of an absence under team production. The firm's managers might be able to take steps to mitigate the effects of an unexpected absence. One possibility is for the firm to substitute a new worker for the absent type-two worker, thereby allowing the remaining members of the team to continue to produce some output. Whether it is worthwhile to hire a substitute depends on the substitute's impact on output and her cost. Consider a situation where the substitute (L_3) is less productive because she does not have the team-specific skills of the team member she is replacing (L_2), but her presence does not alter the

productivity of the other inputs. The 'replacement' team's production function becomes:

$$Y = \min\{aL_1, hL_3, dK\}$$

where h is the productivity parameter of the replacement worker. The new cost function becomes

$$C = Y(w_1/a + w_3/h + v/d)$$

where w_3 is the replacement worker's wage.

When a firm decides whether to hire the replacement worker, w_1 and v are sunk costs; the team members made idle by the absence of the type-two worker will be paid whether or not a replacement is hired. The replacement worker will therefore be hired if $w_3/h < p$, where p is the output price. In the example above, if the replacement worker's wage is \$6 and p is \$10, then the replacement worker will be hired if $h > 0.6$. That is, if the replacement worker is at least 30% as productive as the absent worker, then the replacement will be hired. If this condition does not hold, the firm may choose to invest money to cross-train employees so they can more ably fill in for absent team members. This would imply that the parameter h is a function of the amount of money the firm invests in team-specific training. Replacing absent team members will be less likely to occur in a situation where there is a substantial amount of team-specific human capital. More generally, the cost of an absence will be the incremental cost of hiring a sufficient number of replacement workers to maintain the team's target output. If it is optimal to hire a replacement worker, then the expected cost of an absence will be reduced, and the team members' wages will rise relative to a situation where the non-absent team members remain idle.

The Leontief production function described above is an extreme form of team production where workers of different types cannot be substituted for one another. With less rigid team production, a firm could pay team members overtime when a team member is absent to ensure that output is maintained at Y^*. Alternatively, the firm could build into each team member's salary sufficient compensation for them to agree to work harder when a team member is absent in order to maintain output at its required level. In either case, the cost of an absence must be larger than the wage of the absent worker and smaller than the value of the reduction in the team's output caused by the absent worker. If team members are willing to permanently cover for a worker for less than the worker's wage, then the firm will replace the worker and raise the remaining team members' wages.

We conclude, therefore, that when there is a team production and substantial team-specific human capital, the value of lost output to the firm from an absence will exceed the wage per day of the absent worker. The loss could be as large as the total output of the team, but could be reduced if replacements are available who are either inexpensive or are reasonably close substitutes for team members.

Penalties for output shortfalls

The other situation where the cost associated with work loss may be much larger than the wage of the absent worker is when a firm incurs a penalty if output falls below a critical level for a given time interval (e.g. a day). Consider a firm that uses a single homogeneous labor input in an individual (non-team) production process and incurs a penalty for output shortfalls. For example, department store sales clerks answer customer questions and process sales. If one clerk is absent and no replacement is available, queues will form at the functioning registers and some potential customers might leave. Disappointed customers not only do not buy on a day they cannot be served, but they might take their future business elsewhere. If a replacement worker can be hired at the same wage as the absent worker, then the cost of the work loss will be the wage per day paid, as before. If, however, the absent worker has substantial firm-specific human capital and is costly to replace, the cost to the firm of the absence will be greater than the sales clerk's wage.

The costs associated with work loss can be particularly large at service firms where the inventory is often perishable. If an airline flight is cancelled because a pilot is absent, for example, the airline will never be able to recoup the lost revenue. The department store, on the other hand, can still sell the inventory that accumulated when the sales clerk was absent. In both cases, dissatisfied customers can affect future revenue so there is a penalty associated with output shortfalls. Managers can take steps to minimize the likelihood of an output shortfall. In the case, where output can be inventoried so that work can be deferred until later, there will still be costs if there

are penalties for holding higher inventories; work-loss is the enemy of 'just in time' processes. But the costs will be larger still if many customers refuse to reschedule.

Consider first a firm that is too small to hire extra workers in anticipation that some fraction $m/250$ will be absent each day, on average. Therefore, it is not costless to find a perfect substitute. The penalty that we analyze is a situation where the firm must pay its workers overtime wage of $(1+\alpha)w$ to remain after their shift to fill in for any absent colleagues in order to ensure that Y^*, the target output, is produced each day.[c] Over the course of the year, the firm will expect to incur overtime costs of $m(1+\alpha)w$ due to worker absences, where m is the expected number of absences per worker per year. This cost will be subtracted from each worker's annual marginal product when determining the wage per day paid.[d] The value to the firm of reducing absenteeism by one day per worker is $(1+\alpha)w$. The cost of an absence is the marginal labor cost (wage plus overtime) of a worker, which is higher than the average wage per day.

Larger firms under pressure to produce a certain output each day may be able to mitigate the effects of absenteeism. If workers were never absent, then a firm would hire L^* workers to produce an output of Y^*. The actual number of workers present on a given day is $L = L^s - a$, where L^s refers to the number of scheduled workers and a is the number of workers who are absent. If there were no penalty associated with producing less than Y^* and output could be stored at zero cost, then the firm would hire $L_{\min} = (1+m/250)L^*$ workers, where $m = E[a]$ – the number of days that each worker is expected to be absent during a 250-work-day year. If m were 10 and L^* were 100, for example, a firm would hire 4% more workers than needed (104 instead of 100) in order to produce Y^*. Over a long time period, L^* workers would be present for work, on average.

If the penalty associated with having fewer than L^* workers each day is sufficiently large, then a firm might hire staff in excess of $L_{\min}$. The probability that a firm will have a shortage of workers on a particular day is

$$\begin{aligned}\Pr(L^s - a < L_{\min}) &= \Pr(a > L^s - L_{\min}) \\ &= [1 - G(a, L^s - L_{\min})]\end{aligned}$$

where G is the cumulative distribution function for absences. The number of workers who are absent each day, a, is a random variable with a mean of m and variance σ^2. The variance in the number of workers who are present, which will be a decreasing function of L^s, will dictate the magnitude of overtime costs. A firm that schedules a large number of workers will have less variance in the number of workers who are present each day relative to a small firm, and therefore, will be less likely to incur a penalty (overtime costs in this case).

Conditional on Y^*, the firm's objective is to choose L^s, the number of scheduled workers per day, to minimize production costs (production costs if there were no absences would be $wL_{\min}$)

$$\min_{L^s} \; w(L^s - L_{\min}) + [1 - G(a, L^s - L_{\min})](1+\alpha)w \qquad (1)$$

The firm will hire workers until $g_L = 1/(1+\alpha)$, where g_L is the effect of scheduling an additional worker on the expected number of absent workers. If overtime wages are twice as high as regular wages ($\alpha = 1$), then a firm will hire an extra worker if he reduces the expected number of absent workers by at least 0.50. The cost to the firm of an absence will be greater than w but smaller than $(1+\alpha)w$ (or else it would not have hired any excess staff).

Now consider the value of a program that reduces the absences for all workers by 1 day at a firm that has been incurring overtime costs. When m, the expected number of absences, is reduced by one, the firm will adjust $L_{\min}$ downward. If 104 workers had been hired in hope that 100 workers would show up, fewer than 104 will now be hired. In expression (1), both the expected value of a and $(L^s - L_{\min})$ will decrease. Overtime costs should also decrease because the variance of workers who are present each day will fall when m falls. Table 1 presents a simple analysis of a firm that hires ten workers. Column one presents the probability distribution for the number of workers who will show up each day assuming that each worker is absent 4% of the time and absences are independent across workers. For example, the probability that all ten scheduled workers show up is 0.67. The variance of the expected workforce, which is a more important determinant of expected overtime costs than the mean number of absences, is presented in the bottom row of Table 1. When absenteeism is reduced to 3% per day per worker (a decrease of m greater than 1), the variance of

Table 1. Probability distribution of employees who are present at a firm with ten scheduled workers

No. of employees who are present	Daily absence rate of 4%	Daily absence rate of 3%
10	0.665	0.737
9	0.277	0.228
8	0.052	0.032
7	0.006	0.003
6	0.0004	0.0001
Variance of present workers	0.620	0.539

workers falls by about 15%. This reduced variance will result in lower overtime costs.

So far we have assumed that absences are exogenous, and cannot be affected by a firm's sick leave policy. More realistically, firms that incur substantial penalties when output varies should consider implementing stricter sick leave policies to reduce the cost of absenteeism. The workers would benefit from such a policy in a competitive market through the form of higher wages.

We conclude that if expected absenteeism is reduced by 1 day per worker at a firm that had been paying overtime, the firm will respond by reducing staffing and/or reducing overtime expenses, and the marginal benefit of the reduced absenteeism will be between the daily wage and the overtime wage. For the particular case described above the homogeneous output but time-sensitive demand, we can conclude that, other things being equal, large firms will suffer less from the work loss of a given average or expected amount than will smaller firms. What is truly relevant here, however, is not firm size *per se* but rather the size of the production unit. A large airline or a large hotel chain with small geographically isolated local units or offices will have the experience of a small firm.

The full-employment case: summary

In some firms, the average wage per day can be a reasonably accurate measure of the cost of lost work time and the benefit of reducing lost work time. In other firms and in other situations, however, the wage will substantially underestimate the cost of lost work time. When will the cost of lost work time exceed the wage and when will the divergence between the two measures be large?

There are three factors that dictate whether the wage will be an accurate approximation of the cost of lost work time. Firms, or production units within firms, can be arrayed along three dimensions, as in Figure 1: the degree to which production is team oriented rather than individual oriented, the cost of replacing an absent worker, and the magnitude of the penalty associated with an output shortfall.

The wage will be a good estimate of the cost of work loss when an absent worker can be replaced with an equally productive substitute at the same wage. This is true even if production is team-oriented and there are large penalties associated with not meeting an output target. If there is team production but no team-specific human capital, then absences can be offset by calling in perfect substitutes. Even if there is a firm-specific human capital, a large enough firm can create a reserve pool of employees familiar with company policies and procedures that can fill in. The cost of replacing absent workers will be relatively large at small firms that cannot afford to create a reserve pool to replace absent workers, firms with substantial firm-specific human capital, and firms that use teams with substantial team-specific human capital.

The first situation where the cost of work loss will exceed the wage is when a firm uses an individual production process, incurs a penalty if they miss an output target, and perfect substitutes are not available to replace absent workers (Case A in Figure 1). For example, a reservation call center that is understaffed due to absences will potentially alienate customers and damage the firm's reputation. If the firm takes no steps to minimize the likelihood of incurring the penalty, the cost of work loss is assumed to be $(1 + \alpha)w$ per worker. The firm might respond to this penalty by overstaffing to minimize the likelihood of an output shortfall, or by paying overtime to its workers to work double shifts when necessary. These actions will reduce the cost of work loss. The divergence between the wage and the actual cost of lost work time will be determined by (1) the size of the penalty for unsatisfied demand; (2) the size of the firm/unit and the variance of firm output relative to its target; and (3) the quality and price of substitute inputs (temps) or strategies (overtime).

The second situation where the cost of work loss will be larger than the wage is when there is a team production and perfect substitutes are not

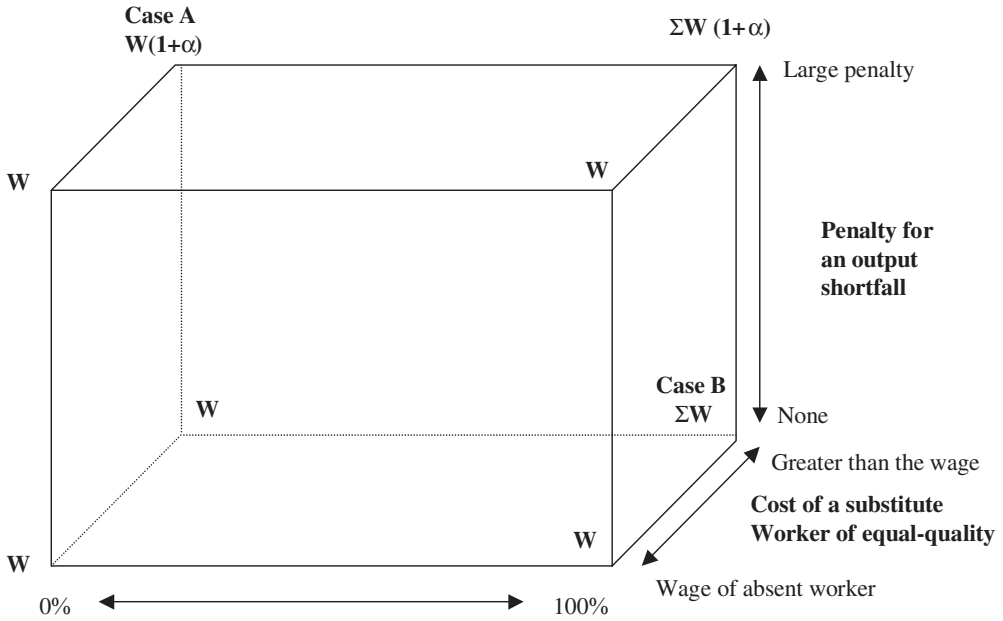

Figure 1. Firm characteristics that cause the cost of work loss to be greater than the wage (cost of work loss is displayed in **bold** at each vertex)

available to replace absent workers (Case B in Figure 1). The maximum cost of work loss in such a situation would be the sum of the factor prices of all the inputs, including capital (ΣW in Figure 1). This would occur when it is prohibitively expensive to replace an absent team member so the optimal response is to allow the entire team to be idle when one member is absent. If the firm is able to replace the absent worker with a less productive substitute, then the cost of work loss will be smaller than in the extreme case. The divergence between the wage per day and the actual cost of lost work time in Case B will be determined by (1) the impact of a team member's absence on team production; and (2) the quality and price of substitute inputs, which are related to the amount of team-specific human capital.

Firms might have all three characteristics – team production, difficulty substituting for absent workers, and penalties for output shortfalls – which will experience the largest costs of lost work. An example might be Boeing, which uses teams to produce airplanes. If Boeing misses a delivery deadline it can still sell the planes, but its reputation is damaged and future sales may be affected. While not all firms will have the crucial combinations of characteristics described above, there are many firms that are like Case A or B.

These concepts are related to the notion of 'friction costs' – the cost of replacing an absent worker – described by Koopmanschap *et al.* [10]. However, in a model of competitive labor markets in (full employment) equilibrium, friction costs are an addition to the lost wage or productivity measures, not a substitute for them. Moreover, if workers make up lost output later in what would otherwise be their spare time (or through a faster pace of work) there is both a cost to society – the value of sacrificed leisure or effort, which could even exceed the wage – and a cost to the employer in the form of higher annual wages for jobs that require such make up or speed up (compared to jobs that do not). Even if physical output (and therefore measured GDP) do not decline when illness strikes, wellbeing does fall, so that a proper measure of welfare (and a proper measure of the value of human capital) should be affected. Thus, lost leisure is a cost that should

be considered, most especially if one takes a correct societal perspective than encompassing all costs and benefits that matter to the citizens [4].

Less than full employment

How is this model altered when there is (non-transitional) involuntary unemployment? In this case, the employer, employee, and societal perspectives diverge. The societal value of avoiding a lost workday, though almost always positive, can fall below the prevailing money wage. Sticky money wages or imperfect macroeconomic policy may cause the prevailing wage to be higher than the level that equates the quantity of labor demanded to the quantity supplied.

In such a case, the marginal revenue product of labor will still equal the money wage, but it will exceed the opportunity cost of labor (the marginal value of leisure). From the societal perspective, the cost of lost work time is now less than the wage. However, the cost of lost work is obviously not zero; instead, it equals the (marginal) reservation wage unemployed workers would require to sacrifice leisure for work. In such a case, even if the loss to an employer from a lost workday exceeds the money wage (e.g. because of team production), the loss to society could be less than the money wage.

In the firm-specific health capital case, however, the employer will still value avoiding lost work time at the same level as in the full employment case. Programs to reduce work loss may still add to profits.

In the case of general health capital, the annual wage that would be paid to healthier workers is difficult to specify since the money wage is no longer being determined in a competitive labor market equilibrium. The question is whether, in a time of unemployment, a set of healthier workers could find a market for the additional labor time they can supply. Probably, the additional payment would be positive but would be less than the additional marginal revenue product.

The size of the difference between the prevailing money wage and the (lower) wage that would equilibrate the labor market determines the extent of the divergence between the private (employer) cost of work loss and the societal cost. If the unemployment rate is only slightly elevated above the transitional level, then the perspectives will still be fairly similar and our earlier analysis will apply. There could be a substantial divergence in the perspectives if the unemployment rate is high. However, in that case not only should lost work time be measured differently from a societal perspective, so also should all other medical 'costs'. For instance, the societal opportunity cost of a hospital stay will not be properly measured by its accounting cost, since money wages for hospital workers will substantially exceed the opportunity cost of their input. In the world of high unemployment, market prices no longer furnish a valid compass for the societal perspective.

Our discussion here should be distinguished from the treatment of lost wages as 'indirect costs' in cost-effectiveness analysis (CEA) studies. In principle, the consequences of reduced workloss can be included either in the numerator or denominator of a 'cost-utility' version of CEA, although Gold *et al*. prefer to include them in the utility effect, presumably by converting wage gains into utility or QALYs, rather than in the cost component of the ratio [11]. Our argument here is concerned only with monetary valuation of the wage gains. In the full-employment case, they should be valued at the money wage or greater. In the less-than-full-employment case, they should be valued at the reservation wage (equal to the value of leisure foregone), which will often be less than the market wage. In both cases, the wage gains will need to be added on to the QALYs measure in some fashion if sick leave benefits are provided to workers, but should not be added if no such benefits are furnished, in order to avoid 'double counting' [12].

Conclusion

This discussion leads to three conclusions. First, the productivity gains from programs or medical interventions that reduce absenteeism due to illness are very likely to be larger than the wage per day or per hour. Second, the incidence of net benefits from such programs (value of productivity gain less program cost) is likely to fall largely on workers in the long run, but on employer profits in the short run. The incidence will shift more rapidly to workers when it is easy for employers to identify employees with improved health. And third, employers who initiate work loss-reduction

programs with positive net benefits will be rewarded, regardless of the incidence of these gains.

Notes

a. The word 'almost' is necessary because a substantial increase in the supply of productive labor will, depending on the elasticity of the demand for labor, potentially reduce the wage slightly. Still, there will be substantial gains to workers.
b. The employer may also retain the gains if the health intervention affects 'inframarginal' workers who, because of seniority or high moving costs, will not actively seek a different job.
c. Presumably, the firm pays overtime rather than allowing customers to walk away because the overtime pay penalty is less than the lost business/bad reputation penalty, or a penalty from bringing in expensive temporary workers who do not know the job.
d. Workers should recoup this amount in overtime payments when they fill-in for their absent colleagues.

Acknowledgements

This project is funded by a grant from Merck & Co., Inc. The views expressed in this paper are those of the authors and not of Merck & Co., Inc.

References

1. Greenberg PE, Finkelstein SN, Berndt ER. Economic consequences of illness in the workplace. *Sloan Manage Rev* **36**(4): 26–38.
2. Conti DJ, Burton WN. The economic impact of depression in the workplace. *J Occup Med* 1994; **36**(9): 983–988.
3. Greenberg PE, Stiglin LE, Finkelstein SN, *et al*. Depression: a neglected major illness. *J Clin Psych* 1993; **54**(11): 419–423.
4. Garber AM, Weinstein MC, Torrance GW, *et al*. Theoretical foundations of cost-effectiveness analysis. In *Cost Effectiveness in Health and Medicine*, Gold MR, *et al*. (eds). Oxford University Press: New York, 1996.
5. Rice DP, Miller LS. The economic burden of affective disorders. In *Advances in Health Economics and Health Services Research*, vol. 14, Scheffler RM (ed.). JAI Press: Greenwich, CT, 1993; 37–53.
6. Ungar WJ, Coyte PC, the Pharmacy Medication Monitoring Program Advisory Board. Measuring productivity loss days in asthma patients. *Health Econ* 2000; **9**(1): 37–46.
7. Hodgson TA, Meiners MR. Cost-of-illness methodology: a guide to current practices and procedures. *Milbank Q* 1982; **60**(3): 429–462.
8. Becker GS. *Human Capital: a Theoretical Analysis with Special Reference to Education*. Columbia University Press: New York, 1964.
9. Pauly MV. *Health Benefits at Work: an Economic and Political Analysis of Employment-Related Health Insurance*. University of Michigan Press: Ann Arbor, MI, 1997.
10. Koopmanschap MA, Rutten FFH, van Ineveld BM, *et al*. The friction cost method for measuring indirect costs of disease. *J Health Econ* 1995; **14**(2): 171–189.
11. Gold M, Siegel JE, Russell LB, *et al*. (eds). *Cost-Effectiveness in Health and Medicine*. Oxford University Press: Oxford, 1996.
12. Johannesson M. Avoiding double-counting in pharmacoeconomic studies. *Pharmacoeconomics* 1997; **11**(5): 385–388.

