Psychological Medicine, 2004, **34**, 1001–1011. © 2004 Cambridge University Press
DOI: 10.1017/S0033291703001843 Printed in the United Kingdom

How accurate is recall of key symptoms of depression? A comparison of recall and longitudinal reports

J. ELISABETH WELLS* AND L. JOHN HORWOOD

Department of Public Health and General Practice and the Christchurch Health and Development Study, Christchurch School of Medicine and Health Sciences, New Zealand

ABSTRACT

Background. Assessment of lifetime major depression is usually made from a single interview. Most previous studies have investigated reliability. Comparison of recall of key symptoms and longitudinal reports shows the accuracy of recall, not just reliability.

Method. At age 25, 1003 members of the Christchurch Health and Development Study cohort were asked to recall key symptoms of depression (sadness, loss of interest) up to age 21. This recall was compared with longitudinal reports at ages 15, 16, 18 and 21 years. Diagnosis was by DSM-III-R and DSM-IV criteria.

Results. Only 4% of those without previous reports recalled key symptoms. Of those with a diagnosis of depression up to age 21, 44% recalled a key symptom. Measures of severity of an episode (number of symptoms, impairment, duration, suicidality) and chronicity (years with a diagnosis, years with suicidal ideation) all strongly predicted recall. Current key symptoms increased recall, even after taking account of severity and chronicity. Being female and receiving treatment also predicted recall, although odds ratios were reduced to 1·6–1·7 when all other predictors were included. Comparison of risk factors for key symptoms showed similar results from longitudinal reports and recall. Sexual abuse, neuroticism, lack of parental attachment, gender, physical abuse and maternal depression were major risk factors in both sets of analyses.

Conclusions. Forgetting of prior episodes of depression was common. Severity, chronicity, current depression, gender and treatment predicted recall. Lifetime prevalence based on recall will be markedly underestimated but the identification of major risk factors may be relatively little impaired.

INTRODUCTION

The accuracy of recall of earlier episodes of depression is important in two areas within psychiatric epidemiology, namely cross-sectional surveys and genetic epidemiology, because of the reliance on a single interview for a lifetime history. It is also relevant to clinicians taking a clinical history for a clinical study or as part of their practice.

* Address for correspondence: Dr J. Elisabeth Wells, Department of Public Health and General Practice, Christchurch School of Medicine and Health Sciences, PO Box 4345, Christchurch, New Zealand.
(Email: elisabeth.wells@chmeds.ac.nz)

For genetic and family studies it is essential to establish if a disorder has ever occurred, not just whether or not it is current, in order to determine the phenotype (Foley *et al*. 1998). Lifetime prevalence is required, namely the prevalence of disorder at any time up to interview. Some individuals may not have developed the disorder but might do so later; this censoring can be dealt with statistically through survival analysis or other modelling incorporating the period at risk of disorder. However, there is no statistical remedy for failure to recall episodes or symptoms.

For cross-sectional surveys a variety of prevalence periods can be used: the last month, the last year, or ever in lifetime so far. Recent

https://doi.org/10.1017/S0033291703001843 Published online by Cambridge University Press

prevalence alone cannot distinguish between factors which affect incidence and those which affect chronicity or recurrence. Nor can it provide information about cohort effects or life course or the extent to which past episodes influence help-seeking for recent episodes. Robins *et al*. (1984) have stressed the need for psychiatric history in order to make diagnoses. Consequently, interview schedules such as the Diagnostic Interview Schedule (DIS) (Robins *et al*. 1981) and the Composite International Diagnostic Interview (CIDI) (Robins *et al*. 1988), which have been widely used in psychiatric epidemiology surveys (Kohn *et al*. 1998), ask if symptoms have ever occurred. Similarly, the semi-structured clinician interviews such as the SCID (Williams *et al*. 1992) and the SCAN (WHO, 1989) enquire about psychiatric history (Robins, 1990).

Recall error has been investigated mainly through studies of reliability, both for structured lay interviews and for semi-structured clinician interviews, beginning with studies of inter-rater reliability (Keller *et al*. 1981) and short-term test–retest reliability (e.g. Robins *et al*. 1981) and subsequently studies of reliability over longer periods (e.g. Bromet *et al*. 1986; Fendrich *et al*. 1990; Kendler *et al*. 2001) including up to five or six years (Prusoff *et al*. 1988; Rice *et al*. 1992; Foley *et al*. 1998). Reliability over longer periods varies with the group studied (patients or relatives or community samples), the base rate of disorder, and the length of the interval, but is seldom high. For example, Kendler and colleagues (Foley *et al*. 1998; Kendler *et al*. 2001) have found kappas of 0·43 and 0·48. The discrepancies are due primarily to lower rates of reporting at follow-up (e.g. Rice *et al*. 1992), indicating forgetting rather than uncertainty.

There are far fewer longitudinal studies that have looked at recall of specific episodes or have followed people over several waves. In a small study of 27 patients re-contacted after 25 years, Andrews *et al*. (1999) found that 70% of these people hospitalized for a depressive episode could recall key symptoms of sadness or loss of interest but only half recalled enough symptoms to satisfy diagnostic criteria. Wilhelm & Parker (1994), with follow-up of teacher trainees after 5 and 10 years, reported that men were more likely than women to forget episodes. Aneshensel *et al*. (1987) repeatedly interviewed a community sample over four years and found that the predominant type of inconsistency in reporting of lifetime prevalence arose from failure to subsequently report earlier reports. Because of the early age of onset of many depressive episodes (Hankin *et al*. 1998), even at first interview some individuals in these adult cohorts will have already had episodes many years earlier.

To assess accuracy of recall, rather than just reliability, requires longitudinal data beginning before the likely age of onset. Accordingly, this study uses data from the Christchurch Health and Development Study (CHDS), a cohort of New Zealand children studied from birth to age 25. Detailed information on depressive symptoms and DSM diagnostic criteria for major depression has been gathered for the period from when they were 14 years old until they were 21. Sample members were subsequently asked at age 25 for their recall of key depressive symptoms prior to age 21. Therefore it is possible to compare longitudinal reports with recall at age 25 to obtain more direct information on the accuracy of recall than is possible with reliability studies or other studies of adults. The aims of this study were:

(1) To investigate the level of recall at age 25 of key symptoms of depression experienced prior to age 21.

(2) To investigate the factors affecting accuracy of recall amongst those who met criteria for major depression prior to age 21.

(3) To investigate the implications of using recall measures of depressive symptoms rather than longitudinal reports for estimates of risk factor associations.

The predictors of recall examined in this study were those found in reliability studies: the severity and chronicity of depressive symptoms (Bromet *et al*. 1986; Rice *et al*. 1992; Williams *et al*. 1992; Foley *et al*. 1998; Kendler *et al*. 2001), gender (Wilhelm & Parker, 1994), and a history of treatment (Fendrich *et al*. 1990; Foley *et al*. 1998; Kendler *et al*. 2001).

METHOD

Sample

Participants were members of a birth cohort, born in a 4-month period in 1977 in the Christchurch Urban Area in New Zealand. The

https://doi.org/10.1017/S0033291703001843 Published online by Cambridge University Press

CHDS has followed this cohort with data collection at birth, 4 months, 1 year, annual intervals to age 16, and then at 18, 21 and 25 years. Data were collected as soon as possible after the participant's birthday. Fergusson *et al.* (1989) give an overview of the design of this study.

Interview

Major depression

At each assessment from age 15 onwards, sample members were interviewed by trained survey interviewers on a structured mental health interview designed to assess mental health and adjustment over the period since the previous assessment. The interviews obtained information on the diagnostic criteria for the assessment of major depression (Fergusson *et al.* 1999). The questions at ages 15 and 16 combined elements of the Diagnostic Interview for Children (Costello *et al.* 1982) and the Diagnostic Interview Schedule (Robins *et al.* 1981) and yielded DSM-III-R diagnoses. Subsequent interviews used questions from the Composite International Diagnostic Interview (WHO Division of Mental Health, 1993) and yielded DSM-IV diagnoses. In each interview participants were questioned about depression over the last month, the previous year, and then the period, if any, back until the previous assessment. For each depressive episode indices of severity included the number of DSM symptoms, the duration of the episode, and impairment in each of seven areas of life functioning. The impairment measure used was the number of areas of functioning reported with 'a great deal' of impairment. Duration and impairment were not recorded at age 15 and age 16 interviews. Participants who met criteria for depression only at ages 15 and 16 were allocated to the lowest categories on these variables ($n=26$). Help-seeking for depression was also recorded (consulting a doctor, psychologist, or counsellor).

Recall of key symptoms of major depression

A diagnosis of major depression cannot be made without symptoms of sadness or loss of interest. Two questions about these symptoms were asked at the start of each period to be reported on. If answers to both questions were negative there were no further questions about that period. At age 25 there were two additional questions which asked:

'Looking back over your whole life before you were 21, did you ever have a period of at least two weeks when you (*a*) felt sad, blue or depressed nearly every day? (*b*) lost interest in most things like work, hobbies or things you usually enjoy?' Sample members who responded affirmatively to either of these items were also asked to report the estimated age of onset of these symptoms.

Suicidal behaviours

From the age 15 interview on, all participants were asked about suicidal ideation and suicide attempts. Questioning about suicidal behaviours was conducted separately from questions on the relevant DSM symptom criteria within a depressive episode. At each interview, participants were asked to report for each 12-month period since the previous assessment whether they had experienced suicidal thoughts, planned suicide or made a suicide attempt (Fergusson & Lynskey, 1995).

Risk factors for depression

A range of social, family and individual factors were selected from the database of the study because they were known from previous analyses of the CHDS data or had been suggested in the literature as being potential correlates of depression.

Family socio-economic status (*SES*)

Family SES was assessed at the time of the survey child's birth, using the Elley–Irving scale of socio-economic status for New Zealand (Elley & Irving, 1976). This scale classifies families into six levels on the basis of paternal occupation. For the present analysis this scale was reduced to a three-level classification: high (professional, managerial); middle (clerical, technical, skilled); low (semi-skilled, unskilled, unemployed).

Parental history of depression or anxiety

When sample members were aged 15 years, parents were questioned about their history of problems with depression or anxiety. This questioning was based on parental self-definition of internalizing problems rather than standard diagnostic criteria.

https://doi.org/10.1017/S0033291703001843 Published online by Cambridge University Press

Maternal depression (*7–13 years*)

At each year from when sample members were aged 7–13 years, a measure of current maternal depressive symptomatology was obtained using a modified form of the Levine–Pilowsky Depression Inventory (Pilowsky *et al*. 1969; Pilowsky & Boulton, 1970). This scale comprised 37 dichotomous items and had high internal consistency with coefficient alpha values in excess of 0·97 in each year.

Cognitive ability (*8 years*)

When sample members were aged 8 years, child cognitive ability was assessed using the revised Wechsler Intelligence Scale for Children (WISC-R; Wechsler, 1974). For logistic reasons the assessment of IQ was restricted to the sample of children who were resident in the Canterbury (New Zealand) region. The full-scale IQ score was used in the present analysis. Reliability, assessed using split-half methods, was 0·93.

Leaving school without qualifications

School Certificate is a national series of examinations that are attempted by most New Zealand children at the end of their third year of secondary school. Sample members who left school without attaining at least one pass grade in School Certificate and who did not attain any higher-level qualification were classified as having left school without qualifications.

Parental attachment (*14 years*)

The quality of parental attachments during adolescence was assessed using the Armsden and Greenberg Parental Attachment Scale (Armsden & Greenberg, 1987), administered when sample members were aged 14 years. Reliability, assessed using coefficient alpha, was 0·91.

Child neuroticism (*14 years*)

This was assessed using a short form version of the neuroticism scale of the Eysenck Personality Inventory (Eysenck & Eysenck, 1964) administered when sample members were aged 14 years. Reliability, assessed using coefficient alpha, was 0·80.

Childhood sexual abuse

At age 18 and 21 years, sample members were questioned about their experience of childhood sexual abuse prior to the age of 16 years (Fergusson *et al*. 1996). For the present analysis, sample members who reported any form of sexual abuse, at age 18 or 21, were classified as having experienced childhood sexual abuse.

Childhood physical abuse

At ages 18 and 21, young people were asked to report on the extent to which their parents used methods of physical punishment during their childhood (prior to age 16 years) (Fergusson & Lynskey, 1997). For the present analysis sample members were classified as having experienced childhood physical abuse if they reported, at age 18 or 21, that either parent had used physical punishment regularly or had used a more severe or harsh form of physical punishment.

Statistical analyses

All analyses were carried out using SAS 8.02. In modelling recall of key symptoms in cohort members who had met criteria for depression up to age 21, the effects of severity and chronicity were investigated first, to take account of the experience of depression. Then the effect of current state was looked at. Finally, gender and then treatment were added, to see what effect these had which could not be accounted for by severity, chronicity and current state. Two measures were used to compare the adequacy of the fitted models. The Akaike Information Criterion (AIC) enables models to be compared taking account of the number of variables in the model. With a binary outcome,

$$\mathrm{AIC} = -2\,\mathrm{Log}\ L + 2(1+s),$$

where Log L is the log likelihood of the fitted model and s is the number of variables in the model (Stone, 1998; SAS Institute Inc., 1999). The model with the lowest absolute value of AIC is the most parsimonious. R^2 is a generalization of the usual coefficient of determination for linear regression. For logistic regression the maximum is less than one so Nagelkerke has developed a rescaled R^2 which does have a maximum of 1 (Nagelkerke, 1991; SAS Institute Inc., 1999). With AIC and R^2 it is possible to see the gains, if any, of adding variables to a model. In addition, the range of outcomes predicted from the lowest to the highest was calculated for some models to

https://doi.org/10.1017/S0033291703001843 Published online by Cambridge University Press

Table 1. *Percentage who experienced key symptoms* (*sadness or loss of interest for 2 weeks*) *or met criteria for a major depressive episode in the year before each interview*

Age at interview	Percentage meeting criteria for depression	Percentage with at least one of the key symptoms	*n*	Percentage of birth cohort (*n* = 1265)
15	5·1	13·5	965	76·3
16	6·8	15·0	953	75·3
18	18·2	25·5	1025	81·0
21	18·2	26·2	1011	79·9

All interviews available for a given age were used. Interviews were as close as possible after a birthday.

indicate how well the models discriminated between individuals.

Sample size and missing data

The analyses in this paper are based upon the 1003 sample members who were assessed at age 25 years. This sample represents 79% of the original cohort of 1265 children who entered the study. The number of young people interviewed in assessments from age 15–21 years has ranged from 953 to 1025 (see Table 1). To compare longitudinal reports of depression with recall of key symptoms requires an assessment of the extent and pattern of missing interviews. Nearly all of the 1003 interviewed at age 25 were also interviewed from age 15 onwards (91% had been interviewed at ages 15 and 16 and 97% and 98% at ages 18 and 21 respectively). Given the low numbers of missing interviews, and the lower prevalence of depression at the earlier ages when there was more missing data, for the present analysis summaries of longitudinal reports used any evidence of key symptoms or depression from the age 15- 16- 18- and 21-year interviews. This meant that all those interviewed at age 25 were included in analyses.

RESULTS

History of depressive symptoms (14–21 years)

To provide an overview of changes in depressive symptomatology in the cohort from age 14–21 years, Table 1 shows the proportion of participants at each interview who reported key symptoms of depression (sadness or loss of interest) in the past year, and the proportion who met criteria for major depression in that year. All available participants were used at each time-point. For both outcomes the pattern is of a steep increase in depression up to age 18, which then remains steady up to age 21. Combining data from all years, including the 'in-between interview' ages of 17, 19 and 20 years, revealed that 54% of the sample reported one of the key symptoms of depression (sadness, loss of interest) in the period from age 14–21 years and 37% met DSM criteria for major depression.

Recall of key symptoms

Only 4% of those without reports of key symptoms prior to age 21 recalled such a key symptom at age 25 (20/462). Two had missed interviews and six gave an age of onset before age 14, so could have been reporting from periods not included in the dataset. Thus there were very few new reports of key symptoms. Among those with a prior history of depressive symptoms, recall was extremely poor. Recall was only 22% for those with symptoms alone (36/167), and only 44% for those who had met criteria for a major depressive episode (165/374). These results show that it was very common for participants with prior reports of symptoms to fail to recall them.

Predictors of recall

A series of analyses was conducted to investigate the predictors of recall of key symptoms in the 374 participants who had met criteria for major depression in the period 14–21 years. The first set of analyses investigated measures of severity and chronicity of depressive symptomatology over the period from 14 to 21 years as predictors of recall. Table 2 shows that for all measures there were strong and highly significant tendencies for rates of recall to increase with increasing severity or chronicity. The strongest association was with years of suicidal ideation: recall ranged from 29% in those without any suicidal ideation to 78% in those with three or more years of ideation.

To investigate the combined effects of severity and chronicity on the prediction of recall, three logistic regression models were fitted to the data: a model including the severity measures only (Model 1); a model including the chronicity measures only (Model 2); and a model including all factors (Model 3). The results of these

https://doi.org/10.1017/S0033291703001843 Published online by Cambridge University Press

Table 2. *Severity and chronicity predictors of recall at age 25 of key symptoms of depression at or before age 21: percentage recall (n = 374 who met criteria for major depression from age 15 to age 21 interviews)*

Predictors	Levels	*n*	Percentage recall of key symptoms[a]	*p*[b]
Severity				
Maximum number of symptoms in an episode	5–6	64	25	<0·0001
	7	82	26	
	8	107	50	
	9	121	62	
Maximum duration of an episode	2–3 weeks	137	28	<0·0001
	4–7 weeks	103	44	
	8+ weeks	134	60	
Maximum number of areas of severe impairment	0–1	125	32	0·0002
	2	93	40	
	3	83	52	
	4+	73	62	
Suicidal behaviour	None	184	29	<0·0001
	Ideation only	125	55	
	Attempt	65	66	
Chronicity				
Years meeting criteria for major depression[c]	1	179	28	<0·0001
	2	109	51	
	3+	86	67	
Years with suicidal ideation	0	184	29	<0·0001
	1	62	48	
	2	87	57	
	3+	41	78	

[a] Symptoms for at least 2 weeks of sadness/depression or loss of interest.
[b] Based on χ^2 tests for contingency tables.
[c] Years 19 and 20 count as only 1 year because of the form of the interview.

analyses are summarized in Table 3. Severity measures were moderately related to each other, particularly the number of symptoms, duration and impairment. Consequently the joint model of severity (Model 1) fitted well but accounted for little more than the univariate analyses. Furthermore, odds ratios (OR) were reduced substantially in comparison to the univariate analyses, with impairment adding nothing to the other severity measures. In contrast, the two measures of chronicity were less interchangeable, with a number of participants reporting no suicidal ideation even with two or three years with a diagnosed depressive episode. Therefore both variables remained strong predictors of recall in a joint model (Model 2). The combined model including both severity and chronicity (Model 3) produced only very slightly better prediction than Model 2 with chronicity alone. Someone with only one year with an episode of major depression and no suicidal ideation had an observed probability of recall of 22% (23/118) and a predicted probability from Model 2 of 19%, whereas for someone with depression for three years and three years of suicidal ideation the observed probability was 91% (21/23) and the predicted was 84%. Using Model 3, contrasting the prediction for someone with the lowest risk on all variables and someone with the highest risk on all variables yielded predicted probabilities of 19% *versus* 86%, which differs little from the predictions for a model including only the chronicity indices.

As expected from reliability studies, current depressive symptomatology at age 25, gender, and history of treatment seeking prior to age 21 were found to have statistically significant univariate associations with recall (Table 4).

It was expected that the association between current symptoms and recall would be explained entirely by the severity and chronicity of past depressive episodes: those with current symptoms would be those with a worse history and their higher recall would be a function only of this history. However, the OR for current symptoms did not decline and remained statistically significant (OR 2·4, $p = 0·02$) once the measures of severity and chronicity were entered into the model. This suggests that there was an effect of current state *per se* on recall of key symptoms.

Young women were more likely than young men to recall key symptoms of depression up to age 21 (49% *v.* 34%, $p = 0·004$), even though all participants had previously met criteria for major depression. Women had more years with a diagnosed depressive episode ($p = 0·005$). They began to experience depression earlier, both for diagnosed depression ($p = 0·05$) and for key symptoms ($p = 0·02$) but there was no difference in the recency of the last diagnosis, up to age 21 ($p = 0·77$). Nonetheless, taking account of prior severity, chronicity and current key symptoms reduced the gender OR only from 1·9 to 1·7. This OR remained at 1·6 (95% CI 1·0–2·7, $p = 0·06$) even when treatment-seeking was added to the model.

Those with a history of treatment-seeking prior to age 21 had substantially better recall than those who had not sought treatment

https://doi.org/10.1017/S0033291703001843 Published online by Cambridge University Press

Table 3. *Severity and chronicity predictors of recall at age 25 of key symptoms of depression at or before age 21: odds ratios and 95% confidence intervals (n=374 who met criteria for major depression from age 15 to age 21 interviews)*

Predictors	Levels	Univariate analyses	Model 1: severity	Model 2: chronicity	Model 3: severity + chronicity
Severity					
Maximum number of symptoms	5–6	**1**	**1**	—	**1**
	7	**1·0** (0·5–2·2)	**0·9** (0·4–0·9)		**0·7** (0·3–1·7)
	8	**2·9** (1·5–6·0)	**2·0** (1·0–4·3)		**1·6** (0·8–3·5)
	9	**4·9** (2·5–9·8)	**2·2** (1·0–4·9)		**1·5** (0·7–3·6)
	p[a]	<0·0001	0·03		0·10
Maximum duration of an episode	2–3 wk	**1**	**1**	—	**1**
	4–7 wk	**2·0** (1·1–3·4)	**1·3** (0·7–2·4)		**1·2** (0·6–2·2)
	8+ wk	**3·8** (2·3–6·4)	**2·0** (1·1–3·7)		**1·7** (0·9–3·1)
	p[a]	<0·0001	0·05		0·20
Maximum number of areas of severe impairment	0–1	**1**	**1**	—	**1**
	2	**1·4** (0·8–2·5)	**0·8** (0·4–1·6)		**0·8** (0·4–1·5)
	3	**2·3** (1·3–4·1)	**1·3** (0·7–2·4)		**1·1** (0·6–2·1)
	4+	**3·4** (1·9–6·3)	**1·2** (0·6–2·5)		**1·0** (0·5–2·2)
	p[a]	0·0003	0·60		0·77
Suicidal behaviour[b]	None	**1**	**1**	—	**1**
	Ideation	**3·1** (1·9–4·9)	**2·0** (1·2–3·5)		
	Attempt	**4·8** (2·7–9·0)	**2·7** (1·3–5·3)		**1·0** (0·5–2·1)
	p[a]	<0·0001	0·0007		0·97
Chronicity					
Years with depression[c]	1	**1**	—	**1**	**1**
	2	**2·7** (1·6–4·4)		**2·1** (1·2–3·5)	**1·7** (1·0–3·1)
	3+	**5·2** (3·0–9·1)		**3·3** (1·8–6·1)	**2·4** (1·2–4·6)
	p[a]	<0·0001		0·0002	0·03
Years with suicidal ideation	0	**1**	—	**1**	**1**
	1	**2·3** (1·3–4·2)		**2·0**(1·1–3·7)	**1·7** (0·9–3·3)
	2	**3·3** (2·0–5·7)		**2·5** (1·4–4·4)	**1·9** (1·0–3·6)
	3+	**8·8** (4·1–20·7)		**5·4** (2·4–13·2)	**4·0** (1·5–11·5)
	p[a]	<0·0001		0·0001	0·04
AIC[d]		476·3[e]	472·4	463·2	468·5
R^2 (rescaled)[d]		0·15	0·21	0·21	0·24

[a] The p value is for the variable as a whole, i.e. for the set of dummy variables covering all categories.
[b] In Model 3 the suicide variable was collapsed to avoid linear dependence with years of ideation.
[c] Years 19 and 20 count as only 1 year because of the form of the interview.
[d] See Method section.
[e] Maximum for a single variable, which was years with suicidal ideation.

(63% *v.* 35%, $p<0·0001$). Taking account of gender reduced the OR for treatment only a small amount, from 3·2 to 3·0. Taking account of severity, chronicity and current symptoms as well produced a substantial reduction to an OR of 1·7 (95% CI 1·0–2·9, $p=0·06$).

Overall, incorporating current symptoms, gender and treatment-seeking into a prediction model with severity and chronicity did not markedly improve levels of prediction ($R^2=0·27$) over the earlier models and AIC (463·4) was slightly worse than for Model 2 with only the chronicity variables. Nonetheless, the range of predicted probabilities increased (from 9% to 88%), because this full model could better distinguish those who did not recall key symptoms.

Risk factor analyses

The above analyses provide clear evidence of the poor recall of key symptoms of depression. Estimates of the lifetime prevalence of depression based on retrospective reports are likely to substantially underestimate the true lifetime prevalence of depression. Moreover, the accuracy of recall was found to vary substantially with severity, chronicity and other factors. Given such highly inaccurate reporting, it is therefore of

https://doi.org/10.1017/S0033291703001843 Published online by Cambridge University Press

Table 4. *Current key symptoms of depression (within last month at age 25 interview), gender and treatment as predictors of recall of key symptoms of depression at or before age 21 (n = 374 who met criteria for major depression from age 15 to age 21 interviews)*

	Current key symptoms of depression[a]		
Statistic	No	Yes	*p* value
n	330	44	
Percentage recall[a]	42%	61%	0·01
Univariate OR (95% CI)	1	2·2 (1·2–4·3)	0·01
OR (95% CI) in a model with severity and chronicity	1	2·4 (1·2–5·1)	0·02
	Gender		
Statistic	Male	Female	*p* value
n	127	247	
Percentage recall[a]	34%	49%	0·004
Univariate OR (95% CI)	1	1·9 (1·2–3·0)	0·004
OR (95% CI) in model with severity, chronicity, and current key symptoms of depression	1	1·7 (1·0–2·9)	0·04
	Treatment for depression up to age 21		
Statistic	No	Yes	*p* value
n	254	120	
Percentage recall[a]	35%	63%	<0·0001
Univariate OR (95% CI)	1	3·2 (2·0–5·1)	<0·0001
OR (95% CI) in model with gender alone	1	3·0 (1·9–4·8)	<0·0001
OR (95% CI) in model with severity, chronicity, current key symptoms and gender	1	1·7 (1·0–2·9)	0·06

[a] Symptoms of at least 2 weeks of sadness/depression or loss of interest.

OR, odds ratio; CI, confidence interval.

interest to examine the impact this may have on estimated associations between risk factors and depression.

Forgetting and differential forgetting are mechanisms which produce misclassification error. The impact of such error has been investigated in the theoretical methodological literature in epidemiology (Goldberg, 1975; Kuha *et al.* 1998). Non-differential error attenuates risk differences and ORs but differential error can produce bias in either direction. Since it is not possible to make general predictions about the impact of misclassification error it is necessary to make empirical comparisons for any given set of risk factors and outcomes.

Therefore a series of parallel analyses was conducted on the full sample to compare risk factor effect sizes obtained from longitudinal reports with those from recall of key symptoms of depression. ORs are reported as they are the usual measure of effect size in psychiatric epidemiology.

Table 5 presents ORs for a series of social, family and childhood risk factors for depression, using both recall and longitudinal reports of key symptoms as outcomes. The results of Table 5 are encouraging for the analysis of risk factors using recall. The major risk factors were the same for both recall and for longitudinal reports of symptoms: childhood sexual abuse, neuroticism, lack of parental attachment, gender, physical abuse and maternal depression. Total IQ at age 8 was unrelated to symptoms. There were small inconsistencies for SES at birth, parental history of depression or anxiety, and school qualifications, but all of these measures were at best only weakly linked to risks of key symptoms. Note that for gender the reduction in OR due to forgetting was compensated for by greater recall by young women so the net result was very similar ORs from recall and longitudinal reports. If recall had been non-differential the OR would have been only 1·5.

DISCUSSION

This study has shown recall of prior symptoms and episodes of depression to be poor. As expected from reliability studies (Bromet *et al.* 1986; Rice *et al.* 1992; Foley *et al.* 1998; Kendler *et al.* 2001) and longitudinal studies (Aneshensel *et al.* 1987; Wilhelm & Parker, 1994; Andrews *et al.* 1999), lifetime prevalence estimated from recall will be substantially lower than that obtained from longitudinal reports. The result is expected; the magnitude is perhaps surprising, although a similar reduction in lifetime prevalence was seen at Wave 2 in the Epidemiological Catchment Area studies (see Dohrenwend, 1990).

Reliability studies have shown some apparently 'new' reports of depression at second interviews but these studies have required full criteria for depression. Therefore the 'new' cases could have arisen from only a small change in recall of symptoms, moving the symptoms count above the criterion number, rather than from a *de novo* appearance of a previous depressive episode. This study showed

https://doi.org/10.1017/S0033291703001843 Published online by Cambridge University Press

Table 5. *Odds ratios for risk factors for key symptoms* (*prior reports or recall*; *n* = *1003*)

Variable	Level	*n*	Prior reports		Recall	
			OR	95% CI	OR	95% CI
Gender	Male	488	**1**		**1**	
	Female	515	**2·4**	(1·9–3·1)	**2·2**	(1·6–3·1)
SES at birth	High	207	**0·9**	(0·6–1·3)	**1·0**	(0·7–1·5)
	Middle	549	**0·9**	(0·7–1·3)	**0·7**	(0·5–1·0)
	Low	247	**1**		**1**	
Parental history of depression or anxiety	No	649	**1**		**1**	
	Yes	282	**1.2**	(0·9–1·6)	**1·3**	(1·0–1·8)
Maternal depression average score[a]	1 unit	985	**1.06**	(1·02–1·09)	**1·05**	(1·02–1·09)
Total IQ at age 8	10 points	778	**0·99**	(0·90–1·08)	**1·04**	(0·94–1·17)
School qualifications	None	182	**1**		**1**	
	Some	795	**0·7**	(0·5–1·0)	**1·0**	(0·7–1·4)
Sexual abuse	None	855	**1**		**1**	
	Some	146	**5·9**	(3·8–9·7)	**4·0**	(2·8–5·8)
Physical abuse	Minimal	823	**1**		**1**	
	Abuse	178	**1·8**	(1·3–2·5)	**1·8**	(1·3–2·6)
Neuroticism at age 14	1 unit	920	**1·16**	(1·12–1·21)	**1·11**	(1·07–1·15)
Parental attachment at age 14	1 unit	920	**0·96**	(0·94–0·97)	**0·96**	(0·94–0·97)

[a] Average Levine–Pilowsky score over the years when the cohort member was 7–13 years old.
OR, odds ratio; CI, confidence interval.

very few apparently new reports of key symptoms of depression at age 25.

The predictors of recall were those found previously from reliability studies. Characteristics of an individual's history of depression such as severity and chronicity were the most important predictors, as has been found in other studies (Bromet *et al.* 1986; Rice *et al.* 1992; Williams *et al.* 1992; Foley *et al.* 1998; Kendler *et al.* 2001). Kendler *et al.* (2001) found suicidal ideation to be one of the two individual symptoms which predicted consistent reporting; in the present study suicidal ideation was an important severity predictor, and years with suicidal ideation was the most important single predictor of recall. Treatment has been found to improve reliability (Aneshensel *et al.* 1987; Fendrich *et al.* 1990), even taking account of number of criteria, number of episodes and impairment (Foley *et al.* 1998; Kendler *et al.* 2001) with OR 1·9 and 2·4. In this study, treatment had a similar effect on recall (OR 1·7). In addition, as Aneshensel and colleagues found, current symptoms were also associated with recall.

Like Wilhelm and Parker's 10-year study of teacher trainees (Wilhelm & Parker, 1994), this study found that young women were more likely than young men to recall prior depression, and this difference was diminished but still persisted even taking account of severity, chronicity, current symptoms and treatment. In contrast Kendler and co-workers in their study of adult twins (Kendler *et al.* 2001) found no differences in reliability between men and women over a 19-month period. One explanation might be that differential recall is particularly marked in adolescence and early adult life and that subsequently there is little difference. Alternatively it might be that the time interval in Kendler's study was too short to show the effect of differential forgetting.

The comparison of risk factor analyses using longitudinal reports or recall as the outcome indicates that major effects are likely to be detected with either outcome measure. However, discrepancies observed on risk factors with small effects should make researchers cautious about such effects if based on recall data.

Strengths

The longitudinal nature of this large and comprehensive study enabled comparison of longitudinal reports of depressive symptoms from age 14 and recall of key symptoms, thus indicating accuracy, not just reliability. Furthermore, this design made possible the comparison of risk factor analyses for longitudinal reports or recall of key symptoms.

https://doi.org/10.1017/S0033291703001843 Published online by Cambridge University Press

Limitations

This study looked only at recall of key symptoms, not at attempted recall for all symptoms of depression, so no diagnosis of depression could be made from recall. However, since the key symptoms of depressed mood or loss of interest are required for diagnosis, the percentage who could have received a lifetime diagnosis of depression from recall must have been lower than the percentage who recalled at least one key symptom.

Both the age of this cohort (25 years at last interview) and the experience of being in a cohort may have influenced the results. Genetic studies and community surveys typically interview adults aged 18 and over, often with no upper age limit. Furthermore, most participants in these studies are interviewed about their psychiatric history for the first time, whereas in this cohort study the 25-year interview was the fifth time the cohort members had been interviewed about psychiatric symptoms. Robins (1985) and Aneshensel *et al.* (1987) have discussed reasons why lifetime recall might be lower at a subsequent interview. Some apparent failure to recall may be a misunderstanding of the task or of the questions, a failure to make the effort to search memory, or a decision not to report symptoms this time.

Implications

For genetic studies that use recall to ascertain lifetime prevalence, the primary consequence of recall failure will be that the phenotype as measured will not be quite the phenotype as defined. Those with more severe and more chronic depression at any time prior to interview will be those who are more likely to meet criteria based on the interview data. People who have experienced less severe, transient episodes of depression will tend to be categorized as never having been depressed. The consequences may be advantageous or disadvantageous for genetic studies, depending on how appropriate the phenotype is for studies of heritability or the relationship between genotype and phenotype. For family studies the lower estimates of lifetime prevalence in relatives may provide encouraging but misleading information for clinicians and patients. However, recall estimates may be appropriate if what is required is an estimate of the percentage likely to experience severe or chronic depression.

For community studies of lifetime prevalence, recall failure clearly leads to underestimation of lifetime prevalence. In spite of this the parallel analyses of risk factors using longitudinal reports or recall indicate that major effects will emerge in both. Nonetheless, small effects should be regarded with caution if based on recall.

Recall is a general problem in epidemiology, not just in psychiatric epidemiology (Friedenreich, 1994). Without it, little can be done; using recall, biases abound. One way forward is to attempt to improve recall through interview modifications with the use of time lines (Lyketsos *et al.* 1994), or probes and a requirement that the respondent commit to a serious memory search (as in the current version of the CIDI) (Kessler *et al.* 1998). Analyses and interpretations of results should take account of the problems of recall failure and unreliability. Additional reliability studies are needed only to test a new instrument or if reliability estimates from a sample are to be used to correct for measurement error (Foley *et al.* 1998; Kuha *et al.* 1998) in other analyses of that sample.

ACKNOWLEDGEMENTS

This research was funded by grants from the Health Research Council of New Zealand, and the National Child Health Research Foundation, Auckland; the Canterbury Medical Research Foundation, Christchurch; and the New Zealand Lottery Grants Board, Wellington, New Zealand.

REFERENCES

Andrews, G., Anstey, K., Brodaty, H., Issakidis, C. & Luscombe, G. (1999). Recall of depressive episode 25 years previously. *Psychological Medicine* **29**, 787–791.

Aneshensel, C. S., Estrada, A. L., Hansell, M. J. & Clark, V. A. (1987). Social psychological aspects of reporting behaviour: lifetime depressive episode reports. *Journal of Health and Social Behaviour* **28**, 232–246.

Armsden, G. C. & Greenberg, M. T. (1987). The inventory of parent and peer attachment: individual differences and their relationship to psychological well-being in adolescence. *Journal of Youth and Adolescence* **16**, 427–454.

Bromet, E. J., Dunn, L. O., Connell, M. M., Dew, M. A. & Schulberg, H. C. (1986). Long-term reliability of diagnosing lifetime major depression in a community sample. *Archives of General Psychiatry* **43**, 435–440.

https://doi.org/10.1017/S0033291703001843 Published online by Cambridge University Press

Costello, A., Edelbrock, C., Kalas, R., Kessler, M. & Klaric, S. A. (1982). *Diagnostic Interview Schedule for Children*. National Institute of Mental Health: Bethesda, MD.

Dohrenwend, B. P. (1990). 'The problem of validity in field studies of psychological disorders' revisited. *Psychological Medicine* **20**, 195–208.

Elley, W. B. & Irving, J. C. (1976). Revised socio-economic index for New Zealand. *New Zealand Journal of Educational Studies* **11**, 25–36.

Eysenck, H. M. & Eysenck, S. B. G. (1964). *Manual of the Eysenck Personality Inventory*. London University Press: London.

Fendrich, M., Weissman, M. M., Warner, V. & Mufson, L. (1990). Two-year recall of lifetime diagnoses in offspring at high and low risk for major depression. *Archives of General Psychiatry* **47**, 1121–1127.

Fergusson, D. M. & Lynskey, M. T. (1995). Suicide attempts and suicidal ideation in a birth cohort of 16-year-old New Zealanders. *Journal of the American Academy of Child & Adolescent Psychiatry* **34**, 1308–1317.

Fergusson, D. M. & Lynskey, M. T. (1997). Physical punishment/maltreatment during childhood and adjustment in young adulthood. *Child Abuse & Neglect* **21**, 617–630.

Fergusson, D. M., Horwood, L. J. & Beautrais, A. L. (1999). Is sexual orientation related to mental health problems and suicidality in young people? *Archives of General Psychiatry* **56**, 876–880.

Fergusson, D. M., Horwood, L. J., Shannon, F. T. & Lawton, J. M. (1989). The Christchurch Child Development Study: a review of epidemiological findings. *Paediatric and Perinatal Epidemiology* **3**, 302–325.

Fergusson, D. M., Lynskey, M. T. & Horwood, L. J. (1996). Childhood sexual abuse and psychiatric disorder in young adulthood: I. Prevalence of sexual abuse and factors associated with sexual abuse. *Journal of the American Academy of Child & Adolescent Psychiatry* **35**, 1355–1364.

Foley, D. L., Neale, M. C. & Kendler, K. S. (1998). Reliability of a lifetime history of major depression: implications for heritability and co-morbidity. *Psychological Medicine* **28**, 857–870.

Friedenreich, C. M. (1994). Improving long-term recall in epidemiologic studies. *Epidemiology* **5**, 1–4.

Goldberg, J. D. (1975). The effects of misclassification on the bias in the difference between two proportions and the relative odds in the fourfold table. *Journal of the American Statistical Association* **70**, 561–567.

Hankin, B. L., Abramson, L. Y., Moffitt, T. E., Silva, P. A., McGee, R. & Angell, K. E. (1998). Development of depression from preadolescence to young adulthood: emerging gender differences in a 10-year longitudinal study. *Journal of Abnormal Psychology* **107**, 128–140.

Keller, M. B., Lavori, P. W., McDonald-Scott, P., Scheftner, W. A., Andreasen, N. C., Shapiro, R. W. & Croughan, J. (1981). Reliability of lifetime diagnoses and symptoms in patients with a current psychiatric disorder. *Journal of Psychiatric Research* **16**, 229–240.

Kendler, K. S., Gardner, C. O. & Prescott, C. A. (2001). Are there sex differences in the reliability of a lifetime history of major depression and its predictors? *Psychological Medicine* **31**, 617–625.

Kessler, R. C., Wittchen, H.-U., Abelson, J. M., McGonagle, K., Schwarz, N., Kendler, K. S., Knauper, B. & Zhao, S. (1998). Methodological studies of the Composite International Diagnostic Interview (CIDI) in the US National Comorbidity Survey (NCS). *International Journal of Methods in Psychiatric Research* **7**, 33–55.

Kohn, R., Dohrenwend, B. P. & Mirotznik, J. (1998). Epidemiological findings on selected psychiatric disorders in the general population. In *Adversity, Stress and Psychopathology* (ed. B. P. Dohrenwend), pp. 235–284. Oxford University Press: New York.

Kuha, J., Skinner, C. & Palmgren, J. (1998). Missclassification error. In *Encyclopedia of Biostatistics*, vol. 4 (ed. P. Armitage and T. Colton), pp. 2615–2621. Wiley: Chichester.

Lyketsos, C. G., Nestadt, G., Cwi, J., Heitoff, K. & Eaton, W. W. (1994). The Life Chart Interview: a standardized method to describe the course of psychopathology. *International Journal of Methods in Psychiatric Research* **4**, 143–155.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika* **78**, 691–692.

Pilowsky, I. & Boulton, D. (1970). Development of a questionnaire-based decision rule for classifying depressed patients. *British Journal of Psychiatry* **116**, 647–650.

Pilowsky, I., Levine, S. & Boulton, D. (1969). The classification of depression by numerical taxonomy. *British Journal of Psychiatry* **115**, 937–945.

Prusoff, B. A., Merikangas, K. R. & Weissman, M. M. (1988). Lifetime prevalence and age of onset of psychiatric disorders: recall 4 years later. *Journal of Psychiatric Research* **22**, 107–117.

Rice, J. P., Rochberg, N., Endicott, J., Lavori, P. W. & Miller, C. (1992). Stability of psychiatric diagnoses. An application to the affective disorders. *Archives of General Psychiatry* **49**, 824–830.

Robins, L. N. (1985). Epidemiology: reflections on testing the validity of psychiatric interviews. *Archives of General Psychiatry* **42**, 918–924.

Robins, L. N. (1990). Psychiatric epidemiology – a historic review. *Social Psychiatry & Psychiatric Epidemiology* **25**, 16–26.

Robins, L., Helzer, J. E., Croughan, J. & Ratcliff, K. S. (1981). National Institute of Mental Health Diagnostic Interview Schedule: its history, characteristics, and validity. *Archives of General Psychiatry* **38**, 381–389.

Robins, L. N., Helzer, J. E., Weissman, M. M., Orvaschel, H., Gruenberg, E., Burke Jr, J. D. & Regier, D. A. (1984). Lifetime prevalence of specific psychiatric disorders in three sites. *Archives of General Psychiatry* **41**, 949–958.

Robins, L. N., Wing, J., Wittchen, H. U., Helzer, J. E., Babor, T. F., Burke, J., Farmer, A., Jablenski, A., Pickens, R. & Regier, D. A. (1988). The Composite International Diagnostic Interview. An epidemiologic instrument suitable for use in conjunction with different diagnostic systems and in different cultures. *Archives of General Psychiatry* **45**, 1069–1077.

SAS Institute Inc. (1999). *SAS/STAT User's Guide, Version 8*. SAS Institute Inc.: Cary, NC.

Stone, M. (1998). Akaike's criteria. In *Encylopedia of Biostatistics*, vol. 1 (ed. P. Armitage and T. Colton), pp. 123–124. Wiley: Chichester.

Wechsler, D. (1974). *Manual for the Wechsler Intelligence Scale for Children – Revised*. Psychological Corporation: New York.

Wilhelm, K. & Parker, G. (1994). Sex differences in lifetime depression rates – fact or artefact. *Psychological Medicine* **24**, 97–111.

Williams, J. B. W., Gibbon, M., First, M. B., Spitzer, R. L., Davies, M., Borus, J., Howes, M. J., Kane, J., Pope Jr, H. G., Rounsaville, B. & Wittchen, H. U. (1992). The Structured Clinical Interview for DSM-III-R (SCID). II. Multisite test-retest reliability. *Archives of General Psychiatry* **49**, 630–636.

WHO (1989). *Schedules for Clinical Assessment in Neuropsychiatry (SCAN)*. World Health Organization: Geneva.

WHO Division of Mental Health (1993). *Composite International Diagnostic Interview*. World Health Organization: Geneva.

https://doi.org/10.1017/S0033291703001843 Published online by Cambridge University Press