Using life history calendars to improve measurement of lifetime experience with mental disorders

William G. Axinn[1], Stephanie Chardoul[1], Heather Gatny[1], Dirgha J. Ghimire[1], Jordan W. Smoller[2,3], Yang Zhang[1] and Kate M. Scott[4]

[1]Institute for Social Research, University of Michigan, 426 Thompson St, Ann Arbor, MI 48104-2321, USA; [2]Department of Psychiatry and Psychiatric and Neurodevelopmental Genetics Unit, Harvard Medical School; Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA; [3]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA and [4]Department of Psychological Medicine, Dunedin School of Medicine, University of Otago, Dunedin, New Zealand

Abstract

Background. Retrospective reports of lifetime experience with mental disorders greatly underestimate the actual experiences of disorder because recall error biases reporting of earlier life symptoms downward. This fundamental obstacle to accurate reporting has many adverse consequences for the study and treatment of mental disorders. Better tools for accurate retrospective reporting of mental disorder symptoms have the potential for broad scientific benefits.
Methods. We designed a life history calendar (LHC) to support this task, and randomized more than 1000 individuals to each arm of a retrospective diagnostic interview with and without the LHC. We also conducted a careful validation with the Structured Clinical Interview for the Diagnostic and Statistical Manual of Mental Disorders-Fourth Edition.
Results. Results demonstrate that—just as with frequent measurement longitudinal studies— use of an LHC in retrospective measurement can more than double reports of lifetime experience of some mental disorders.
Conclusions. The LHC significantly improves retrospective reporting of mental disorders. This tool is practical for application in both large cross-sectional surveys of the general population and clinical intake of new patients.

Introduction

Retrospective reports of lifetime experience with mental disorders greatly underestimate actual experiences of disorder because recall error biases reporting of earlier life symptoms downward. Only the most severe or enduring disorders are accurately reported many years later (Wells and Horwood, 2004). A careful comparison of prospective longitudinal measurement to retrospective measurement indicates this underreporting is so severe only half of disorders are reported retrospectively—and only through age 32 (Moffitt *et al.*, 2010). At older ages, retrospective recall bias is likely to be much worse. This fundamental obstacle to accurate reporting has adverse consequences for the study and treatment of mental disorders.

This study applied the survey methodology of improving retrospective recall in a population-scale experiment to test the potential for reducing bias in recall of lifetime mental disorders. We designed a life history calendar (LHC) for this task. LHC methods are designed to make the cognitively challenging task of recall of the timing of life events easier by providing a matrix of visual cues respondents can use to help them recall the timing of those life events (Freedman *et al.*, 1988; Petersen and Kerwin, 1992). These timing cues are both standardized, (e.g. as column headings marked with years and ages), and flexible, usually composed of respondents' reports of other personal life events, which increase the power of respondents' autobiographical memory (Belli, 1998). Specifically around traumatic events, a structured experiment demonstrated the LHC method elicited more reports of lifetime intimate partner violence – even more strongly for abuse that happened early in life – than a standard interview survey (Yoshihama *et al.*, 2005). These examples demonstrate the broad social science relevance and application of the LHC.

Two decades ago careful adaptation of the methods produced highly successful LHC tools for use across multiple ethnic groups in rural Nepal (Axinn *et al.*, 1997, 1999), the setting of the present study. This research was particularly important for multicultural use of the LHC method because the multiple ethnic groups in rural Nepal not only speak mutually unintelligible languages, they also represent a wide range of conceptualizations of time and use of calendars in daily life (Axinn *et al.*, 1999; Axinn and Pearce, 2006). These successes give

the LHC tool a favorable profile for pairing with multicultural psychiatric diagnostic interviews to improve recall of episodes of psychiatric stress or mental disorder.

In a subsample of a long-running panel in Nepal we randomized more than 1000 individuals to each arm of a retrospective diagnostic interview (the World Health Organization's (WHO) Composite International Diagnostic Interview (CIDI)) with and without the LHC. Our aims were to compare screening and diagnosis rates of lifetime experience of selected mental disorders among those randomized to each arm of the study and to examine whether there were sex, age, or education differences in the effect of LHC administration. Additionally, we conducted a clinical validation of the the Diagnostic and Statistical Manual of Mental Disorders-Fourth Edition (DSM-IV) diagnoses produced by the LHC-CIDI instrument. Prior validation of the Nepal-CIDI without the LHC calendar demonstrates that instrument is conservative (Ghimire *et al*., 2013).

Methods

Measures

The LHC

LHCs are typically designed in the form of a matrix where the columns represent time units (weeks, months, years) and the rows contain the domains of life being studied (Axinn *et al*., 1999).[†1] The paper calendar used in this investigation (where the interviewer and respondent sit together and use the visual cues and matrix format of the calendar) is described in online Supplementary Appendix A. The top lines begin with Nepal-specific visual cues used in this LHC. This includes Nepalese calendar years, the timing of important national events that occurred at the same time for all of those in the study population (Axinn *et al*., 1999), and the timing of highly memorable local neighborhood events that are pre-edited for each specific neighborhood before the calendar is administered (Axinn *et al*., 1997). Visual aspects of the LHC method help ensure that interviewers collect complete data. The interviewer's work begins with the respondent's age and important personal experiences (online Supplementary Appendix A). This section of the calendar is administered to all respondents prior to the section of the diagnostic instrument that screens for key symptoms of mental disorders (the CIDI—described below). It includes a complete residential history, marital and childbearing history, and a history of educational and job change events. Recording this information can take 15 minutes, but this information becomes a crucial person-specific set of memory cues and the process of remembering and recording all of them creates cognitive engagement in the detailed recall of prior life experiences.

All respondents then begin the screening section of the CIDI interview. Many respondents do not enter additional information on the calendar. The only respondents who return to the LHC are those who proceed to the diagnostic portion of the interview, and are unable to remember their age at an episode of mental disorder. Interviewers then help these respondents pinpoint the age at which the episode was experienced using the information previously recorded on the LHC. Age is recorded on the appropriate row of the LHC for the disorder (rows 9–20), and the

interviewer then returns to the CIDI to enter the age into the computer and proceed to the next question.

Mental disorders

DSM-IV disorders were measured using selected modules from the Nepal version of the WHO-CIDI (Wittchen, 1994; Horwath and Weissman, 2000; Kessler and Üstün, 2004; Pennell *et al*., 2008). Following standard CIDI administration practice, professional interviewers were rigorously trained in the administration of the CIDI using computer-assisted personal interviewing, and then went to respondents' homes, obtained consent and privacy, and administered the CIDI. The modules included were Depression, Mania, Panic Disorder, generalized anxiety disorder (GAD), post-traumatic stress disorder (PTSD), intermittent explosive disorder, and alcohol use disorders (AUD). The high effort to construct culturally and linguistically appropriate translations of diagnostic measures forced us to limit the focal disorders to a subset of all disorders. The selection of these specific disorders was informed by prior studies as well as pilot data indicating the disorders anticipated to be most prevalent and/or clinically significant. Careful translation and adaptation were used in a multistep process to create this Nepal CIDI (Ghimire *et al*., 2013). Clinical validation of the Nepal CIDI against the gold standard of the clinician-administered Structured Clinical Interview for DSM-IV (SCID-IV) demonstrates high concordance, comparable to validation studies of the USA and European CIDI instruments (Ghimire *et al*., 2013). Analyses of pilot data from Nepal demonstrated expected associations of these conditions and potentially traumatic experiences with Nepal CIDI measures of mental disorder (Axinn *et al*., 2013, 2015).

Socio-demographics

We include *sex*, *age*, *birth cohort* (1957–71, 1972–81, 1982–91, 1992–2001), *ethnicity*, and *education* as covariates. Ethnicity is coded into the five most common ethnic groups in this area of Nepal and one 'other' category: Brahmin/Chhetri, Hill Janajati, Newar, Terai Janajati, Dalit, or 'others' (see Axinn *et al*., 1999). Though the data include details of educational attainment and literacy, we summarize variability in education with a single dichotomous indicator of achieving a 'School Leaving Certificate' (SLC). The SLC is awarded to those scoring highly enough on a nationally standardized exam offered after the successful completion of 10th grade and variance in this attainment reflects recent changes across cohorts in access to schools.

Experimental design

To evaluate the performance of the LHC, we selected a subsample from the ongoing Chitwan Valley Family Study (CVFS) in Nepal. The CVFS launched in 1995 with a *general population sample* of 151 neighborhoods (clusters of 5–15 households) fully representative of rural Chitwan in Nepal. All individuals, including migrants from the selected neighborhoods, were then entered into a unique household registry system using state-of-the-art survey methods to achieve high rates of re-contact and re-interview through 2016 (Axinn *et al*., 2012). These methods featured frequent re-contact and mixed mode data collection to retain 95% of the original respondents. Both household and individual interviews on social, economic, and demographic topics have been repeated multiple times. These interviews generated high-quality data with >90% response rates, low refusal rates, and low item missing data (Thornton *et al*., Forthcoming).

[†]The notes appear after the main text.
[1]For examples of different designs of LHCs, please refer to Freedman *et al*. (1988), Belli (1998), and Axinn & Pearce (2006).

We chose a sub-sample of 50 of the 151 CVFS neighborhoods, and randomly selected 25 neighborhoods to receive an integrated LHC and CIDI interview and the other 25 neighborhoods to receive a CIDI interview with no LHC. This community-level randomization was used to create a geographically matched set of neighborhoods with and without LHCs in the CIDI interview. The sample yielded 1404 individuals, aged 15–59, who were interviewed with CIDI only and 1089 individuals who were interviewed with a LHC first, then the CIDI. Note that randomization took place at the community level, creating unequal sample of individuals, but that both samples attained response rates of over 94%.

Clinical validation

We followed the validation protocol published by Haro et al. (2006) and employed in other CIDI clinical validation studies. These clinical validation studies repeatedly demonstrate the CIDI's high success in matching survey measures from the general population with diagnoses by clinicians (Haro et al., 2006; Kessler and Üstün, 2008; Kessler et al., 2009, 2013). The protocol features clinical re-interview of selected CIDI respondents with the Axis I research version, non-patient edition of the SCID-IV (First et al., 2002), performed by professional mental health clinicians (MDs, nurses, or clinical social workers trained and certified on SCID administration).

To implement this protocol we selected a systematic sample of 120 respondents who completed the LHC-CIDI. Following best practices, the sample was stratified to ensure large numbers of cases who had screened positive for each disorder, with roughly equal numbers of those who subsequently diagnosed positive and negative. The clinical validation interviews were performed with the SCID by a highly trained team of one psychiatrist and one psychiatric nurse, cross-checked by a psychiatric clinical social worker with extensive experience conducting SCID validation studies across multiple CIDI projects.

Statistical analysis

This investigation focuses on the four most prevalent disorders to maximize robustness of estimates: major depressive disorder (MDD), GAD, AUD (alcohol abuse or dependence), and PTSD. The CIDI questionnaire features both screening for key symptoms of disorders and then, among those who screen positive, detailed diagnostic assessments. The diagnostic portion of these interviews asks respondents about details of their experiences including specific symptoms, frequency and lengths of time experiencing these symptoms, and level of impairment caused by the symptoms. These details are used in a complex statistical algorithm mapped to the criteria of the DSM-IV to determine whether or not each individual meets the threshold for a diagnosis.

Our analyses have several steps. In step 1, we compare proportions and use proportion z-tests to compare CIDI screening and diagnosis rates for individuals who were interviewed with the CIDI only and individuals who were interviewed with the LHC and the CIDI (online Supplementary Appendix B, Table 1). We present the percentage of individuals screening positive and the percentage of individuals with a DSM-IV diagnosis by disorder for both samples. In step 2, we estimate multivariate logistic regression to assess the interacting effects between the use of LHC and multiple factors (sex, age, education) on the likelihood of screening positive or diagnosing with a specific disorder. Step 1

Table 1. Percent meeting the criteria for mental disorder, with and without use of life history calendar (LHC)

	Without LHC (n = 1404), %	With LHC (n = 1089), %
Screening Positive		
Depression	19.59	29.38***
Generalized anxiety disorder	17.09	21.03*
Alcohol use disorder	29.56	28.65
Post-traumatic stress disorder[a]	78.21	83.84***
DSM-IV Diagnoses (Life Time)		
Depression	1.85	14.60***
Generalized anxiety disorder	1.78	7.35***
Alcohol use disorder	5.06	5.60
Post-traumatic stress disorder[a]	1.71	3.67**

[a]Respondents who experienced at least one trauma in their lifetime will be screened positive for post-traumatic stress disorder

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$ (proportion tests for significant differences between the two subsamples).

and 2 statistical analyses were performed using STATA 15. In step 3, we focus on the comparison of the SCID validation interviews to the LHC-CIDI interviews. Appropriate statistical weights were used to derive estimates for the entire group of LHC-CIDI respondents and calculate many different commonly used validation metrics. These analyses were conducted using SUDAAN.

Results

Sample characteristics

The two groups (the 'LHC group' and the 'non-LHC group') did not significantly differ on age or sex, but they did differ in ethnic group distribution (online Supplementary Appendix B, Table 1). They also differed in educational attainment, with a higher proportion of the LHC group attaining a greater degree of education (having the SLC).

Comparison of CIDI screening and diagnoses rates with v. without a LHC

Receiving the LHC prior to the CIDI interview resulted in significantly higher rates of screening positive, and meeting lifetime diagnostic criteria for MDD, GAD, and PTSD (Table 1). The difference between the two groups for lifetime MDD diagnosis was especially pronounced, with only 1.85% of the non-LHC group meeting diagnostic criteria, compared with a diagnosis rate of 14.60% among the LHC group ($p < 0.001$). The use of the LHC did not influence rates (either positive screens or diagnoses) of AUD.

Interactions of LHC use with sex, age, and education

Significant differences in the effect of the LHC on disorder prevalence by sex, by age, and by education were found (online Supplementary Appendix B, Table 4 shows interaction model results). Tables 2–3 (and online Supplementary Appendix B, Table 2) show the prevalence of screen positives and diagnoses

Table 2. Percent meeting the criteria for mental disorder, with and without use of life history calendar (LHC), by sex

	Male (n = 1141)		Female (n = 1352)		
	Without LHC (n = 656) %	With LHC (n = 485) %	Without LHC (n = 748) %	With LHC (n = 604) %	Sex × LHC[a]
Screening Positive					
Depression	16.46	17.53	22.33	38.91***	$p < 0.001$***
Generalized anxiety disorder	15.09	9.48**	18.85	30.30***	$p < 0.001$***
Alcohol use disorder	58.99	60.62	3.74	2.98	$p = 0.359$
Post-traumatic stress disorder[b]	71.49	87.63***	84.09	80.79	$p < 0.001$***
DSM-IV Diagnoses (Life Time)					
Depression	1.22	4.54***	2.41	22.68***	$p = 0.021$*
Generalized anxiety disorder	1.22	1.24	2.27	12.25***	$p = 0.004$**
Alcohol use disorder	10.67	12.16	0.13	0.33	$p = 0.539$
Post-traumatic stress disorder[b]	0.46	1.65*	2.81	5.30*	$p = 0.390$

[a]Logistic results of the interaction between sex and the use of LHC on meeting criteria for mental disorder.
[b]Respondents who experienced at least one trauma in their lifetime will be screened positive for post-traumatic stress disorder.
* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$ (proportion tests for significant differences between the two subsamples).

Table 3. Percent meeting the criteria for mental disorder, with and without use of life history calendar (LHC), by age (age 30 is approximate median age of the sample)

	Age<30 (n = 1230)		Age>= 30 (n = 1263)		
	Without LHC (n = 678) %	With LHC (n = 552) %	Without LHC (n = 726) %	With LHC (n = 537) %	Age × LHC[a]
Screening Positive					
Depression	18.14	23.55*	20.94	35.38***	$p = 0.038$*
Generalized anxiety disorder	14.90	13.04	19.15	29.24***	$p = 0.001$***
Alcohol use disorder	21.68	21.38	36.91	36.13	$p = 0.930$
Post-traumatic stress disorder[b]	77.58	82.43*	78.79	85.29**	$p = 0.502$
DSM-IV Diagnoses (Life Time)					
Depression	2.06	8.15***	1.65	21.23***	$p = 0.002$**
Generalized anxiety disorder	0.74	2.72**	2.75	12.10***	$p = 0.659$
Alcohol use disorder	4.57	3.26	5.51	8.01	$p = 0.048$*
Post-traumatic stress disorder[b]	1.03	1.99	2.34	5.40**	$p = 0.729$

[a]Logistic results of the interaction between age (age group is stratified by the median of age) and the use of LHC on meeting criteria for mental disorder.
[b]Respondents who experienced at least one trauma in their lifetime will be screened positive for post-traumatic stress disorder.
* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$ (proportion tests for significant differences between the two subsamples)

among the two LHC groups stratified by each of these demographic variables in turn.

For sex, Table 2 illustrates that the effect of the LHC in increasing rates of positive screens and lifetime diagnoses was greater among women relative to men, at least for some disorders. The interaction of sex and LHC was significant for the screen positives for MDD, GAD, and PTSD (shown by the *p*-values in the final column; interaction odds ratios shown in online Supplementary Appendix 1), and it was also significant for lifetime MDD and GAD diagnoses. For lifetime PTSD the interaction with sex was not significant ($p = 0.390$); the use of the LHC increased diagnosis of PTSD among men and women equally. There were no sex differences in the effect of the LHC for AUD.

For age, Table 3 shows that the effect of the LHC tended to be greater among older participants (>=30) relative to those under the age of 30. This interaction with age was significant for the screen positives for MDD and GAD, for lifetime diagnoses of MDD and AUD (with a similar, albeit non-significant, pattern evident for lifetime GAD). This stronger effect of the LHC method among older participants is further illustrated in Fig. 1, which shows the prevalence of MDD (screen positives and lifetime diagnosis) in the two groups, stratified by birth cohort. Particularly for MDD diagnosis, use of the LHC dramatically enhanced recall among older cohorts compared to younger cohorts. The prevalence of the other three disorders, by birth cohort, is shown in online Supplementary Appendix B, Table 3;
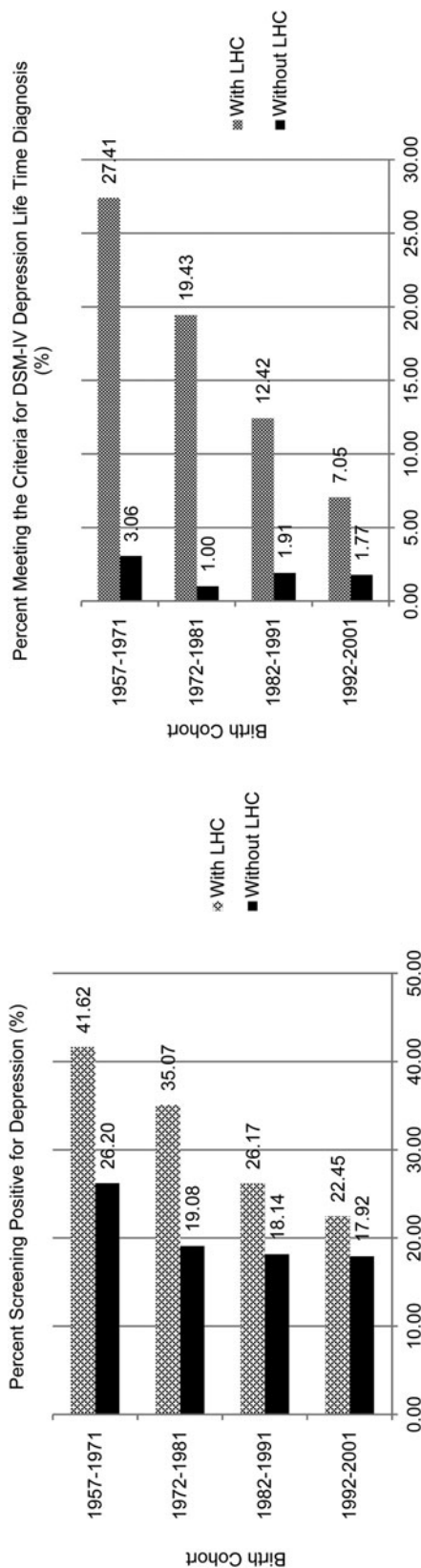
the same birth cohort pattern shown for MDD is apparent for GAD and PTSD, although it is less pronounced.

The LHC method appears to have a greater impact among those with less education (online Supplementary Appendix B, Table 2). Although the interaction with education did not reach significance for any of the screen positives, it was close ($p = 0.061$) for MDD screen positives, and the same pattern was evident for GAD and PTSD screen positives. The interaction with education was significant for MDD diagnoses and close to significance for GAD, such that there was a much higher prevalence of these disorders among the LHC group compared to the non-LHC group among those with less education, but a more muted effect of the LHC on diagnosis rates among those with more education. Similarly, it was only among those with less education that the LHC group had significantly higher rates of lifetime PTSD compared to the non-LHC group (however, in the absence of a significant interaction this finding must be interpreted with caution).

Clinical validation

Table 4 displays several validation statistics for comparisons between survey measures of mental disorders and clinical SCID interviews: McNemar chi-square (χ^2), area under the receiver operating characteristics curve (AUC), Cohen kappa (kappa), sensitivity (SN), specificity (SP), positive predictive value (PPV), and negative predictive value (NPV). To aid in interpretation of these validation statistics we compare them to the US CIDI validation statistics published by Haro et al. (2006). Focusing on the AUC summary measure, for example, the AUC for MDD validation in the Nepal LHC-CIDI is good and matches the AUC for the USA CIDI. The Nepal LHC-CIDI AUC for GAD is comparable to the AUC statistics for specific forms of anxiety published by Haro et al. (2006) although they do not provide AUC for GAD. The Nepal LHC-CIDI AUC for AUD is also strong and similar to the AUC for AUD in the US CIDI. The Nepal LHC-CIDI AUC for PTSD is higher than the AUC for PTSD in the US CIDI; this finding is likely because levels of traumatic experience are so high in rural Nepal. The high incidence of traumatic experience not only means a high proportion of the population has been exposed to trauma, but that, on average, those exposed to trauma experienced that trauma much more recently than individuals in populations with a low incidence of traumatic experience. This recency, on average, means respondents are likely more successful recalling details of their reactions to a specific traumatic experience. This may contribute to the high validity of CIDI measures of PTSD in Nepal.

Note that the increased rates with the LHC represent improved SN and positive predicted values as opposed to reduced SP or increased false positives. Comparison with gold standard SCID interviews is consistent with this conclusion, showing that even with the LHC, rates are still conservative compared to the SCID. CIDI diagnostic measures are generally more conservative than SCID diagnostic measures, though CIDI screening measures provide greater inclusion than other screeners, such as the PHQ-9 (Wittchen, 1994; Haro et al., 2006; Kessler and Üstün, 2008; Kessler et al., 2013). Overall, it is highly unlikely engagement in the LHC would cause people to falsely recall psychiatric disorder episodes.

Discussion

Accurate lifetime retrospective reports of mental disorders are an important scientific priority. Though longitudinal studies with

Fig. 1. Percent screening positive and meeting lifetime criteria for depression, with and without use of the life history calendar, by birth cohort.

Table 4. Measures of the consistency of lifetime DSM-IV diagnoses obtained with the LHC-CIDI and the SCID, in comparison to similar validation in the USA.[a]

		χ²	AUC	Kappa	(se)	SN	(se)	SP	(se)	PPV	(se)	NPV	(se)
Major depressive disorder (MDD)	Nepal	4.4*	0.75	0.56	(0.13)	52.80	(14.20)	96.60	(1.80)	84.00	(4.00)	85.90	(6.70)
	USA[b]	7.2*	0.75	0.54	(0.06)	55.30	(6.80)	93.70	(1.90)	73.70	(7.00)	86.80	(2.70)
Generalized anxiety disorder (GAD)	Nepal	3.7*	0.67	0.45	(0.12)	36.20	(13.50)	98.40	(1.10)	77.60	(12.40)	91.10	(5.50)
	USA[c]												
Alcohol use disorder (AUD)	Nepal	20.7*	0.95	0.79	(0.11)	100.00	(0.00)	90.80	(4.60)	71.40	(13.00)	100.00	(0.00)
	USA[b]	7.3*	0.81	0.70	(0.06)	64.10	(7.40)	98.10	(1.00)	88.10	(5.60)	92.70	(2.00)
Post-traumatic stress disorder (PTSD)	Nepal	1.4	0.83	0.64	(0.11)	76.50	(11.80)	90.50	(4.30)	68.40	(11.70)	93.40	(4.70)
	USA[b]	11.4*	0.69	0.49	(0.10)	38.30	(11.80)	99.10	(0.50)	86.10	(7.70)	91.30	(3.00)

[a]A statistically significant χ² indicates a significant difference between the survey and SCID measures and larger the χ² the larger the difference. An AUC of 1.0 indicates a perfect match; and AUC of 0 indicates they are perfectly unmatched. A Cohen's kappa measure of concordance ranges from 0 to 1, with 0 meaning matching only by chance and 1 meaning a perfect match. The SN is percentage of SCID diagnosed cases correctly classified by CIDI and SP is the percentage of SCID non-diagnosed cases correctly classified by CIDI. If SN is low and SP is high, the CIDI is conservative in diagnosing relative to SCID, but if SN is high and SP is low the CIDI then the CIDI diagnoses more aggressively than SCID. PPV is proportion of CIDI cases that are confirmed by SCID, and NPV is the proportion of CIDI noncases that are confirmed by the SCID. Please note the high standard errors on some measures, such as SN and PPV, because of the relatively small numbers of diagnoses.
[b]Results from 'Concordance of the Composite International Diagnostic Interview Version 3.0 (CIDI 3.0) with standardized clinical assessments in the WHO World Mental Health surveys,' by J.M. Haro *et al.,* 2006, *International Journal of Methods in Psychiatric Research,* 15,4, p. 173–174, Table 1.
[c]There is no WMH-CIDI comparative result from the USA for generalized anxiety disorder.
*McNemar Chi Square Test significant at the 0.05 level, two-sided test.
χ², McNemar Chi Square; kappa, Cohen kappa; AUC, area under the receiver operating characteristics curve; SN, sensitivity; SP, specificity; PPV, positive predictive value; NPV, negative predictive value.

frequent measurement are desirable, they are rare because of the tremendous effort required. Limiting epidemiological research to longitudinal studies would largely eliminate population monitoring across time, international comparative studies, and population-scale investigation of high-priority sub-populations. Moreover, to launch a new longitudinal study, accurate baseline assessment of prior lifetime experiences is needed to understand current conditions. This issue matches a clinical perspective on recall bias. If patients can more accurately report the life history of experience with disorders, this may optimize treatment planning. Better tools for accurate retrospective reporting of mental disorder symptoms have the potential for broad scientific and clinical benefits.

This randomized experiment with an LHC produced significantly higher reports of lifetime experience of mental disorders when used in conjunction with the CIDI. The higher lifetime prevalence of disorders in the LHC group was evident for MDD, GAD, and PTSD, but not AUD. The LHC-CIDI proved to have good clinical validity, improving detection of disorders relative to the conventional (non-LHC) CIDI, but it remained conservative relative to the SCID, suggesting that the higher rates primarily reflected improved SN rather than false positives. The LHC increased measurement of disorders in men and women, but for MDD and GAD the LHC effect was significantly stronger among women. The LHC increased reports among younger and older respondents, but there was a significantly stronger effect among older respondents, varying in magnitude across disorders. Finally, administration of the LHC increased disorder detection markedly among those with less education, with less of an effect among those with more education (but an effect nonetheless, at least for MDD).

Completing the LHC, including cognitive engagement in detailed recall of life events, results in enhanced recall of mental disorder symptoms through their temporal association with key life events. It is not just the occurrence of symptoms that must be recalled, but also their duration. For MDD, participants must recall whether '*symptoms lasted most of the day, nearly every day*' for the threshold of '*two weeks or longer*'. This demands recalling the beginning and ending of symptoms accurately enough to know the spell covered 2 weeks or longer. We examined whether the higher rates of MDD in the LHC group were partly because of improved recall of duration of symptoms, and found that among those who received the LHC-CIDI interview and screened positive for MDD, there was a significantly higher rate (71.1%) of reporting symptoms that met the duration threshold compared to those without the LHC (28.6%). By contrast, diagnosis of AUD is partially dependent on count measures ('*Did you ever have three or more of these problems in the same 12 month period?*'). It is possible that recall of duration of episodes benefits more from the LHC than recall of counts of experiences: this may explain why we found the negligible effect of the LHC on rates of AUD.

The enhancing effect of the LHC method on disorder detection was, at least for MDD and GAD, stronger for women. The reason for this is unclear. Women tend to have better recall of disorders but it is unclear whether this is a function of better memory *per se*, or greater willingness to engage in the memory search task. Although the administration of the LHC only adds 15 min to an interview, this connection between interviewer and participant around lifetime events could influence motivation to engage in the demanding recall tasks, and there may be sex differences in these motivational effects. But, it is also noteworthy that in this

Nepalese sample there was a significant overlap in being female and being less educated (particularly among older participants).

The significantly stronger effect of the LHC in older ages is not surprising: recall is especially challenging as individuals age, and therefore we expected older respondents would experience a larger difference in reporting lifetime experience of disorders when the LHC was used. We also found benefits to those under age 30. The enhanced recall among older participants is important because it helps to reduce the age-related recall bias that is a pronounced feature of retrospective studies. The dramatically higher recall observed among the less educated is also important because it suggests that the inverse education gradient of disorder prevalence observed in prior studies based on lifetime recall probably understates the true degree of education-related social inequality in mental disorders.

A key limitation of this study is that the study population is in Nepal. Prior LHC innovations initially designed in Nepal have proved successful elsewhere (Axinn and Pearce, 2006), and the population in Nepal is similar to India and China, as well as many other Asian settings, representing much of the world population. Nevertheless, additional tests of the tool in settings quite different are a high priority. The limitation to only four disorders is also important. Though MDD, GAD, AUD, and PTSD are prevalent outside Nepal as well, application of the LHC approach to the full range of mental disorders is another priority. For example, because measures of suicidal thoughts and behaviors have similarities to depression, the LHC holds high promise for improved lifetime reporting. Likewise, though some prior research indicates clinical applications hold high promise (Caspi et al., 1996), formal testing of the LHC results against existing records of prior disorders would strengthen the empirical basis for clinical applications.

Fully removing the problems of retrospective reporting of mental disorder symptoms is unlikely to ever be achieved. However, the application of the strongest scientific tools for assisting people in their recall efforts has strong merit for both general population research and clinical practice. The LHC tools have proved successful in other areas of science. Through a large-scale randomized trial, we demonstrate the benefits of applying the LHC tool to retrospective reporting of mental disorders in a diagnostic interview. Results demonstrate that use of an LHC in retrospective measurement can more than double reports of lifetime experience of some mental disorders – a result comparable to the improvements achieved through intensive panel studies (Moffitt et al., 2010). Careful validation demonstrates this increased reporting is still conservative. Moreover, the LHC tool is practical for application in both large surveys of the general population and clinical intake of new patients.

References

Axinn WG and Pearce LD (2006) *Mixed Method Data Collection Strategies.* New York, NY: Cambridge University Press.

Axinn WG, Barber JS and Ghimire DJ (1997) The neighborhood history calendar: a data collection method designed for dynamic multilevel modeling. *Sociological Methodology* **27**, 355–392.

Axinn WG, Pearce LD and Ghimire DJ (1999) Innovations in life history calendar applications. *Social Science Research* **28**, 243–264.

Axinn WG, Ghimire D and Williams NE (2012) Collecting survey data during armed conflict. *Journal of Official Statistics* **28**, 153–171.

Axinn WG, Ghimire DJ, Williams NE and Scott KM (2013) Gender, traumatic events and mental health disorders in a Rural Asian setting. *Journal of Health and Social Behavior* **54**, 444–461.

Axinn WG, Ghimire DJ, Williams NE and Scott KM (2015) Associations between the social organization of communities and psychiatric disorders in rural Asia. *Social Psychiatry and Psychiatric Epidemiology* **50**, 1537–1545.

Belli RF (1998) The structure of autobiographical memory and the event history calendar: potential improvements in the quality of retrospective reports in surveys. *Memory (Hove, England)* **6**, 383–406.

Caspi A, Moffitt TE, Thornton A, Freedman D, Amell JW, Harrington H, Smeijers J and Silva PA (1996) The life history calendar: a research and clinical assessment method for collecting retrospective event-history data. *International Journal of Methods in Psychiatric Research* **6**, 101–114.

First MB, Spitzer RL, Gibbon M and Williams JBW (2002) *Structured Clinical Reinterview for DSM-IV Axis I Disorders, Research Version, Non-Patient Edition (SCID-I/NP).* New York: Biometrics Research, New York State Psychiatric Institute.

Freedman D, Thornton A, Camburn D, Alwin D and Young-DeMarco L (1988) The life history calendar: a technique for collecting retrospective data. *Sociological Methodology* **18**, 37–68.

Ghimire DJ, Chardoul S, Kessler RC, Axinn WG and Adhikari BP (2013) Modifying and validating the composite international diagnostic interview (CIDI) for use in Nepal. *International Journal of Methods in Psychiatric Research* **22**, 71–81.

Haro JM, Arbabzadeh-Bouchez S, Brugha TS, De Girolamo G, Guyer ME, Jin R, Lepine JP, Mazzi F, Reneses B, Vilagut G, Sampson NA and Kessler RC (2006) Concordance of the Composite International Diagnostic Interview Version 3.0 (CIDI 3.0) with standardized clinical assessments in the WHO World Mental Health Surveys. *International Journal of Methods in Psychiatric Research* **15**, 167–180.

Horwath E and Weissman MM (2000) The epidemiology and cross-national presentation of obsessive-compulsive disorder. *Psychiatric Clinics of North America* **23**, 493–507.

Kessler RC and Üstün TB (2004) The World Mental Health (WMH) Survey Initiative version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). *International Journal of Methods in Psychiatric Research* **13**, 93–121.

Kessler RC and Üstün TB (2008) *The WHO World Mental Health Surveys: Global Perspectives on the Epidemiology of Mental Disorders.* New York: Cambridge University Press.

Kessler RC, Avenevoli S, Green J, Gruber MJ, Guyer M, He Y, Jin R, Kaufman J, Sampson NA, Zaslavsky AM and Merikangas KR (2009) National comorbidity survey replication adolescent supplement (NCS-A): III. Concordance of DSM-IV/CIDI diagnoses with clinical reassessments. *Journal of the American Academy of Child & Adolescent Psychiatry* **48**, 386–399.

Kessler RC, Santiago PN, Colpe LJ, Dempsey CL, First MB, Heeringa SG, Stein MB, Fullerton CS, Gruber MJ, Naifeh JA, Nock MK, Sampson NA, Schoenbaum M, Zaslavsky AM and Ursano RJ, the Army STARRS Collaborators (2013) Clinical reappraisal of the composite international diagnostic interview screening scales (CIDI-SC) in the army study to assess risk and resilience in servicemembers (army STARRS). *International Journal of Methods in Psychiatric Research* **22**, 303–321.

Moffitt TE, Caspi A, Taylor A, Kokaua J, Milne BJ, Polanczyk G and Poulton R (2010) How common are common mental disorders?

Evidence that lifetime prevalence rates are doubled by prospective versus retrospective ascertainment. *Psychological Medicine* **40**, 899–909.

Pennell B-E, Mneimneh ZN, Bowers A, Chardoul S, Wells JE, Viana MC, Dinkelmann K, Gebler N, Florescu S, He Y, Huang Y, Tomov T and Saiz GV (2008) Implementation of the world mental health surveys. Part I. Methods. In Kessler RC and Üstün TB (eds), *The WHO World Mental Health Surveys: Global Perspectives on the Epidemiology of Mental Disorders*. New York: Cambridge University Press, pp. 33–57.

Petersen LS and Kerwin J (1992) *Testing a Life History Calendar as a Visual Aid in the National Survey of Family Growth*. no. Paper presented at the Annual Meeting of the Population Association of America, April 30-May 2, Denver.

Wells JE and Horwood LJ (2004) How accurate is recall of key symptoms of depression? A comparison of recall and longitudinal reports. *Psychological Medicine* **34**, 1001–1011.

Wittchen H-U (1994) Reliability and validity studies of the WHO-Composite International Diagnostic Interview (CIDI): a critical review. *Journal of Psychiatric Research* **28**, 57–84.

Yoshihama M, Gillespie B, Hammock AC, Belli RF and Tolman RM (2005) Does the life history calendar method facilitate the recall of intimate partner violence? Comparison of two methods of data collection. *Social Work Research* **29**, 151–163.