e-Data analysis

The supplemental file for the study entitled: *"Incidence and nature of cognitive decline over one year among HIV-infected former plasma donors in China",* includes a detailed description of the neuropsychological (NP) methodological approach to define cognitive change in individual cases and results supporting the validity of the approach. It also includes two Tables and two figures to illustrate the methodology used.

Reference sample to derive norms for change

We randomly selected 101 HIV-negative (HIV−) participants to complete a follow-up assessment over a one year interval. This sample was representative of the total baseline group with respect to demographic and overall NP status. These HIV− participants served as a reference sample to develop normative standards for NP change.

Data inspection and transformation

Among the HIV− and HIV+ individuals who participated in the one year followup assessment, there were no missing data on the 17 individual NP measures. Moreover, test-retest interval was slightly shorter for the HIV− group than the HIV+ group (323 ± 52 vs. 336 ± 37; $p < .03$). Nonetheless, the test-retest interval was not associated with practice effect on the mean scaled scores in the HIV− group ($r = -.003; p = .98$).

To define NP decline in the HIV+ sample, the 101 HIV− individuals were used as a reference sample to derive normative regression change scores and cutoffs that were then applied to the HIV+ group. Because outlier values can have a disproportionate effect on the regression [1], and to normalize data distributions, we transformed raw scores into normally distributed scaled scores (with a mean of 10 and a standard deviation of 3; see [2] for details). To test whether there was any substantial loss of predictive power when using scaled scores as compared to raw scores, we computed final regression models (see below) based on both scaled scores and raw scores and found that for most measures, the overall prediction for scaled scores was similar and superior for data that were originally skewed (see Table e-1; Figure e-1).

Models determination in reference sample

To determine which factors influenced followup NP performance in the HIV-controls, we conducted a series of hierarchical regression analyses using the followup scaled score as the outcome variable and a predetermined list of candidate predictors. The selection of these factors and their importance for followup NP performance (which informed their entry order into a hierarchical regression model [1]) was based on previous research literature in the study of longitudinal NP performance [3−6].

Infection with Hepatitis C Virus (HCV) was common in both the HIV− and HIV+ groups in Anhui, and was a significant contributor to baseline NP impairment status (see [2]). However, HCV infection was not retained as a predictor in the current longitudinal models because exploratory analyses determined that it was not associated with decline in any individual NP test measure or summary measure.

The predictor variables that were considered in the multivariate analyses included (in order of entry): 1. baseline performance on the individual NP measure in question (scaled scores), 2. Baseline overall NP competence (defined as the individual's mean scaled score at baseline including all NP measures *with the exception of the one being currently tested in the regression model);* 3. Test–retest interval (in days); 4. Age (in years), 5. Education (in years); and 6. Gender (male or female, coded as 0 and 1). Any predictive factor that accounted for significant R^2 change above and beyond the previously entered predictors ($p < .10$) was retained. Baseline performance was found to be highly predictive of followup performance for all the 17 NP measures (positive associations). Overall baseline competence added to the prediction of followup performance (better overall competence was associated with better followup performance) for Grooved Pegboard (both Dominant hand and non–dominant hand), PASAT 50, Animal Fluency, Category Test, Trail Making Test A, Color Trails 1 & 2, Digit Symbol, Symbol Search, Spatial Span and Brief Visuospatial Memory Test-Revised (BVMT-R) total learning. Test-retest interval never reached significance for prediction *(ps>.10)*. Age was a significant predictor for Digit Symbol, Trail Making Test A, Color Trails 1, and Symbol Search (negative associations or less "practice effect" in performance with older age), as well as Animal Fluency and Action Fluency (positive associations). Education was a significant predictor only for BVMT-R total learning (lower education was associated with more improved learning at followup). Sex was a significant predictor for PASAT 50 (males improved more than females), Digit Symbol (females improved more than males), and Stroop Color (females improved more than males).

Final reference models applied to the HIV+ sample

Significant predictors were retained in a series of standard multivariate regression analyses to derive regression formulas for each individual NP measure. These formulas served to compute the 17 individual predicted followup scaled scores (see (1)). Each formula was then applied to the HIV+ sample:

$$Y_p = \beta_1 X_1 + \beta_2 X_2 + \beta_n X_n + a \tag{1}$$

Where Y_p is the predicted scaled score at *followup*, β_1 is the regression coefficient (slope) for predictor X_1, β_2 for predictor $X_2,\ldots,$ β_n for predictor X_n X_1 is the observed baseline score, X_2 through X_n may include the overall competence, or demographic factors entered into the model, and a is the intercept.

The summary regression-based change score (sRCS)

The development of the *summary regression-based change score (sRCS)* follows the next three steps: 1. the standard deviation of the residuals (i.e., error term of the regression model in SD units) for each of the 17 final regression models in the HIV− sample was computed. 2. Seventeen individual Z-scores were computed, by dividing the difference between predicted and obtained followup scaled scores by the error term of the corresponding regression model (see (2)). The resulting Z-score reflects how well or poorly the participant did at the 12 months followup, relative to normal expectations for someone with his/her baseline score and other variable-specific baseline predictors:

$$Zscore = (X_o - X_p)/(SD_{residual}) \qquad (2)$$

In this formula, X_o is the obtained scaled score at followup; X_p is the predicted followup scaled score derived from the reference group regression equation (1), and $SD_{residual}$ is the standard deviation of the residuals from the reference HIV− group regression model. Note that this regression based change score can be negative (if the obtained followup score is worse than the predicted) or positive (if the obtained score is greater than the predicted score).

3. The final step in the development of the norms for change was to sum the 17 individual regression based change scores (Z-scores) to compute the sRCS and determine a 90% confidence interval to define "no change" on the test battery. That is, the cut-off for the top 5% of the sRCS distribution of the HIV− controls defined the "improved" range and the cut-off for the bottom 5% defined the "decliners" range. This was applied to the HIV+ sample (see Figure 1). For the following analyses, we used the sRCS to define NP decline as 1-tailed 95% confidence interval in the reference HIV− sample to classify decliners and nondecliners.

Extended data analysis

The HIV+ decliners and non–decliners (as defined by the sRCS) were then compared on baseline and followup demographic, HIV disease-related laboratory measures, AIDS status, Global Deficit Score (GDS), treatment-related variables, cognitive complaints, IADL, BDI-II using t-test and Chi-square as appropriate.

The Global Deficit Score (GDS) and ability domain T-scores at followup were corrected for practice effect using the following procedure. For each NP scaled score measure, we computed the *median* of the differences between the followup score and the baseline score in the HIV− sample. Then this median practice effect was subtracted from the actual followup score in all individuals, providing a "practice effect corrected" followup scaled score for both the HIV− and HIV+ participants; this corrected followup score represents our best estimate of what the test performance would have been, absent of any prior exposure to the test instrument. Finally, T-score transformations and deficit score transformations were then applied to the *corrected* scaled scores, following the same procedure described in detail for the baseline results [2,7,8] in order to compute the corrected followup ability domain T-scores and corrected GDS. Analyses were conducted using SPSS 16.0 version and JMP 7.0 version (SAS, Inc).

Results supporting the validity of the sRCS approach

Some practice effect was observed in both the HIV− sample [mean scaled score baseline: 10.04 ± 1.81; mean scaled score followup: 10.94 ± 1.66; Mean diff $= 0.91 \pm 0.98$, p < .0001; Dunlap's d = .52] and the HIV+ sample [mean scaled score baseline: 8.71 ± 2.00; mean scaled score followup: 9.36 ± 1.96; Mean diff $= 0.65 \pm 1.13$, p < .0001; Dunlap's d = .33].

See also Figure e-2 for illustration of mean scaled score change from baseline to follow-up in the HIV− sample, the HIV+ nondecliners sample and the HIV+ decliners sample.

Our results demonstrate that the regression change score approach can be validly applied to NP data from a non–Western context. Specifically, by deriving norms from the same risk group (here HIV− former plasma donors), we were able to detect disease-related cognitive change over a one year followup interval. Indeed, our estimate of cognitive decline was more than fives times greater in the HIV+ than in the HIV− group, and affected a sizeable minority of the HIV+ sample. We further demonstrated that without taking into account practice effect correction (derived from the control group), impairment at followup may be wrongly estimated in 18% of cases (in the total HIV+ group, 25% were NP-impaired using uncorrected followup scores and 43% using the corrected followup scores). Also without correcting for practice effect, our estimates of NP-impairment in the HIV+ sample at followup would be lower by more than 10% compared to what has been observed cross-sectionally in this same sample at baseline (i.e., 37%); this cannot reflect real improvement since the HIV+ group gave no indication of improving using norms for change developed with the stable HIV− group. Moreover, the approximate 15% rate of impairment in controls that

should be projected per our use of a one–SD normative cut–off would falsely disappear [9]. This suggests that the failure to account for practice effects in longitudinal NP studies is highly likely to reduce sensitivity to impairment after baseline.

References

1. Tabachnik BG, Fidell LS. **Using Multivariate Statistics**. 5th ed. Boston: Person International Edition; 2007.
2. Heaton RK, Cysique LA, Jin H, Shi C, Yu X, Letendre S, *et al*. **Neurobehavioral effects of human immunodeficiency virus infection among former plasma donors in rural China**. *J Neurovirol* 2008; **7**:1−14.
3. Basso M, Bornstein R, Lang J. **Practice effects on commonly used measures of executive function across twelve months**. *Clin Neuropsychol* 1999; **13**:283−292.
4. Duff K, Beglinger LJ, Schultz SK, Moser DJ, McCaffrey RJ, Haase RF, *et al*. **Practice effects in the prediction of long-term cognitive outcome in three patient samples: a novel prognostic index**. *Arch Clin Neuropsychol* 2007; **22**:15−24.
5. Heaton R, Temkin N, Dikmen S, Avitable N, Taylor M, Marcotte T, **Grant I Detecting change: A comparison of three neuropsychological methods, using normal and clinical samples**. *Arch Clin Neuropsychol* 2001; **16**:75−91.
6. Levine A, Miller E, Becker J, Selnes O, Cohen BA. **Normative data for determining significance of test–retest differences on eight common neuropsychological instruments**. *Clin Neuropsychol* 2004; **18**:373−384.
7. Carey CL, Woods SP, Gonzalez R, Conover E, Marcotte TD, **Grant I, Heaton RK. Predictive validity of global deficit scores in detecting neuropsychological impairment in HIV infection**. *J Clin Exp Neuropsychol* 2004; **26**:307−319.
8. Heaton R, Miller W, Taylor MJ, Grant I. *Revised Comprehensive Norms for an Expended Halstead Reitan Battery: Demographically Adjusted Neuropsychological Norms for African American and Caucasian Adults* Lutz, FL: Psychological Assessment Resources; 2004.
9. Taylor MJ, Heaton RK. **Sensitivity and specificity of WAIS–III/WMS–III demographically corrected factor scores in neuropsychological assessment**. *J Int Neuropsychol Soc* 2001; **7**:867−874.
10. Steiger JH. Statpower: **Intelligent Decisions Through Statistical Analysis, R2**. In: http://www.statpower.net/index.html; 2009.

Table e-1: Multiple correlation coefficient (R^2) and 95% confidence interval (CI) for raw scores* versus scaled scores standard multiple regression models in 101 HIV- individuals.

	Raw scores R^2	95% CI R2test[1]	Scaled scores R^2	95% CI R2test[1]
Color Trail II	0.616	[.476−.717]	0.605	[.463− .708]
Category Test	0.415	[.254 .589]	0.380	[.220−.517]
Animal fluency	0.250	[.100− .385]	0.246	[.100− .381]
Action fluency	0.166	[.050− .301]	0.156	[.040− .290]
PASAT-50	0.470	[.304 − .593]	0.450	[.283 − .575]
WMS-III Spatial Span	0.340	[.182 .480]	0.350	[.191 .490]
HVLT-R Learning	0.265	[.125− .414]	0.263	[.123− .412]
BVMT-R Learning	0.292	[.132− .428]	0.305	[.151− .467]
HVLT-R Delayed Recall	0.233	[.100− .381]	0.216	[.090− .364]
BVMT-R Delayed Recall	0.394	[.241− .535]	0.356	[.205− .500]
Grooved Pegboard DH	0.364	[.197−.497]	0.385	[.225 − .522]
Grooved Pegboard NDH	0.436	[.268 − .563]	0.440	[.287 − .575]
WAIS-III Digit Symbol	0.752	[.642−.818]	0.766	[.659− .829]
WAIS-III Symbol Search	0.368	[.201− .501]	0.387	[.219−.518]
Trail Making Test A	0.384	[.216−.516]	0.363	[.196−.496]
Color Trail I	0.567	[.419−.678]	0.515	[.353− .631]
Stroop color	0.608	[.467−.711]	0.582	[.436− .690]

*In these analyses, raw scores were winsorized. The distribution of the 17 NP measures at baseline and followup (raw scores) for the 101 HIV− was checked using Normal Quantile Plots. When outsider values were observed we applied a 98% 2-tailed winsorization (Grooved Pegboard non-dominant hand, Action Fluency, the Color Trail 1 and 2 at baseline and Action Fluency, Color Trail 1, Digit Symbol, Spatial Span at follow-up). Then using the winsorized scores, we determined which NP factors or demographic characteristics influenced follow-up performance by conducting a series of hierarchical regression analyses in a similar fashion as detailed in data analysis section for the scaled scores. Similarly to the method used for the scaled scores, the significant predictors $(p < .10)$ were retained in a series of standard multivariate regression analyses to derive regression formulas for each individual NP measure (for which the R^2 is displayed in this table). [1]R2 Test: R2 is a computer program, which implements some statistical procedures not generally available for inference on multiple correlation coefficients. R2 allows calculation of confidence intervals around R^2 (Steiger, 2009 <http://www.statpower.net/index.html> [10]). CI computations are based on number of observations (101) and number of variables in the model including depending variable.

Table e-2: Uncorrected and corrected scaled scores, GDS and impairment status at baseline and follow-up and in HIV− and HIV+ persons who decline versus those who did not decline.

		HIV− (n = 101)			HIV+ decliners (n = 53)			HIV+ non-decliners (N = 139)		
		Baseline	FU	Corrected FU	Baseline	FU	Corrected FU	Baseline	FU	Corrected FU
Executive Function	Color Trails II	9.75 (3.40)	10.76(3.50)	9.76 (3.50)	8.21 (3.76)	7.26(3.41)	7.26(3.41)	8.67 (3.37)	9.60 (3.29)	8.60 (3.29)
	Category Test	10.59(3.18)	11.82(2.85)	10.82(2.85)	8.55 (3.28)	9.08 (2.78)	9.08 (2.78)	9.03 (3.63)	10.81 (3.14)	9.81 (3.14)
Verbal Fluency	Animal fluency	10.17(2.88)	10.77(3.02)	9.77 (3.02)	8.38 (2.88)	7.68 (2.72)	6.68 (2.72)	9.26 (2.95)	9.84 (2.92)	8.84(2.91)
	Action fluency	10.15(2.89)	10.90(3.04)	10.90(3.04)	9.50 (3.07)	8.87(3.11)	8.87(3.11)	9.06(3.31)	10.00(3.03)	10.00(3.03)
Attention/WM	PASAT-50	10.15(3.20)	11.54(3.41)	10.54(3.41)	8.36 (2.59)	7.87 (2.28)	7.87 (2.28)	9.04 (3.29)	10.23(3.25)	9.23 (3.25)
	WMS-III Spatial Span	10.05(2.76)	10.44(2.86)	10.44(2.86)	8.36 (3.08)	8.24 (2.42)	8.24 (2.42)	9.29 (2.97)	9.84 (3.05)	9.84(3.05)
Learning	HVLT-R Learning	9.68 (3.06)	10.79(3.39)	9.79 (3.39)	8.64 (3.39)	7.58 (2.37)	7.58 (2.37)	8.82 (3.42)	9.88(3.10)	9.88(3.10)
	BVMT-R Learning	10.36(3.05)	11.49(2.85)	10.49(2.85)	7.83 (2.72)	7.55 (2.63)	6.55 (2.63)	8.83 (3.29)	10.31 (3.00)	9.31 (3.00)
Memory	HVLT-R Delayed Recall	9.70 (3.05)	10.52(2.99)	0.52 (2.99)	8.41 (3.14)	6.66(3.12)	5.66(3.12)	8.77 (3.34)	9.74 (3.28)	8.74 (3.28)
	BVMT-R Delayed Recall	10.13(3.29)	11.14(2.98)	10.13(2.98)	7.34 (2.90)	6.86 (2.60)	5.87 (2.60)	8.51 (3.20)	9.56 (2.94)	8.56 (2.95)
Motor	Grooved Pegboard DH	9.66 (2.78)	10.60(3.00)	9.60 (3.00)	8.55 (2.98)	8.45 (2.94)	7.45 (2.93)	8.87(3.14)	10.31 (3.06)	0.31 (3.06)
	Grooved Pegboard NDH	9.86 (3.03)	10.20(3.01)	10.20(3.01)	8.88(3.15)	8.00(3.01)	8.00(3.01)	8.63(3.10)	9.65 (3.05)	9.65 (3.05)
SIP	WAIS-III Digit Symbol	10.19(2.98)	10.86(3.01)	0.86(3.01)	7.41 (2.55)	7.11 (2.23)	6.11(2.34)	8.88(3.14)	9.53 (3.04)	8.53 (3.04)
	WAIS-III Symbol Search	9.93 (2.86)	10.59(2.79)	9.59 (2.79)	7.26(2.18)	7.43 (2.23)	6.43 (2.23)	8.68 (2.60)	9.39 (2.80)	8.39 (2.80)
	Trail Making TestA	9.96 (3.02)	11.92(3.03)	9.92 (3.03)	7.94 (3.08)	7.87 (2.36)	5.87 (2.36)	8.93 (3.30)	10.86(3.20)	8.86 (3.20)
	Color Trails I	9.90(3.13)	11.17(2.97)	10.17(2.97)	7.64 (2.72)	7.56 (2.84)	6.56 (2.84)	9.24 (3.43)	10.31 (3.14)	9.31(3.14)
	Stroop color	10.39(2.93)	10.54(3.05)	10.54(3.05)	7.64 (2.69)	7.55 (2.52)	7.55 (2.52)	8.93 (3.36)	9.77 (2.86)	9.77(2.86)
GDS		0.22 (0.26)	0.12(0.14)	0.21 (0.22)	0.52 (0.45)	0.60 (0.37)	0.91 (0.48)	0.48 (0.50)	0.24 (0.29)	0.40 (0.40)
% Impaired		13.8%	2%	7.0%	37.7%	51%	79.2%	36.7%	15%	28.8%

FU: Follow-up; SIP: Speed Information Processing; WM: Working Memory; HVLT-R: Hopkins Verbal Learning Test-Revised, BVMT-R: Brief Visuospatial Memory Test-Revised Corrected FU= scaled scored at follow-up - HIV- median practice effect GDS: Global Deficit Score When available, alternate versions of the NP tests were used in order to minimize practice effects (i.e., Hopkins Verbal Learning Test-Revised, Brief Visuospatial Memory Test-Revised)
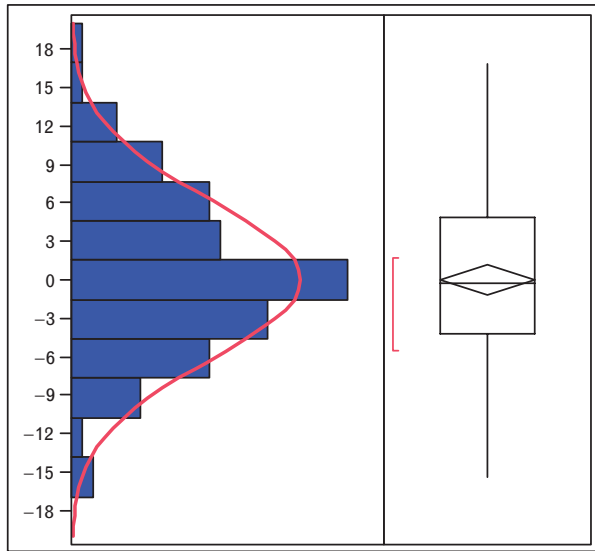
Figure e-1: Summary regression change score (sRCS) distribution in 101 HIV− former plasma donors in Anhui. Mean = 0; SD = 6.25. To define a 90% confidence interval in this sample, 5% with the most negative scores (cut off <-9.2) and 5% with the most positive scores (>10.5) were defined. This was then applied to the HIV+ sample.
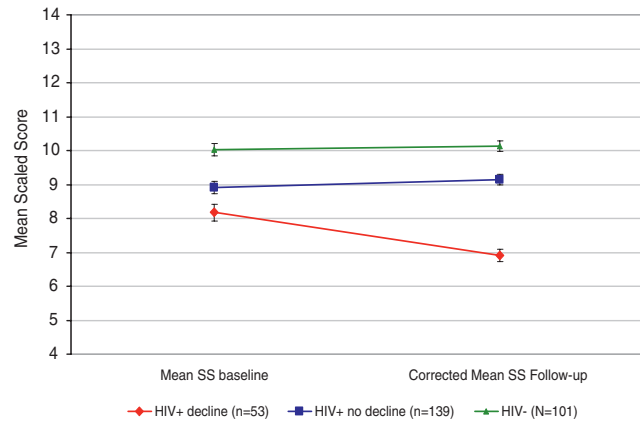

Figure e-2: Mean scaled score at baseline and practice effect corrected mean scaled scores at followup for the HIV+ decliners and non–decliners, and the HIV– group. Correction for practice effect was made using the *median* difference between baseline and followup in the HIV− sample and then applied to the HIV+ sample.