PNAS
www.pnas.org

Supplementary Information for

Using a Generative Model of Affect to Characterize Affective Variability and its Response to Treatment in Bipolar Disorder

Erdem Pulcu PhD[1], Kate E.A. Saunders MD[1,2], Catherine J. Harmer PhD[1,2], Paul J. Harrison MD[1,2], Guy M. Goodwin MD[1,2], John R. Geddes MD[1,2], Michael Browning MD[1,2*]
[1]Department of Psychiatry, University of Oxford, UK
[2]Oxford Health NHS Trust, Warneford Hospital, Oxford, UK
*michael.browning@psych.ox.ac.uk, +447961340428

This PDF file includes:

In the supplementary materials we provide a formal description of the Bayesian filer, explore the effect of alternative link functions, assess the performance of the filter and describe a series of sensitivity and exploratory analyses that assess potential confounding factors and alternative explanations for the results reported in the main paper.

Supplementary Methods

The Bayesian Filter

Overview: A grid-based recursive Bayesian filter (1) was developed to estimate the causal processes described in the generative model (Figure 1). Note that the filter has no free parameters and so is not fitted to participant ratings, rather it uses the ratings to estimate the values of the causal processes defined by the generative model.

As illustrated in Figure 1a, the filter assumes that affect ratings at a given time point t, y_t, are generated from a Gaussian distribution with an unknown mean, mu_t, and standard deviation, $\exp(SD_t)$, with the later producing noise in the ratings (Figure 1b). Note that all variance parameters in the filter are represented in log space to ensure they lie on the infinite real line and that it is these log transformed parameters that are analysed in the current paper. Formally, the production of ratings is described by:

$$y_t \sim \mathcal{N}(mu_t, \exp(SD_t)) \quad (1)$$

The mean of this distribution may change between time points, leading to volatility of the ratings (Figure 1b), with this change described by a second Gaussian distribution, centered on the current mean and with a standard deviation of $\exp(vmu_t)$.

$$p(mu_{t+1}) \sim \mathcal{N}(mu_t, \exp(vmu_t)) \quad (2)$$

Both the SD_t and vmu_t parameters can also change between time points with their change governed by Gaussian distributions centred on their current value with standard deviations of $\exp(vSD)$ and $\exp(kmu)$ respectively. These higher-level parameters allow the model to conceptualise periods in which noise and volatility are high and other periods in which they are low.

$$p(vmu_{t+1}) \sim \mathcal{N}(vmu_t, \exp(kmu)) \quad (3)$$

$$p(SD_{t+1}) \sim \mathcal{N}(SD_t, \exp(vSD)) \quad (4)$$

The filter seeks to estimate the joint posterior probability of the five causal parameters, given the affect ratings it has observed. The joint probability distribution at time point t is defined as:

$$p(joint_t) = p(mu_t, vmu_t, kmu, SD_t, vSD \mid y_{t-1}, y_{t-2}, \ldots y_1) \quad (5)$$

This joint probability distribution can be thought of as the filter's belief about the values of each parameter in the generative model. The filter uses this distribution as a sufficient statistic of the affect generating process (i.e. it assumes that the magnitude of future affect

ratings is completely described by this distribution). Formally, this is equivalent to a Markovian assumption (i.e. that the joint probability on the next trial depends only on the joint probability of this trial and the observed rating) so that that recursive update performed by the filter can be stated as:

$$p(joint_t) = p(mu_t, vmu_t, kmu, SD_t, vSD \mid y_{t-1}, joint_{t-1}) \qquad 6$$

We initialize the joint posterior, before observation of any ratings, $p(joint_1)$, as a flat distribution. In the next section we describe how the recursive updates are performed between time points.

Belief update1: the effect of the observed rating: The update of the joint probability distribution between time points is split into two broad components. First the filter has to use the rating it observes to update its belief. This is achieved using Bayes rule:

$$p(joint_{t-1} \mid y_{t-1}) = \frac{p(y_{t-1} \mid joint_{t-1})\, p(joint_{t-1})}{p(y_{t-1})} \qquad 7$$

The important points to note here are 1) the likelihood term, $p(y_{t-1} \mid joint_{t-1})$, is a function only of the mean, mu_t, and standard deviation, SD_t, of the generative model and 2) the process described by this equation does not account for the change in values of the mu_t, vmu_t or SD_t parameters between time points described by equations 2-4.

Belief update2: the effect of parameter shifts: The next component of the update deals with the change in the mu_t, vmu_t or SD_t parameters across time. In order to illustrate this in an intuitive fashion, we first describe the update of a single parameter, mu_t, from time $t-1$ to time t, given the volatility, vmu_t.

$$p(mu_t \mid vmu_t) = \int p(mu_t \mid mu_{t-1}, vmu_t)\, d(mu_{t-1}) \qquad 8$$

The integral in this equation sums together the probabilities, across every possible starting value of the mean, mu_{t-1}, that would have led it now to be the new value, mu_t, as defined by $p(mu_t \mid mu_{t-1}, vmu_t)$. In this way the original value of the mean is "integrated out" and the probability distribution across values of mu_{t-1} is updated. The same process is applied to the other drifting parameters in the model and is combined with the update in response to the observed rating described in equation 7 to give:

$$p(joint_t \mid y_{t-1}) = \iiint p(joint_{t-1} \mid y_{t-1}) p(SD_t \mid SD_{t-1}, vSD)\, p(vmu_t \mid vmu_{t-1}, kmu) \ldots p(mu_t \mid mu_{t-1}, vmu_t), dSD_{t-1}, dvmu_{t-1}, dmu_{t-1} \qquad 9$$

The recursive nature of the filter and the separation of the updating process into two components provides a natural approach for dealing with missing data. On days when no rating was returned the first phase of the update (equation 7) is omitted, but the second phase (remaining terms in equation 9) is applied, leading to a dispersion of the filter's

belief about the current parameter values (see supplementary methods for an illustration of this and for a sensitivity analysis).

The filter's belief about the value of each node is derived at every time point by marginalising over all but the relevant dimension of the joint probability distribution and calculating the expected value of that dimension.

Note that the filter estimates magnitudes in an unbounded space, whereas affect ratings are produced on a bounded scale (e.g. from 1-7). The magnitude ratings were therefore logit transformed before being passed to the filter. The transformation was achieved by first scaling the ratings by dividing by 8 (i.e. so the ratings range from 1/8 to 7/8), and then transforming this $scaled_{score}$ score using the logit function:

$$transformed_{score} = -\ln(\left(\frac{1}{scaled_{score}}\right) - 1)$$

An alternative approach to this issue is to specify an alternative link function for the filter. In the next section we describe such a link function and show that it produces equivalent results.

The filter was implemented as a five dimensional matrix in matlab (R2020). The filter's code is available at: https://osf.io/j7md3/

Supplementary Results

The demographic details of participants in both studies is summarized in Table S1 below.

Table S1. Demographic details of patients included in the two studies.

Cohort Study				
	Bipolar Disorder	**Borderline Personality Disorder**	**Control**	**p value for group difference**
n	51	33	51	
Sex (F/M)[a]	32/19	30/2	32/18	0.004*
Age (mean(SD))	39.47 (13.03)	33.72 (10.42)	38.18 (12.98)	0.12
Educational Achievement (mean(SD))[b]	4.81 (0.96)	4.1 (0.88)	5.1 (1.02)	<0.001*
Previous Hospitalisation (Y/N)	21/28	14/17	0/48	

Subtype of Bipolar Disorder (I/II)	30/17	N/A	N/A	
On Any Psychoactive Medication (Y/N)	48/3	24/8	1/50	
On Lithium (Y/N)	22/29	0/32	0/51	
On Anticonvulsant (Y/N)	19/32	1/31	0/51	
On Antipsychotic (Y/N)	34/17	6/26	0/51	
On Antidepressant (Y/N)	16/34	23/9	1/50[d]	
On Anxiolytic (Y/N)	2/48	8/24	0/51	
On Hypnotic (Y/N)	4/46	2/29	0/51	
Current or Previous Psychotherapy (Y/N)[e]	42/7	31/1	11/33	
Current or Previous Recreational Drug Use (Y/N)	29/18	21/10	12/38	

Experimental Medicine Study		
	Lithium Group	**Placebo Group**
N	19	16
Sex (F/M)	8/11	7/9
Age in years (mean(SD))	28.84 (9.81)	35.14 (13.79)
Diagnosis (BP I/BP II/BP NoS)	3/16/0	3/11/2

[*] Borderline group differed significantly from both other groups. Bipolar and control groups did not differ. [a] Demographic data was missing from some participants: one patient in the borderline group did not provide any demographic data, data on educational and employment level was missing from 3 patients in the bipolar group, and 2 patients in the borderline group. Binary outcomes (e.g. sex, medications taken etc) are reported for all participants who provided the relevant data. [b] Educational achievement was rated on a six-point scale from primary school to post graduate level. SD= Standard deviation, BP I/II/NoS = Bipolar disorder type 1, type 2, not otherwise specified. [d] One participant in the control group was taking low dose amitriptyline for pain. [e] Including any form of therapy or counselling. Note participants were excluded from the experimental medicine study if they were taking other psychoactive medications.

<u>An Alternative Link Function</u>

While the filter described above acts on data points from the infinite real line, the affective ratings collected from participants are constrained between the upper and lower bounds of the scale used (e.g. the individual items from the MoodZoom Scale require discrete responses between 1-7, which are averaged into positive or negative scores). A logit transform of the data is therefore used before it is presented to the filter. An alternative approach to this issue is to specify a link function which explicitly describes how the ratings are generated. For ordered, discrete responses, one such link is the ordered-probit function (2). In brief, this assumes that a gaussian latent process underlies the (discrete, ordinal) data, but that the data are generated using a set of cut points of this function (e.g. if data take the form 1:n, then n-1 cut points are defined with samples below cut point one producing a data value of "1", between cut points 1-2 producing a value of "2" etc). We incorporated a version of the ordered-probit link function to the Bayesian filter (details of function provided below). In figure S1, we show that the results using this link function on untransformed data (Figure S1 c-d) is similar to the filter described in the paper on the transformed data (Figure S1a-b). As can be seen, using the ordinal probit link function resulted in the same pattern of diagnostic specificity for affective variability (Figure S1c-d; Group x Type of Variability $F(2,122)=7.84$, $p<0.001$) as that reported in the main paper (data reproduced in Figure S1a-b), with the estimates of the two versions of the filter correlating at $r>0.8$ for volatility and at $r>0.75$ for noise. This suggests that the results reported in the main paper are not sensitive to the specific approach used to transform the data.

Details of Ordinal Probit Link Function: Ordinal probit functions have a number of different parameters that may be varied (e.g. the position of the cut points, the position and variance of the latent process). If all of these are allowed to vary simultaneously, the function is not identifiable (e.g. increasing the SD of the latent process is equivalent to moving the cut points closer together). This requires some of the parameters to be set to specific values. For example, for analyses of stationery processes the value of the lower cut point is often set to 0, and the SD of the latent process is set to 1. In the current analyses we are interested in how the SD and the variability of the mean of the latent process differ between groups and therefore cannot fix these. We therefore fix the positions of the cut points to the set values of 1,2,3,...n-1. This is implemented in the filter by adjusting the calculation of the likelihood (Equation 7), with the likelihood of a data point between two cut points (e.g. lying in the range 1-2) being the total probability density of the latent process between these two values.

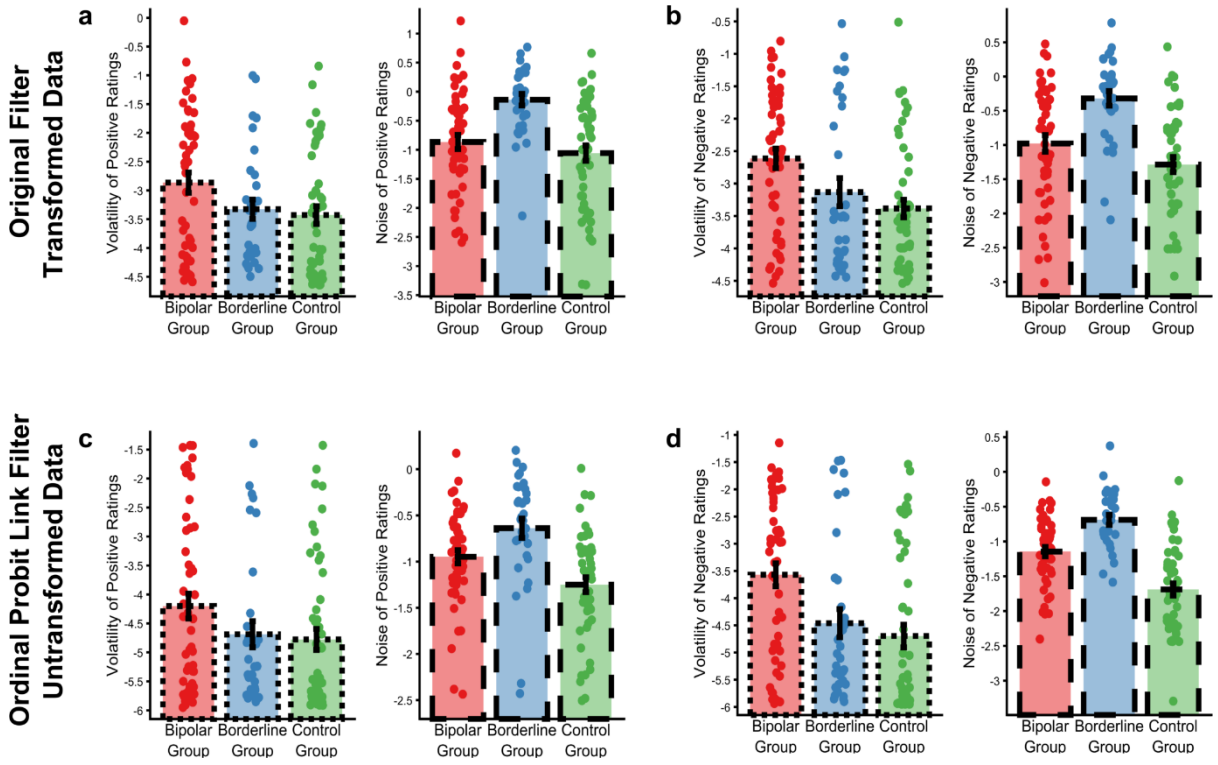

Figure S1: Estimated types of affective variability from the cohort study when the Bayesian filter is applied to transformed ratings data (panels a-b, reproduced from the main paper) and when an ordinal probit link function is used, so that the filter may be applied to untransformed data (panels c-d). As previously, the volatility and noise for the positive and negative ratings are presented separately. Both versions of the filter estimate volatility to be raised in patients with a diagnosis of bipolar disorder, with noise raised in patients with a diagnosis of borderline personality disorder. Data from the bipolar group is summarized with red, the borderline group with blue and the control group with green bars. Bars reporting volatility have dotted edges, those reporting noise have dashed edges.

<u>Analysis of Female Participants</u>

As described in the main text, the groups from the cohort study were not matched for gender, with all but two participants in the borderline personality group being female. In order to assess whether this might account for the reported group differences in affective variability, a sensitivity analysis was run in which only female participants were included. This analysis was identical to that reported in the main paper, with the exception that gender was not included as an explanatory variable. The analysis included 29 participants from the bipolar group, 28 from the borderline group and 32 from the control group. The same relationship between participant group and type of affective variability was found [group x type of variability; $F(2,84) = 3.1$, $p = 0.05$] in this smaller sample.

Analysis of the influence of zero scores on the reported results

"Zero-inflation" is defined as an excess of zero values, over that which would be expected given the range of scores recorded from a participant (3). By this definition, zero-inflation is produced by the behavior of the participant, who returns a rating of 0, when a higher rating is expected. A related but distinct issue is that, when a participant frequently provides an answer at the bound of a scale (e.g. 0), then the apparent variability of the scores is reduced, which may skew measures such as those provided by the filter. Note that this second process is a function of the scale rather than the participant and is true whether or not the 0 scores are "inflated" relative to the participant's other answers. In the following section we assess these two issues separately.

As a point of clarification, the data analyzed in the paper consist of summary positive and negative scores derived by averaging a number of different positive and negative ordinal questions. A first question is therefore to define what a zero-score is on this scale. In the following analysis all data are rounded to the nearest integer (having subtracted 1, so that the minimum score is 0) and thus a zero score is defined as any score that is closer to 0 than to 1. We believe that this approach provides the most conservative estimate (as it treats all low scores as zero, rather than just those scores on which a participant answered 0 for every sub-question) and also allows a formal test of the deviation from the expected rate of zero scores relative to a Poisson distribution.

Do participants show formal zero-inflation and how does this impact the study results?

We derived a measure of zero inflation using the score statistic described by van den Broeck et al. (3). The logic of this test is that counts of scores are assumed to follow a stationery Poisson distribution, with a lambda parameter set to the observed mean of the scores. This generates an expected number of 0 scores, which can be compared to the observed value of 0 scores in a test statistic (described in van den Broeck (3)) with a chi-squared distribution. Applying this test to the 135 participants from the current study, indicated significant 0 inflation (i.e an excess of observed 0s with an uncorrected p value of <0.05 at the participant level) for 10 participants for the positive ratings (4 in bipolar group, 4 in borderline group and 2 in control group) and 7 participants for negative ratings (2 in bipolar group, 2 in borderline group and 3 in control group). Rerunning the main analysis excluding any participant who had inflation in either positive or negative scores (14 participants were excluded as 3 had inflation of both positive and negative scores) did not change the observed pattern of results from the study [group x type of variability; $F(2,108) = 7.61$, $p < 0.001$].

In summary, there was no strong evidence for zero-inflation in participant scores, and excluding those participants who did show evidence for this did not change the pattern of results reported.

What proportion of scores are 0, how is this related to the filter metrics and can it account for the study results?

Whether or not a participant's responses contain more zeros than expected, the bounded nature of the rating scales used mean that the apparent variability is likely to be reduced

when ratings are frequently zero (this was a motivation for assessing the probit model described above). In order to further investigate the impact of zero scores, we summarise their frequency in the three groups, assess their relationship with the filter derived metrics and their impact on the reported results.

Table S2 reports the proportion of the 50 days in which a zero rating was returned in the three cohorts (note, if participants were equally likely to give a rating at the 7 different levels of the scale, the expected proportion of zero responses would be 0.14). As can be seen, control participants were much more likely to give 0 ratings on the negative scale than other participants or for positive ratings in any group.

Table S2: Proportion of zero scores across the 50 days of the study by diagnostic group

	Average proportion of 0 scores for positive ratings	Average proportion of 0 scores for negative ratings
Bipolar group	0.15	0.22
Borderline group	0.16	0.06
Control group	0.09	0.57

Table S3 reports the within group correlations between the proportion of zero scores returned and the filter derived metrics. A general pattern of negative association is observed (i.e. the more zero answers present, the lower the estimated variability), with a stronger effect for measures of noise than volatility. As expected from Table S2, negative ratings in the control group show a particularly strong association, including with negative volatility.

Table S3: Pearson correlation coefficients between the proportion of zero responses from a rating and the filter derived metrics from the same ratings

	Positive volatility	Negative volatility	Positive noise	Negative noise
Bipolar group	-0.26	-0.04	**-0.35**	-0.23
Borderline group	-0.15	0.14	**-0.68**	0.34
Control group	0.1	**-0.6**	-0.22	**-0.64**

Numbers in bold are statistically significant at $p<0.05$ (with no correction for multiple comparisons)

The above results indicate that, as expected, a high proportion of zero scores will tend to result in reduced estimates of variability, particularly for noise. This effect is more pronounced for negative ratings in the control group. As a final step we assessed whether adding a participant's proportion of zero ratings as covariates to the study analysis influenced the reported pattern of results. The same pattern of results reported in the main paper were obtained when these additional covariates were added [group x type of variability; $F(2,120) = 6.34$, $p = 0.002$], with the expected effect of the covariates themselves apparent as main effects [main effect of proportion of positive zero responses; $F(1,120) = 4.35$, $p = 0.04$, main effect of proportion of negative zero responses; $F(1,120) = 10.4$, $p = 0.002$].

In summary, the bounded nature of the rating scale used had the expected effect on filter derived measures of affective variability. This was particularly prominent for measures of noise and for negative ratings in the control group. Consistent with the results obtained from the probit link function, we did not find evidence that the reported results from the paper were related to the frequency with which participants returned zero responses.

Performance of the Filter

The ability of the filter to recover parameters, when those parameters are drifting across time, was assessed using a generate recover procedure. The generative model described in main Figure 1 was used to produce 90 sets of artificial data. This was achieved by selecting a starting state of the generative model by randomly sampling the state of each model node from a Gaussian distribution with mean and SD set to those of the cohort sample of participants. The generative model was then allowed to run for fifty time points ("days"), with the values of the mean, volatility and noise drifting between time points as described by the generative model (note that the other nodes do not drift). Synthetic affective ratings were sampled at each time point from the model's mean and SD on that day. These data were then fed back to the Bayesian filter. The performance of the filter was assessed at day fifty (i.e. the same timepoint used in the study), by correlating the actual value of each drifting parameter with that recovered by the filter. The results of this process are illustrated in Figure S2. As can be seen, all three parameters are generally well recovered, with very low levels of volatility being the least accurately reproduced. This suggests that the filter is able to estimate these dynamic processes with a reasonable precision.

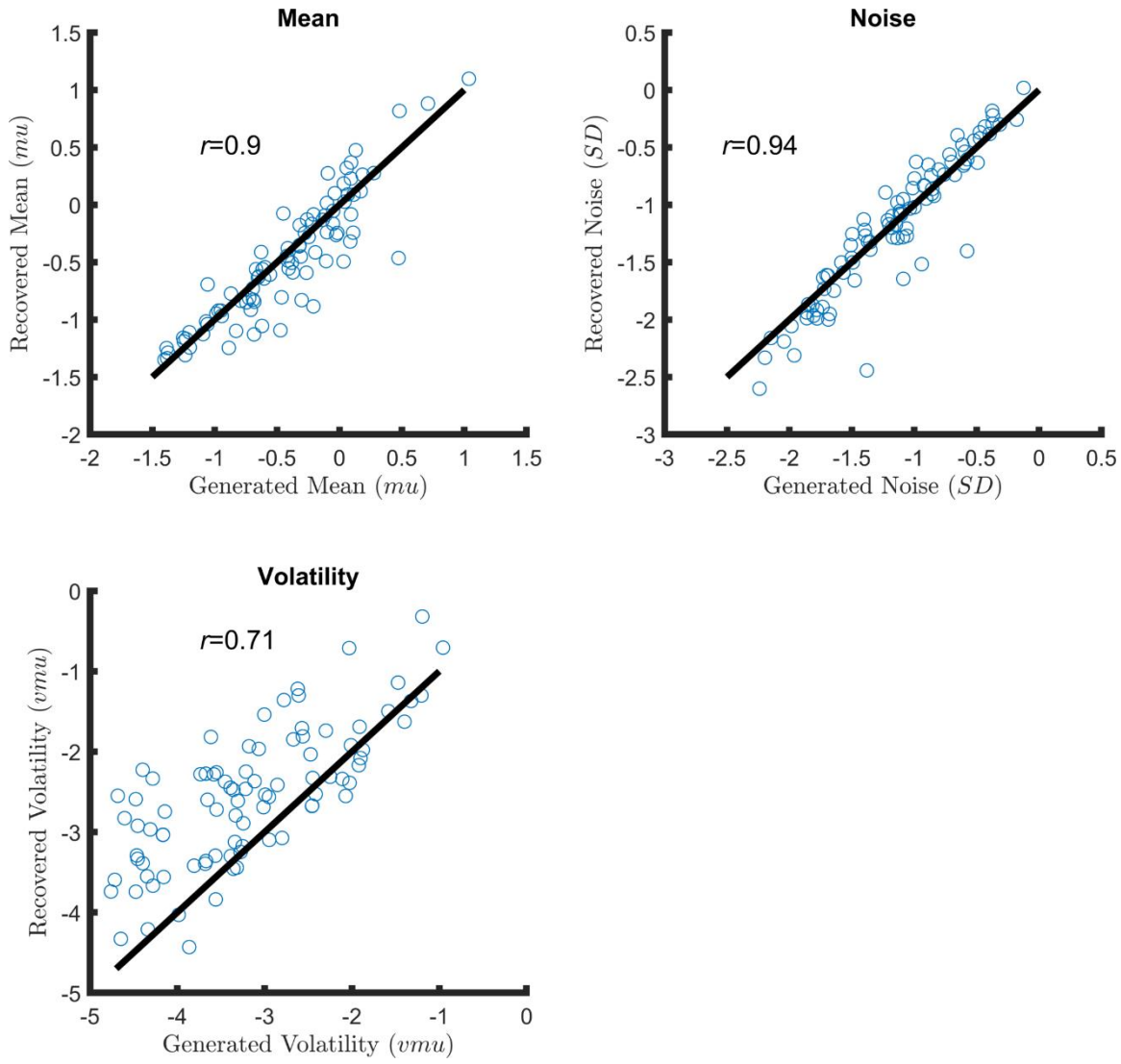

Figure S2: Results of a generate recover process. 50 days of artificial data were created for 90 pseudo-participants using the generative model described in Figure 1. The starting state of the model was set using the values from study participants, with the values of the mean (mu), noise (SD) and volatility (vmu) nodes allowed to drift over time as specified by the generative model. The filter was applied to the data generated using this process, with the filter-based estimates of mean, noise and volatility on day 50 correlated with the actual values of these parameters. As can be seen, the recovered values for the three parameters track the true underlying values. Black lines on each plot are y=x, that is, the line of perfect recovery.

Comparison of the filter with other measures of affective variability

A range of other measures of affective variability have been described (4). The measures which are most closely linked to those described in the current paper are affective inertia, commonly defined as the slope of an AR1 model, which is similar to the volatility term of the filter, and the standard deviation of the residual data, having fit the AR1 model, which is similar to the noise of the filter. Deriving the individual slopes from an AR1 model applied to data from the cohort study, as well as the standard deviation of the residuals, confirmed the overall results from Bayesian Filter with a significant group x variability type interaction ($F(2,114)=13.3$, $p<0.01$; note that missing data, or low levels of variability on one of the scales prevented fitting of the AR1 model in 8 participants, and so reduced the size of the groups by 2 in the group of patients with bipolar disorder, 1 in the group of patients with borderline personality disorder and 5 in the control participants). Post hoc tests found a significant difference in AR1 slope between the bipolar and control groups ($p_{bonf}<0.009$), although the difference between the bipolar and borderline groups found for volatility was not replicated (whether corrected or not for multiple comparisons). Similar to the volatility results, the AR1 slope did not differ between borderline and control groups. Group differences for the standard deviation of the residuals were consistent with those of noise from the filter, with all group differences being significant at $p_{bonf}<0.045$.

Analysis of metrics from the AR1 model for the experimental study was limited as missing rating data prevented the AR1 model being applied in five participants. The overall group x type of variability x valence effect here was not significant $F(1,28)=2.17$, $p=0.15$.

Overall these results indicate a degree of similarity between the filter derived measures and those obtained using an AR1 model, although there are differences, particularly in how the two approaches deal with missing data (see below for further analysis of missingness).

A second model, that bears some similarity to the filter reported here, is the DynAffect model reported by Kuppens and colleagues (5), which considers changes in affect from a dynamical systems perspective. The DynAffect model incorporates a number of sophisticated features, such as treating time as continuous (rather than as discrete, as done here), capturing the interaction between the valence of emotion and its intensity and modelling the effect of time varying covariates, which make it difficult to directly compare to the filter used in this paper. However, the DynAffect model conceptualizes changes in affect as arising from two main processes; first perturbations occur which move affect away from a set point, with the size of these perturbations being controlled by the parameter gamma; and second, an attractive influence pulls affect back to the set point and is controlled by the parameter beta. This framing is conceptually similar to that used for the Bayesian filter, with the gamma parameter being somewhat similar to noise and the beta parameter being somewhat similar to (inverse) volatility. It would therefore be of interest to assess whether the effects reported in the current paper are apparent in data appropriate to the DynAffect model.

Missing data

Treatment of Missing Data by the Filter: When the filter encounters missing data it omits the Bayesian update defined in equation 7 but does apply the shift in parameter values defined by the generative model, as described by the remaining terms in equation 9. The

effect of missing data is therefore to cause the filter to become progressively less certain about the values of the generative parameters. The rate at which this reduction of certainty occurs is rationally influenced by the model's belief about how changeable the parameters are. This process is illustrated with synthetic data in Figure S3 panels a and b for the mu_t parameter. The solid black lines in these plots are the data fed to the filter, which is either volatile (panel a) or stable (panel b). The data from time point 50 onwards is missing. Superimposed on the line is the marginal distribution of the filter's belief about mu_t. As can be seen, this belief is concentrated around the observed data. The key aspect of these figures is what happens to the filter's belief when it encounters missing data. In panel a, where the filter has learned that the data is volatile, its belief quickly disperses so it becomes rapidly uncertain about the likely value of mu_t. In contrast, when the filter has learned that the data is stable, it maintains its belief about the value of mu_t for a longer period. Thus the filter appropriately adapts its belief about the generative parameters of the data it observes as a function of its estimates about the changeability of these parameters.

Missing Data Sensitivity Analysis: In the cohort study there was no overall group difference in the amount of missing data [$F(2,132)=1.49$, $p=0.23$]. With an average (SD) of 7.5 (9.9) missing observations for the bipolar group, 6.2 (8.7) for the borderline group and 4.5 (7.1) for the control group). We conducted an additional sensitivity analysis to investigate whether the group differences in types of affective variability might be caused by some difference in the distribution of data missingness. This analysis simply censored a participant's data as soon as that participant misses a day's rating. The results of this analysis are illustrated in Figure S3 panels c and d. As can be seen, while the removal of participants increases the estimated error of the mean (the number of participants remaining in the three groups by day 50 are: Bipolar group, 10; Borderline group, 13; Control group, 19), the same overall pattern of group differences remains apparent. In other words, the group differences in cause of affective variability cannot be attributed to differences in data missingness.

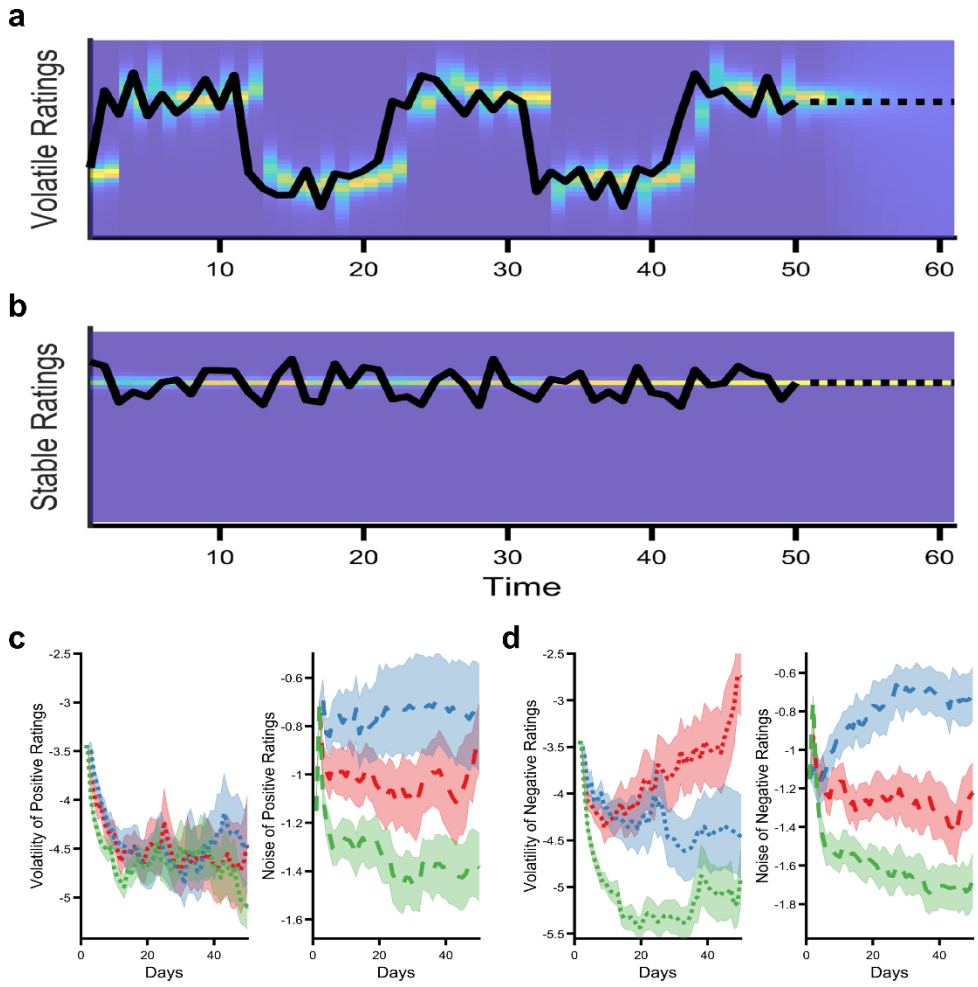

Figure S3: Effect of missing data on the estimated type of affective variability. The filter adapts rationally to missing data, dispersing its belief more quickly about parameters it believes to be changeable e.g. the mu_t parameter when it has been volatile as shown in panel **a** compared to when it is stable as shown in panel **b**. See text for detailed explanation. Solid black lines indicate synthetic data fed to the filter, dotted lines are missing data. The background colour illustrates the filter's marginal belief about mu_t which disperses more quickly in the volatile than stable example. Data missingness does not account for group differences in the cohort study. Panels **c** and **d** show the evolution of the filter's belief about the causes of affective variability as previously shown in Figure 2. In the current figure a participant's data is censored as soon as a single day's data is missing. The same pattern of group differences are apparent indicating that this effect is not related to differences in data missingness. Data from the bipolar group is summarized with red, the borderline group with blue and the control group with green lines. Lines reporting volatility are dotted, those reporting noise are dashed.

Duration of affective responses

The filter attributes change in affect that persist across ratings to volatility and change that do not to noise. This process is illustrated, using synthetic data, in Figure S4 panels a and b. Synthetic timeseries of affect ratings were generated by applying a series of perturbations that caused the affect rating to change, with this change gradually wearing off over time. These perturbations were intended to mimic the effect of events occurring in a patient's life that induced a change in affect. We varied both the frequency with which the perturbations occurred and the rate at which their effect wore off. Characteristic time courses of events at the four extremes (i.e. rare vs. frequent events, brief effect on ratings vs. sustained effect) are illustrated at the appropriate corners of Figure S4 panels a and b. The synthetic data was then passed to the filter which estimated the volatility (panel a) and noise (panel b) of the timeseries. As can be seen, estimates of both volatility and noise increase as the frequency of perturbations increase, whereas the duration of the perturbation influences the attribution of the variability, with long lasting perturbations being attributed to volatility and short lasting perturbations being attributed to noise.

This suggests that patients with bipolar disorder may have a longer duration of affective response than patients with borderline personality disorder (at least for negative affect) and that lithium acts to increase the duration of positive affective responses. In the main study we analysed daily affect ratings and therefore cannot assess timescales shorter than this. However, a significant proportion of participants in the cohort study provided daily ratings for longer than a year, allowing us to assess whether the observed group differences in affective variability are limited to data sampled once a day or whether they are also apparent at longer timescales. The results of these analyses are illustrated in Figure S4 panels c-f. In these analyses the first rating returned every week (panels c-d) or every three weeks (panels e-f) were passed to the filter. As can be seen the group differences in cause of affective variability remains apparent at the longer time periods. These results suggest that the duration of shifts of negative affective in the bipolar group have a duration of at least three weeks. An outstanding question concerns the duration of shifts in the borderline group, assessing this would require collecting affect ratings at a frequency greater than once daily.

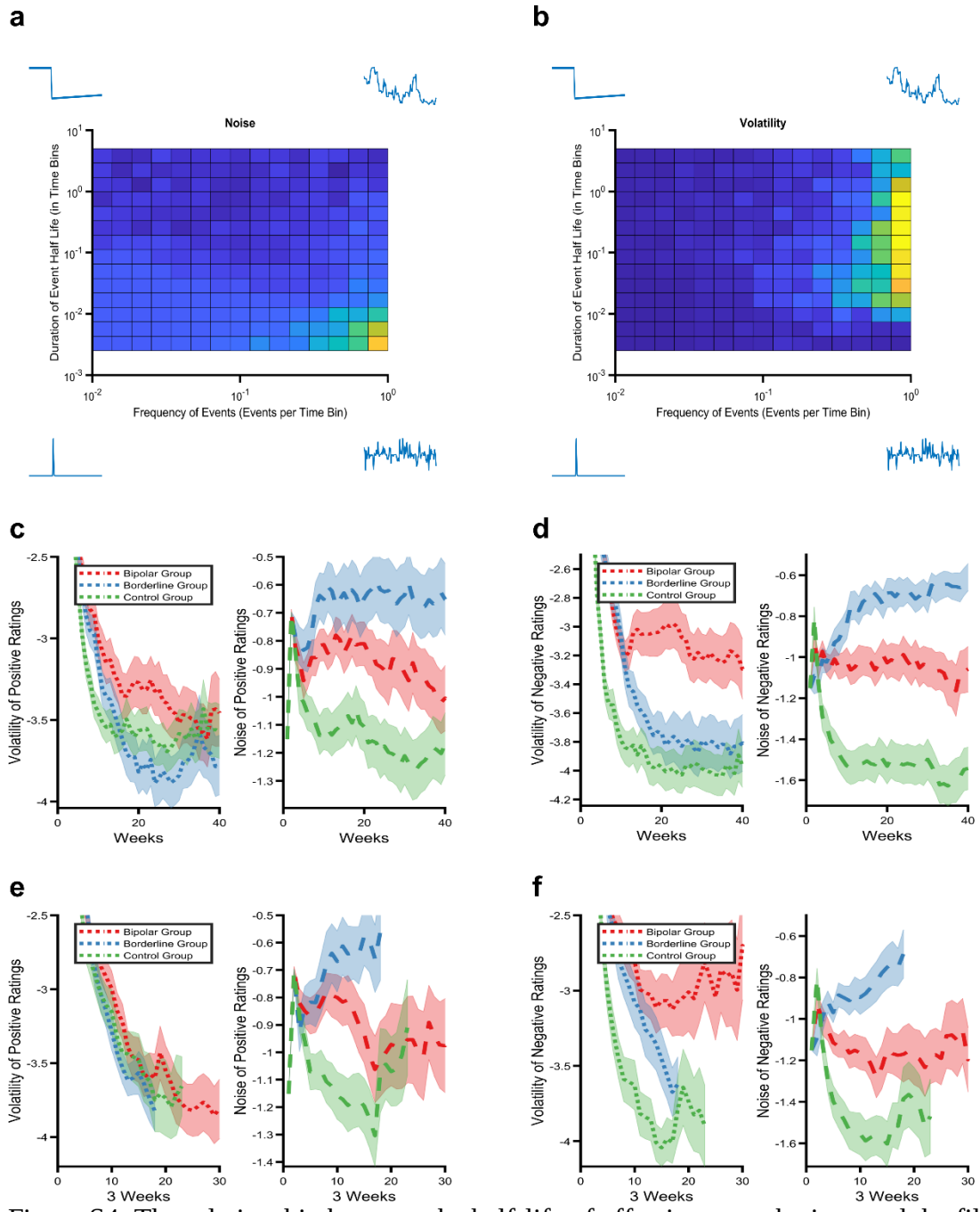

Figure S4: The relationship between the half-life of affective perturbations and the filter derived metrics. Panels **a** and **b** illustrate the effect of the frequency of affective perturbations (x axis) and the duration of their effect (y axis) on estimates of volatility **a** and noise **b** from synthetic data. High levels of volatility and noise are represented by yellow colors and low levels by blue colors. As can be seen, both volatility and noise increase as the frequency of affective perturbations increase (a similar effect on both volatility and noise is seen if the magnitude of affective perturbations increase), whereas the duration of the perturbations influences whether affective variability is attributed to volatility (for longer lasting perturbations) or noise (for shorter perturbations). Illustrative timecourses of the synthetic data are shown in the corner of each panel. Panels **c-f**. Group

differences in the cause of affective variability in the cohort study are apparent when a lower sampling frequency of one (panels **c** and **d**) or three (panels **e** and f) weeks is used. The x-axis reports the number of time points (i.e. for the three week data time point 2 occurs three weeks after time point 1). Mean (SEM) of filter derived estimates are displayed as per the figures in the main paper, lines are censored when fewer than 10 participants are left in a group (i.e. participants remained in the study for variable lengths of time). Data from the bipolar group is summarized with red, the borderline group with blue and the control group with green lines. Lines reporting volatility are dotted, those reporting noise are dashed.

Effect of other medications on affective volatility and noise

In the cohort study of the main paper, receipt of lithium was found to account for the increased volatility of positive affect seen in the bipolar group, but not for the increase in negative volatility. It was not associated with affective noise (which was raised in both patient groups). As the groups differ on a range of other medications received, we ran additional sensitivity analyses in which we added explanatory variables for each class of medication to assess a) whether that class of medication competed with lithium in terms of accounting for positive volatility (i.e. was receipt of the medication significantly associated with positive volatility, or when including it as an explanatory term was lithium no longer significantly associated with positive volatility), b) whether the class of medication could account for group differences in negative volatility (i.e. was receipt of the medication significantly associated with negative volatility, or when including it as an explanatory term was diagnostic group no longer significantly associated with negative volatility) and c) whether the class of medication could account for group differences in either positive or negative noise (i.e. was receipt of the medication significantly associated with positive or negative noise, or when including it as an explanatory term was diagnostic group no longer significantly associated with positive or negative noise). None of the different forms of medication were significantly associated with either positive or negative volatility or noise. Similarly, as summarized in table S4 below, none influenced the effect of lithium on positive volatility or diagnostic group on negative volatility, or on positive or negative noise.

Table S4: Association of lithium treatment and positive volatility, and group membership with negative volatility, when the effects of different classes of medication are added to the analysis.

Medication type	Main Effect of Lithium on Positive Volatility	Main Effect of Group on Negative Volatility	Main Effect of Group on Positive Noise	Main Effect of Group on Negative Noise
Anticonvulsants	$F(1,120)=5.28$, $p=0.02$	$F(2,120)=4.04$, $p=0.02$	$F(2,120)=7.6$, $p=0.001$	$F(2,120)=27.2$, $p<0.001$
Antipsychotics	$F(1,120)=4.97$, $p=0.03$	$F(2,120)=4.4$, $p=0.01$	$F(2,120)=9.7$, $p<0.001$	$F(2,120)=28.1$, $p<0.001$

Antidepressants	F(1,119)=5.71, p=0.02	F(2,119)=5.35, p<0.01	F(2,119)=7.7, p=0.001	F(2,119)=23.9, p<0.001
Anxiolytics	F(1,119)=5.12, p=0.02	F(2,119)=6.6, p<0.01	F(2,119)=8, p=0.001	F(2,119)=27.6, p<0.001
Hypnotics	F(1,118)=5.1, p=0.03	F(2,118)=7.14, p<0.01	F(2,118)=7.5, p=0.001	F(2,118)=29.3, p<0.001

As a final sensitivity analysis, we tested whether, within the group of patients with bipolar disorder in the cohort study, the pattern of affective volatility and noise differed between the groups of patients with Bipolar Disorder type I versus type II. No difference between these groups was found for any interaction including bipolar disorder sub-type (all $p>0.3$).

1. T. E. J. Behrens, M. W. Woolrich, M. E. Walton, M. F. S. Rushworth, Learning the value of information in an uncertain world. *Nat. Neurosci.* **10**, 1214–1221 (2007).

2. T. M. Liddell, J. K. Kruschke, Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology* **79**, 328–348 (2018).

3. J. van den Broek, A Score Test for Zero Inflation in a Poisson Distribution. *Biometrics* **51**, 738–743 (1995).

4. E. Dejonckheere, *et al.*, Complex affect dynamics add limited information to the prediction of psychological well-being. *Nat Hum Behav* **3**, 478–491 (2019).

5. P. Kuppens, Z. Oravecz, F. Tuerlinckx, Feelings change: Accounting for individual differences in the temporal dynamics of affect. *Journal of Personality and Social Psychology* **99**, 1042–1060 (2010).