Supplementary Information

Additional details on linear mixed-effects model

A linear mixed-effects model was fitted to compare the effectiveness of the different treatment methods (see main text). Random intercept terms were included for (i) horse, and (ii) treatment number (i.e. repeated treatments within the study period) nested within horse. The use of a linear mixed-effects model for this type of analysis allows reliable significance tests to be calculated, even in the presence of multiple possibly correlated measurements from the same observational unit (for example patient or horse). By contrast, a standard linear regression, analysis of variance (ANOVA), or analysis of covariance (ANCOVA) model assumes that all measurements are independent, which can lead to incorrect test outcomes when this assumption is not met. The model fitting was performed with the 'nlme' package [1] for the statistical programming language R.

In order to select appropriate terms for inclusion in the final model, a series of candidate models containing various combinations of terms was initially fitted, and their results compared. Terms were considered for (i) treatment method, (ii) previous pocket depth prior to treatment, (iii) days since previous treatment, (iv) interaction effect between treatment method and previous pocket depth, (v) interaction effect between treatment method and days since previous treatment, and (vi) whether additional chlorhexidine based mouthwash ('Hexarinse') was provided to the client for later use. The response variable was the improvement in pocket depth between visits. By pairwise calculation of likelihood ratio tests and comparison of Akaike's information criterion (AIC) between models, it was found that only the main effect terms for (i) treatment method, and (ii) previous pocket depth prior to treatment were significant. These terms were included in the final model (Supplementary Table S1). As discussed in the main text, previous pocket depth prior to treatment was found to be strongly correlated with the choice of treatment method, representing an important confounding effect.

In Supplementary Table S1, the intercept and treatment terms represent the estimated additional effect of each treatment method, after the previous pocket depth has been taken into account. The intercept represents method CD, which is used as a reference category. The other terms (M, PVS and DB) each represent an increment relative to the intercept, i.e. the estimated additional improvement for that method compared to CD. The terms for CD (the intercept), M and PVS are not statistically significant, while the term for DB is statistically significant and negative. This result was further investigated by calculating post-hoc tests comparing all possible pairs of treatment methods (using the 'multcomp' R package [2], which ensures tests are corrected for multiple testing). The post-hoc tests confirmed the result and showed that, after taking into account previous pocket depth, DB was associated with smaller improvements than each of the other treatment groups. No other statistically significant pairwise differences were found among the other groups (CD, M and PVS), after previous pocket depth was taken into account.

However, further analysis also showed that the significance of the term for DB was strongly influenced by a single influential data point (in the PVS group). Due to the small sample size in the DB group (see Table 1 in main text) and the sensitivity to a single data point, the result for DB should not be interpreted as strong evidence.

Residual analysis showed no serious problems in the model. To investigate the robustness of the results, a linear regression / analysis of covariance (ANCOVA) model was also fitted to the data, ignoring possible correlations among measurements from each horse. The results were similar, except that terms for days since previous treatment and provision of additional chlorhexidine based mouthwash ('Hexarinse') were also statistically significant, and were therefore kept in the model. After taking into account previous pocket depth, DB was associated with significantly smaller improvements than CD, as in the previous model. No other pairwise comparisons between treatment methods were significant (including DB compared to M or PVS). However, comparison of AIC values confirmed that this model was inferior to the linear mixed-effects model in Supplementary Table S1 (AIC = 1429.6, compared to AIC = 1378.0 in Supplementary Table S1).

References

1. Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. and R Core Team. (2014) *nlme: Linear and Nonlinear Mixed Effects Models*. R package, version 3.1-118.

2. Hothorn, T., Bretz, F. and Westfall, P. (2008) Simultaneous Inference in General Parametric Models. *Biometrical Journal*. **50**(3), 346-363.

Supplementary Table S1. Results of linear mixed-effects model fitted to compare the effectiveness of the different treatment methods. Random intercept terms were included for horse and treatment number nested within horse (see text). The intercept represents treatment method CD, and the other treatment terms each represent increments relative to the intercept. CD = cleaning and disinfecting; M = additional use of metronidazole antibiotic tablets; PVS = additional use of polyvinyl siloxane temporary filling; DB = additional use of diastema burring; CI = confidence interval; AIC = Akaike's information criterion.

	Estimate (Std. error)	Approx. 95% CI	Approx. p-value
Intercept	-0.561 (0.509)	[-1.556, 0.433]	0.3
treatmentM	0.140 (0.454)	[-0.748, 1.027]	0.8
treatmentPVS	-0.368 (0.594)	[-1.528, 0.791]	0.5
treatmentDB	-3.774 (1.099)	[-5.921, -1.627]	<0.001
previous_depth	0.507 (0.056)	[0.397, 0.617]	<0.001

Model criterion: AIC = 1378.0