Appendix A. Scalar extraction vs. scalar field statistics

The purpose of this Appendix is to demonstrate how scalar extraction can bias non-directed hypothesis testing. To this end we developed and analyzed an arbitrary dataset (Fig.S1). We caution readers that we have constructed these data specifically to demonstrate particular concepts. The reader is therefore left to judge the relevance of this discussion to real (experimental) datasets.

The specific goal of this Appendix is to scrutinize the similarities and differences between: (a) a typical univariate two-sample t test, and (b) a scalar field two-sample t test.

Consider the simulated scalar field dataset in Fig.S1. In Fig.S1a, arbitrary true mean fields are defined for two experimental conditions: "Cond A" and "Cond B". The Cond B mean was produced using a half sine cycle. The Cond A mean was produced by adding a small Gaussian pulse (at time= 85%) to the Cond B mean. This Gaussian pulse is evident in the true mean field difference (Fig.S1b).

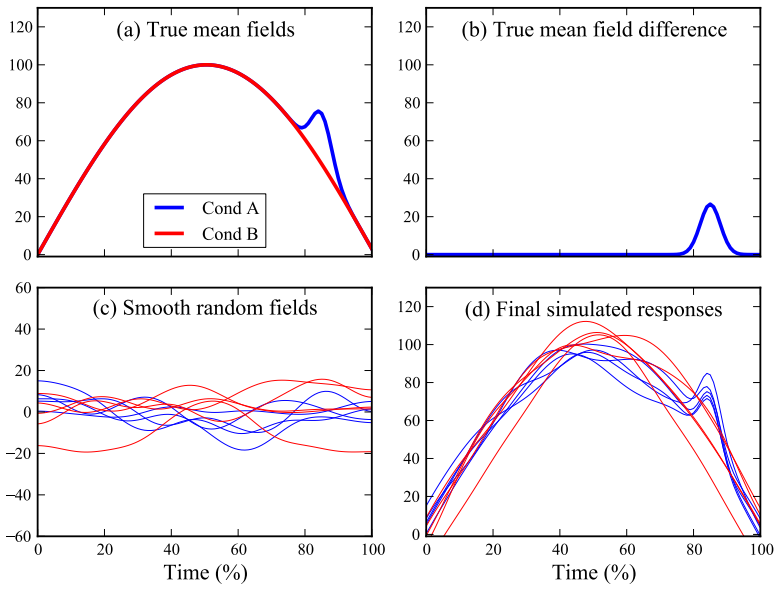

Figure S1: Simulated scalar field dataset depicting two experimental conditions: "Cond A" and "Cond B" (arbitrary units).

We next simulate smooth random fields: five for each condition (Fig.S1c). These random fields were constructed by generating ten fields, each containing 100 random, uncorrelated and normally distributed numbers, then smoothing them using a Gaussian kernel. Adding the random fields to the true field means (Fig.S1a) produced the final simulated responses (Fig.S1d). For interpretive convenience, let us assume that these data represent joint flexion.

Imagine next that we wish to test the following (non-directed) null hypothesis: "Cond A and Cond B yield identical kinematics". Consider first scalar extraction: after observing the data (Fig.S1d) one might decide to extract and analyze the maximum flexion, which occurs near time = 50%:

$$y_{\mathrm{A}} = \left[\begin{array}{ccccc} 100.0 & 91.2 & 92.2 & 95.5 & 97.1 \end{array}\right]$$

$$y_{\mathrm{B}} = \left[\begin{array}{ccccc} 97.2 & 101.9 & 104.8 & 106.3 & 111.7 \end{array}\right]$$

A two-sample t test on these data yields: t=3.16, p=0.013. We would reject the null hypothesis at α=0.05, and we would conclude that Cond B produces significantly greater maximal flexion than Cond A.

An alternative is to use Statistical Parametric Mapping (SPM) (Fig.S2). The SPM procedures are conceptually identical to univariate procedures (Table S1). The only apparent difference is that SPM uses a different probability distribution (Steps 4 and 5). This probability distribution is actually not different because it reduces to the univariate distribution when Q=1 (i.e. if there is only one time point).

SPM results find significant differences between the two conditions near time = 85% (Fig.S2d). We would therefore reject our null hypothesis, with the caveat that significant differences were only found near time = 85%.

Although univariate t testing and SPM t testing are conceptually identical, they have yielded (effectively) opposite results. The univariate test found significantly greater maximal flexion in Cond B, but SPM found significantly greater flexion in Cond A (near time=85%).

Table S1: Comparison of computational steps for univariate and SPM two-sample t tests ("st.dev." = standard deviation).

Step	**(a) Univariate two-sample t test**	**(b) SPM two-sample t test**	Figure
1	Compute mean values $\overline{y}_{\mathrm{A}}$ and $\overline{y}_{\mathrm{B}}$.	Compute mean fields $\overline{y}_{\mathrm{A}}(q)$ and $\overline{y}_{\mathrm{B}}(q)$	S2(b)
2	Compute st.dev. values s_{A} and s_{B}.	Compute st.dev. fields $s_{\mathrm{A}}(q)$ and $s_{\mathrm{B}}(q)$	S2(b)
3	Compute the t test statistic: $$t = \frac{\overline{y}_{\mathrm{B}} - \overline{y}_{\mathrm{A}}}{\sqrt{\frac{1}{J}(s_{\mathrm{A}}^2 + s_{\mathrm{B}}^2)}}$$	Compute the t test statistic field: $$\mathrm{SPM}\{t\} \equiv t(q) = \frac{\overline{y}_{\mathrm{B}}(q) - \overline{y}_{\mathrm{A}}(q)}{\sqrt{\frac{1}{J}\left(s_{\mathrm{A}}^2(q) + s_{\mathrm{B}}^2(q)\right)}}$$	S2(c)
4	Conduct statistical inference. First use α and the univariate t distribution to compute t_{critical}. If $t > t_{\mathrm{critical}}$, then reject null hypothesis.	Conduct statistical inference. First use α and the random field theory t distribution to compute t_{critical}. If SPM$\{t\}$ exceeds t_{critical}, then reject null hypothesis for the suprathreshold region(s).	S2(d)
5	Compute exact p value using t and the univariate t distribution.	Compute exact p value(s) for each suprathreshold cluster using cluster size and random field theory distribution(s) for SPM$\{t\}$ topology.	S2(d)

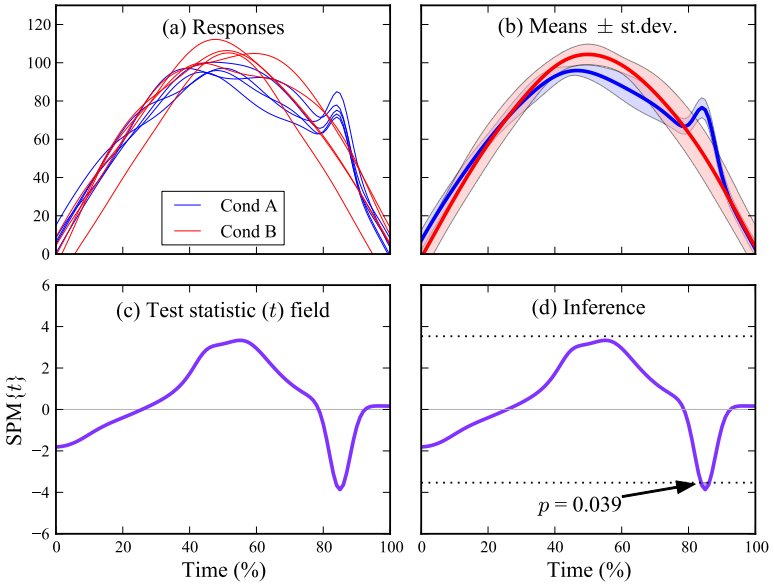

Figure S2: Scalar field analysis using Statistical Parametric Mapping (SPM). In panel (d) the thin dotted lines depict the critical random field theory threshold of $|t_{\mathrm{critical}}|$=3.533. The (incorrect) Šidák threshold is $|t_{\mathrm{critical}}|$=5.595.

This discrepancy can be resolved through standard probability theory regarding multiple comparisons, through a consideration of 'corrected' and 'uncorrected' thresholds. First consider conducting one statistical test at α=0.05. The choice: "α=0.05" means that we are accepting a 5% chance of incorrectly rejecting the null hypothesis, or, equivalently, a 5% chance of a 'false positive'. If we conduct more than one test, there is a greater-than 5% chance of a false positive. Specifically, if we conduct N statistical tests, the probability of at least one false positive is given by the family-wise error rate $\overline{\alpha}$:

$$\overline{\alpha} = 1 - (1 - \alpha)^N$$

For N=2 tests, there is a $\overline{\alpha}$=9.75% chance that at least one test will produce a false positive. For N=100 tests, $\overline{\alpha}$=99.4%.

To protect against false positives, and to maintain a constant family-wise error rate of $\overline{\alpha}$=0.05, we must adopt a corrected threshold. One option is the Šidák threshold:

$$p_{\text{critical}} = 1 - (1 - \overline{\alpha})^{1/N}$$

For N=2 and N=100 tests, the Šidák thresholds are p_{critical}=0.0253 and p_{critical}=0.000513, respectively.

Herein lies one problem: our scalar extraction analysis has used an uncorrected threshold of p_{critical}=0.05. Even though we have formally conducted only one statistical test, the data were extracted from a dataset that is 100 times as large. Since we observed the data before choosing which scalar to extract, we effectively conducted N=100 tests, albeit visually, then chose to focus on only one test. By failing to adopt a corrected threshold, we have biased our analyses.

Although the Šidák correction helps to avoid false positives, it is not generally a good choice because it assumes that there are 100 independent tests (i.e. one for each time point in our dataset). The points in this dataset are clearly not independent because the curves are smooth, changing only gradually over time. Thus the Šidák correction is too severe, lowering $\overline{\alpha}$ well

below 0.05. An overly severe threshold produces the opposite bias: an increased chance of false negatives.

SPM employs a random field theory (RFT) correction to more accurately maintain $\overline{\alpha}$=0.05. The precise threshold is based not only on field size (Q=100), but also on field smoothness — which is estimated from temporal derivatives. Computational details for RFT corrections are provided in the SPM literature.

Unfortunately, even if our scalar analysis had employed a corrected threshold, it still would have been biased, but for a separate reason. By focussing only on maximal flexion (which did not appear in our null hypothesis), we have neglected to consider the signal at time = 85%, and have therefore not detected the true field difference (Fig.S1a). In contrast, SPM was able to uncover the true signal because it both adopted a corrected threshold and considered the entire field simultaneously (Fig.S1d).

The aforementioned sources of bias — (1) failing to adopt a corrected threshold, and (2) failing to consider the entire field — are referred to collectively in the main manuscript as 'regional focus bias'.

Last, we reiterate that this Appendix is relevant only to non-directed hypotheses. If we had formulated a (directed) hypothesis regarding only maximal flexion — prior to observing the data — then our scalar extraction analyses would not have been biased because our null hypothesis would not have pertained to the entire time domain 0–100%.

In summary, regional focus bias can be avoided by:

1. Specifying a directed null hypothesis — before observing the data — and then extracting only those scalars which are specified in the null hypothesis.

2. Analyzing the data using SPM or another field technique which both considers the entire temporal domain and which adopts a corrected threshold.