Available online at www.sciencedirect.com

SCIENCE DIRECT®

Experimental Gerontology 40 (2005) 101–113

Experimental Gerontology

www.elsevier.com/locate/expgero

Re-evaluating a test of the heterogeneity explanation for mortality plateaus☆

David Steinsaltz*

Department of Demography, University of California, 2232 Piedmont Ave., Berkeley, CA 94720-2120, USA

Received 23 April 2004; received in revised form 24 November 2004; accepted 24 November 2004
Available online 25 December 2004

Abstract

[Drapeau, M.D., Gass, E.K., Simison, M.D., Mueller, L.D., Rose, M.R., 2000. Testing the heterogeneity theory of late-life mortality plateaus by using cohorts of *Drosophila melanogaster*, Experimental Gerontology, 35 71–84.] tested, in populations of *Drosophila melanogaster*, a prediction of the heterogeneity explanation for mortality plateaus. They concluded that heterogeneity could not explain their results. We contend here that the statistical analysis was flawed. It was declared that there was no difference between the mortality plateaus of three different strains, on the basis of averaged outcomes. In fact, the results for the different strains were quite different. Most trials showed the expected lowering of the mortality plateaus for the flies selected for robustness, but these effects were washed out by a small number of very large opposing deviations. There is ample reason to believe that the opposing deviations are artifacts of fitting an overly restrictive hazard-rate model. When we fit more appropriate models, the evidence points toward a rejection of the null hypothesis (of identical plateaus), hence toward modest support for the heterogeneity explanation.

Keywords: Heterogeneity explanation; Mortality plateaus; *Drosophila melanogaster*

1. Introduction

The analysis of experimental lifespan data is fraught with statistical difficulties. The difficulties are multiplied when we examine mortality rates for the 'oldest old', when sample sizes have dwindled. Wang et al. (1998) have argued, for example, that inappropriate parametric estimators, combined with an inadequate treatment of data-censoring, led Brooks et al. (1994) to favor an unwarranted fixed-frailty interpretation of mortality deceleration in nematodes. (For an account of general biodemographic issues related to mortality deceleration, see (Vaupel et al., 1998; Pletcher and Curtsinger, 1998)).

Drapeau et al. (2000) present the results of an experiment on *Drosophila melanogaster*, which, they claim, are inconsistent with the heterogeneity theory of mortality plateaus. A more nuanced analysis of the data, we contend, points in the opposite direction. Not only does this force us to reevaluate the conclusions of this experiment, it may serve as a paradigm for some pitfalls in survival-data analysis.

The authors aim to test the prediction that 'populations that are greatly differentiated for stress resistance should show great differences in their late-life mortality schedules,' without stating explicitly what those differences should be. There is, in any case, no single consensus 'heterogeneity model'—some mathematical treatments of heterogeneity models may be found in Vaupel et al. (1979, 1998), Vaupel and Carey (1993) and Service (2000). Since heterogeneity produces its plateau gradually and indirectly—and transiently, if the variation is only in the initial mortality, and is bounded—the definition of the plateau will inevitably be ambiguous. Service et al. (2000), in a critique of this same work, has argued that reasonable versions of the heterogeneity model could produce plateaus that are fairly insensitive to selection. In a more extensive work Service (2000), the same author has shown that a population with Gaussian-form heterogeneity should see the plateau levels *rise* under

☆ Supported by NIA Grant K12-AG00981-02:449470.

* Tel.: +1 510 643 0246; fax: +1 510 643 8558.
E-mail address: dstein@demog.berkeley.edu.

0531-5565/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.exger.2004.11.010

selection for increased robustness. On the other hand, consider a population composed of just two strains, each with mortality rate $k\,e^{0.05x}$, with one having $k=10^{-4}$, the other $k=10^{-6}$. The heterogeneity of k will produce a transient plateau, and it is easy to see that increasing the proportion of robust flies will lower the plateau. In other words, any change in plateau level, or no change at all, is still consistent with the heterogeneity explanation.

Even on the standard set by Drapeau et al. though, the heterogeneity theory acquits itself well. Re-analyzing the data, we find clear evidence for plateaus that are lower for the robust strain. We contend that the authors of the original study applied an inappropriate statistical model, one which is badly misspecified for the observations, and would consequently—as we show by applying it to simulated data in Section 4—be incapable of confirming even quite large and unquestionable differences in plateau level. The plateaus are also more shallow in the selected population, consistent with plateaus produced by population heterogeneity. The plateau timing, on the other hand, seems not to have been altered. As a test of heterogeneity, this must be seen as quite indirect, aside from the general principle that a positive result is less probitive than a negative. At the same time, it should influence the discussion of how, and whether, interventions may affect mortality plateaus. The weight of the experimental data shifts from strong evidence against heterogeneity to weak evidence in favor.

The experimenters followed three different strains of *D. melanogaster*: the SO strain, maintained from an original wild type progenitor after many generations of selection for starvation resistance; the CO strain, which had been subject to no special selection; and the RSO strain, which had been derived from the SO strain, but released from selection for the past 25 generations. (A more extensive description of these strains, their history, and the selection regime may be found in Mueller et al. (2003).) Five replicates of each population (males and females separated: thus, 30 populations in all) were followed until all had died.

The main conclusion of the paper is derived from Tables 1 and 2. (The boxes have been added, for emphasis.) These numbers are based on the computation of a maximum-likelihood estimator of the hazard rate, chosen from the class of two-part functions which begin gompertzian (exponentially increasing), and then become constant for all times after a 'breakday'. The paper claims:

Table 1
Breakday of mortality plateau

	Male			Female		
Replicate	CO	RSO	SO	CO	RSO	SO
1	42	$\boxed{76}$	44	50	$\boxed{78}$	46
2	42	44	44	50	$\boxed{66}$	52
3	46	68	70	46	52	$\boxed{72}$
4	44	44	44	50	52	46
5	44	44	$\boxed{72}$	52	52	46

Table 2
Level of mortality plateau

	Male			Female		
Replicate	CO	RSO	SO	CO	RSO	SO
1	0.10	$\boxed{0.21}$	0.07	0.10	$\boxed{0.14}$	0.07
2	0.10	0.07	0.05	0.13	$\boxed{0.15}$	0.08
3	0.13	0.11	0.10	0.12	0.07	$\boxed{0.16}$
4	0.12	0.08	0.07	0.15	0.10	0.09
5	0.10	0.08	$\boxed{0.19}$	0.12	0.09	0.07

> The data presented here do not indicate any clear relationship between late-life mortality rates and large genetic differences in stress resistance. In this respect, our results provide a refutation of the heterogeneity theory of late-life mortality.

A perusal of the data suggests, though, that the plateau levels for the three groups *are*, in fact, significantly different. It is true that the *average* plateau levels for the three strains are indistinguishable. Observe, however, that the values for the SO flies are strongly bimodal. In most of the trials—four out of five for each sex—the SO flies had the lowest plateau level of the three groups. This is balanced out in the mean by two exceptionally high values, male replicate 5 and female replicate 3. Observe, too, that these two trials (as well as three exceptionally high plateau levels for the RSO strains) also correspond to outliers of the breakday, values between 66 and 78. Twenty-three of the 30 breakdays, including all breakdays for the CO strain, are between 42 and 52, while the rest are at least 66. It requires no advanced statistical methodology to recognize the potential for misdirection when statistical tests on means are applied to such bimodal data.

2. Bias of the plateau estimator

What is the origin of these very high plateau levels, and why do they correspond to late breakdays? Consider, for example, the empirical hazard rates for the first CO and RSO male replicate, as shown in Fig. 1. It is hard to see from these plots why the RSO strain—which is below the CO nearly everywhere—should be assigned a substantially higher mortality plateau level. We suggest that this discrepancy is an artifact of the statistical estimation procedure.

Plotting the empirical hazard rates for different trials, we see that the parametric model chosen by Drapeau et al.—Gompertz up to some point, constant thereafter—does not

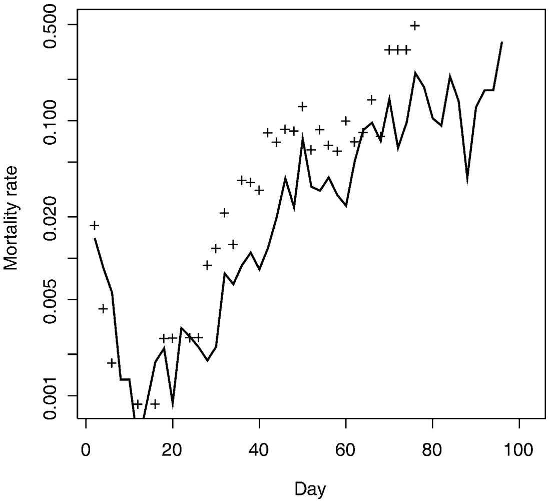

Fig. 1. Empirical hazard rates for first replicate of CO and RSO male flies. The lines are for the RSO flies, while the '+' symbols represent the CO flies.

fit terribly well. After the 'breakday', mortality rates continue to increase, albeit more slowly. The maximum-likelihood algorithm is thus in the position of approximating the latter part of an increasing function by a constant. If this tail becomes longer, the algorithm seeking a flat plateau would be expected to strike later, and thus (since mortality rates are still increasing), higher. More to the point, the length of the tail, depending as it does on the extended survival of a very few flies, will be particularly prey to random fluctuations. Since the RSO and SO strains have lower late-life mortalities—for example, on day 70, when more than one-sixth of SO males are still alive, over 98% of CO males have gone the way of all flesh—there will be a bias toward higher estimated plateau levels.

We sketch this situation in Fig. 2. The black curve portrays a hypothetical empirical log-hazard-rate curve, which might be observed if the true underlying mortality rate were exponentially increasing at one rate up to a fixed age, and then switched to a lower exponential rate. The red curve shows the same mortality rates, reduced by a constant factor. One might reasonably attribute to each mortality curve a 'plateau', with the plateau for the black population being simultaneous with, but higher than the plateau for the red.

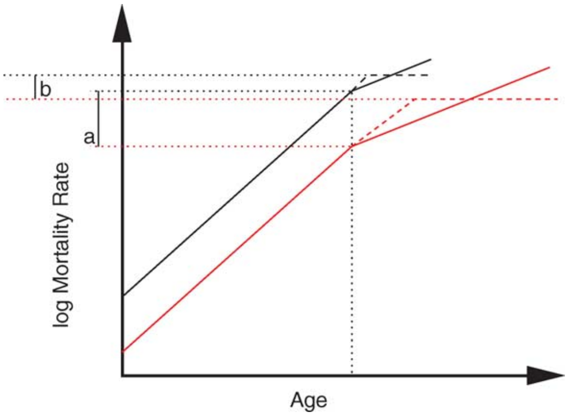

Fig. 2. Illustrating the potential bias in the flat-plateau estimate. The black curve represents a hypothetical mortality rate, the red curve the same rate decreased by a constant factor. The dashed lines show the estimated three-parameter fit, and the dotted lines show the real and estimated plateau levels. The discrepancy 'a' is the true plateau difference, while 'b' marks the (smaller) difference between the plateau estimates.

We have drawn dashed lines to approximate fitting a flat-plateau curve to these data. Because the mortality levels are generally reduced in the red curve, more flies remain alive, and the post-plateau phase continues longer. The experience of early days, when more flies remain alive, weighs heavy on a maximum-likelihood estimator, while the time after the final demise has no weight at all. Consequently, the red estimator, representing a population with more late survivors, is more strongly influenced by late mortality than the black estimator: its plateau will strike later, hence also relatively high. Instead of the true difference in plateau levels 'a', we estimate the smaller difference 'b'.

It is not merely that an increasing-plateau curve may provide a better fit. More important, we show that the flat-plateau estimator is biased when the mortality rates do not really flatten out, so as to compress the plateau levels, and conceal those differences that are present. (This effect may be exacerbated by the substantial differences in starting numbers between different strains.) In addition, the red-curve plateau estimates, depending as they do on a longer period of stochastic survival of small numbers of individuals, may be expected to be more unstable and variable than the black-curve estimates.

This objection has some similarities to the forceful criticism of the same paper made by de Grey (2003b). de Grey rejects maximum likelihood estimation altogether, along with any procedure that gives more weight to larger numbers of deaths, thus exaggerating the influence of early mortality. He suggests that maximum-likelihood procedures with misspecified models is always wrong-headed. He prefers to minimize the sum of absolute differences to the total survival, though he fails to offer, either here or in (de Grey, 2003a), evidence for believing that his procedure would do better. Applying the alternative procedure to the Drapeau et al. data, he concludes that the excellence of fit is 'readily seen' in a plot of total survival, ignoring the fact that total-survival plots are notoriously coarse tools for recognizing differences in mortality rates.

de Grey dismisses the advantages of MLE as being only a mathematical 'seduction', appropriate only in ideal cases, when the true mortality curve is represented exactly by the model. While this argument is consonant with our own—like any estimation procedure, the effectiveness of MLE fitting depends on the 'true' result being close to the class of available models—it exaggerates the details of the estimation procedure, and ignores the particular virtues of MLE in compensating for random fluctuations in death rates while the at-risk population is shrinking. (de Grey's other applications of his method, in de Grey (2003a), were to populations large enough that random fluctuations in death numbers through most of the life course would be negligible.) Furthermore, MLE interpolates weightings smoothly and naturally as the population dies out. de Grey's approach, by contrast, assigns equal weight to each day's mortality, until the population has died out completely. At this point, the weight (unavoidably) plummets to 0. This singularly fails to address a major source of bias, which is the longer total lifetime—hence a more extended plateau, slowly rising—for the low-mortality strains. Adding to this de Grey's arbitrary choice to minimize the sums of first powers, rather than the infinity of other possibilities, we see considerable cause for skepticism of de Grey's approach.

Our alternative, described in Section 3, is to maintain the fitting procedure in its essence, but to improve the class of models, making it less badly misspecified. We propose two alternative models. In one, we alter the crucial plateau region, allowing the 'plateau' to increase, albeit at a slower rate. This model has the same number of parameters as the original (four), because we do not allow the mortality to jump at the breakday, but maintain a continuous mortality curve. It seems to us more sensible to have the plateau mortality start at the same point where the early-life mortality leaves off. Our second alternative model, also continuous, and also with four free parameters, is a version of the popular logistic Gompertz mortality curve.

We demonstrate the success of this approach in several ways. Both alternative models drastically reduce the spread, and eliminate the bimodality, of estimates for different replicates of the same strain. We also show, by simulations, in Section 4, that wild fluctuations and biases, like those observed in Drapeau et al. would indeed appear when mortality curves with rising plateaus are pressed into a flat-plateau mold.

The problem of maximum-likelihood model-fitting has also been analyzed at length, through simulation studies similar to those presented here, by Pletcher (1999). While the specific questions considered were different—Pletcher considered, on the one hand, the ability of statistical tests to distinguish a logistic mortality curve from an unbent Gompertz mortality, and, on the other hand, the ability to discern the differences in Gompertz mortality between two different strains—some of the conclusions are relevant to our study as well. In particular, he considered the effect of fitting a Gompertz mortality curve to data which actually followed the Gompertz-Makeham pattern (with an added age-independent mortality term) or logistic, finding that the misspecified mortality model 'results in parameter estimates that are highly biased.'

3. Alternative models

We demonstrate the distortions of the flat-plateau model in two ways. To begin with, we show that an alternative model of plateau behavior provides more stable results, and reveals clear differences between the populations. It is not, we reiterate, a matter of finding a 'better' fit. Rather, we show that our models avoid both the wide divergences in breakdays, and the odd disparity in variability of plateau levels among the populations. There are still far more data points than parameters, so there is no danger of overfitting. It would seem that a claim of 'no difference' between populations must be abandoned if a reasonable alternative model does uncover such a difference. (This assumes, of course, that one has not tested and abandoned many alternative models.)

The 'flat-plateau' model fitted in the original paper was

$$h_{\mathrm{F}}(t) = \begin{cases} k\,\mathrm{e}^{\alpha t} & \text{for } t \le T, \text{ and} \\ k' & \text{for } \ t > T \end{cases}. \tag{FP}$$

Our alternative models are

$$h_{\mathrm{P}}(t) = \begin{cases} k\,\mathrm{e}^{\alpha_1 t} & \text{for } \ 18 < t \le T, \\ k\,\mathrm{e}^{\alpha_1 T + \alpha_2 (t-T)} & \text{for } T < t; \end{cases} \tag{IP}$$

and

$$h_{\mathrm{L}}(t) = k\frac{1 + A\,\mathrm{e}^{\alpha t}}{1 + B\,\mathrm{e}^{\alpha t}}, \tag{LG}$$

In the first alternative model, which we call the 'increasing plateau' (IP), the hazard is growing exponentially at rate α_1 up to the breakday T, and after that at rate α_2. Even this does not suit the general shape of the hazard-rate curve. The declining mortality early on, which is evident in Fig. 1, is typical of nearly all the replicates. This has serious consequences when we try to fit a model like (IP). The line which fits the early life will be an egregiously poor fit: a sharp decrease followed by a much longer sharp increase is represented by a gradual increase. In many cases, this muddled compromise slope will be less than the slope in late life. That is, instead of a mortality deceleration, we appear to have mortality acceleration.

What we are interested in, when we consider the question of 'mortality plateaus', is the change in slope of the log-mortality curve from midlife to late life. In nearly all the replicates, the decrease in the slope is obvious, but would be masked by mixing the midlife increase with the early-life decrease. One solution would be to fit a three-part curve: a decreasing piece at the start, followed by two increasing pieces. One must ask, though, what function would the early piece serve? It is an irrelevant distraction from the behavior that we seek to analyze. The alternative, which we have preferred, is simply to drop the early piece, and analyze the mid- and late-life mortality data. We have chosen to analyze the data starting after day 18.

For some, this procedure may ring some alarm bells. It is close to a credo of honest statistical analysis that we need to accommodate 'all the data', that to select favorite data points is an opening to manipulation and deception. Recall that this study is not intended as a test of the Gompertz-with-breakpoint model, but rather to use the model as a tool for teasing out the differences in plateau behavior. Our point is that the original statistical test effectively lost power by shoehorning data into an inappropriate model; we boost the discriminating power in the test by applying a more suitable model to only the relevant portion of the data. This is not ‘data snooping’; we are simply electing to analyze only the mid- and late-life data, in all replicates equally, where it is apparent that the chosen models (both ours and that of the original paper) do not suit the early mortality data. This is consistent with de Grey’s warnings (discussed above) of the errors which could be generated by allowing early-life mortality to dominate the analysis of late-life trends. The success of our alternative may be seen in its eliminating the wild fluctuations in the estimated parameters from one replicate to the other, within the same strain. (We note, as well, that a recent analysis of Drosophila survival data by Miyo and Charlesworth (Miyo) found the same difficulty with decreasing early-life mortality, and they also chose to manage it by truncation.)

The second model is the four-parameter logistic Gompertz (LG) model. This has been proposed frequently (Vaupel et al., 1979; Horiuchi and Wilmoth, 1998) as a model, though usually with one fewer parameter, taking our A to be 1. Beginning as it does with a period of mortality acceleration, this model fits the early data without unduly compromising the late-life fit. The clearly defined plateau level is surely an advantage as well, when the plateau levels are what we wish to compare. We define the plateau level in (IP) to be the level when the exponential rate of increase changes from α_1 to α_2 (the parameter T), essentially because this is the only level which is clearly linked to the plateau behavior. There remains a certain degree of arbitrariness to this choice, so that it is valuable to compare these results to those from the (LG) model.

We note that all the models have the same number of parameters. In addition, the alternative models have the advantage of being continuous, as opposed to the flat-plateau model of Drapeau et al. which has a jump at the breakday. We show an example of fitting all three models to the experimental results for the first replicate of male CO flies in Fig. 3.

In each case, we applied the optima obtained by the default algorithm of the *optim* function in the R statistical language, an implementation of the Nelder–Mead algorithm (Nelder and Mead, 1965). For each model, we have set four different starting points, and selected the best

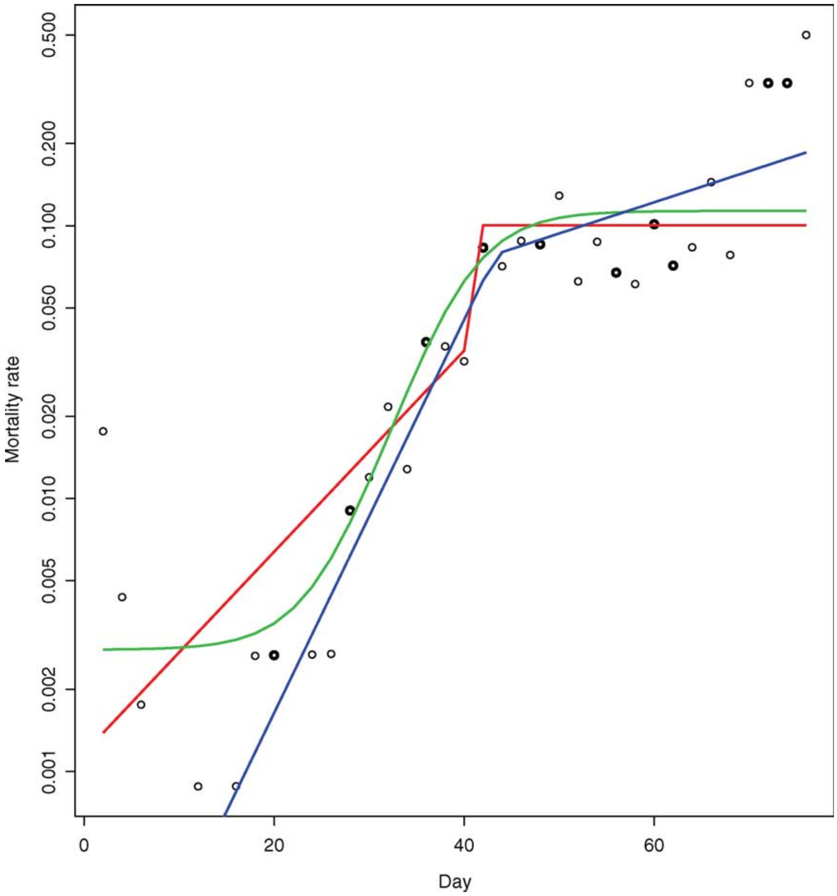

Fig. 3. Fitting three different models to the mortality data from the first replicate of male CO flies. The red curve is the FP fit, the blue is IP, and the green is LG. The circles show the empirical hazard rate.

fit (highest likelihood) for each replicate. We then double-checked these results by comparing them to those obtained from a quasi-Newton optimization algorithm, which is realized in R by the *nlm* function. In each case, the same result was found. We have also checked our results for the 30 datasets with an exhaustive grid search, on a grid of about 1.5 million points. (The grid was fine enough to match all four parameters with an error of about 10%.) In no case did the grid search turn up a better fit—that is, a higher likelihood—than the standard algorithms.

We point out here that the picture looks very different when we apply the two algorithms to fit the (FP) model. For many of the 30 data sets the two algorithms found substantially different curves, though the log likelihoods at these local optima differed only slightly. This is another defect of the (FP) model, which is related to its being seriously misspecified for these data. In Section 4, when we need to fit the (FP) model to simulated data, we apply both algorithms with all four starting points to each simulation run, in order to give ourselves the best chance of finding the true maximum-likelihood estimates.

3.1. The increasing-plateau model

Fitting the data to the (IP) model, we obtained the breakdays given in Table 3. (We have set $T=2T_O+19$, where T_O was the parameter delivered by the MLE computation. The choice of 19 here is slightly arbitrary, since the data points come only every 2 days, but differences between estimates should be meaningful.) Note that these are more stable than the original estimates, and there are no extreme outliers in these estimates. (We must mention, though, that we could have found a slightly better fit for replicate 3 of the SO females, if we had allowed decreasing plateaus.) If we call the mortality at day T the plateau level, we obtain the levels in Table 4. Basic summary statistics for each replicate are included at the bottom. (Here, and elsewhere, we use the standard deviation with $\sqrt{n-1}$ in the denominator.)

It is not immediately obvious how these numbers should be compared. What is an appropriate statistical test, given that we have a small number of measurements. and little idea of what 'error' distribution ought to be supposed? Which of the parameters ought to be given most weight?

Table 3
Breakday of mortality plateau in the piecewise linear model (IP)

Replicate	Male			Female		
	CO	RSO	SO	CO	RSO	SO
1	42.3	46.7	45.0	50.0	50.5	49.3
2	45.0	44.9	44.1	45.6	42.5	46.2
3	45.9	45.9	44.5	46.4	51.1	52.4
4	45.1	46.1	44.8	40.5	49.6	46.3
5	44.1	43.4	44.7	45.7	51.6	49.6

Table 4
Plateau mortality levels in the piecewise log-linear model (IP)

Replicate	Male			Female		
	CO	RSO	SO	CO	RSO	SO
1	0.0786	0.0306	0.0454	0.118	0.0354	0.0541
2	0.109	0.0515	0.0329	0.0755	0.0428	0.0519
3	0.124	0.0301	0.0221	0.101	0.0540	0.0544
4	0.115	0.0602	0.0547	0.0660	0.0767	0.0749
5	0.0919	0.0577	0.0336	0.0594	0.0733	0.0545
Mean	0.104	0.0460	0.0377	0.0840	0.0564	0.0580
SD	0.0185	0.0147	0.0126	0.0248	0.0182	0.00951

Fortunately, we find that various approaches tell essentially the same story: among the males, there is a significant difference between the plateau behavior of the selected strains (SO and RSO) from that of the unselected (CO), while the difference between SO and RSO is not statistically significant.

We begin by noting that, in contrast to the estimates of Drapeau et al. we obtain roughly the same breakdays for all strains. On the other hand, the plateau levels for the CO strains are higher than those for either of the selected strains (RSO and SO), while the two selected strains have roughly the same plateau levels. For the males, the difference is statistically highly significant; for the females, the difference is not so pronounced. For example, if we follow Drapeau et al. in performing an ANOVA test on the three male strains, testing for equality of means, we find an F-statistic of 27.2 for the males, leading to a rejection of the null hypothesis with a p-value of 3.4×10^{-5}. Applying the same test to the female strains yields an F-statistic of 3.48, with a corresponding p-value of 0.064. If we test the strains pairwise, using Welch's two-sample t-test, we find that equality of means between the male CO and RSO is rejected, with a p-value of 0.0007; between the CO and SO the p-value is 0.0003. For the females, the corresponding p-values are 0.08 and 0.078.

Of course, we have no reason to think that these replicates are anything like draws from a normal distribution. If we apply the above tests to the logarithms of the plateau levels, the results are qualitatively the same. More reasonably, we may apply a nonparametric test. We note that, for the males, the lowest plateau level for the CO strains is higher than the highest of either the RSO or SO strains. If we apply the standard Wilcoxon rank-sum test, we obtain, of course, the maximum W-statistic of 25, telling us to reject equality of means between CO and either the RSO or SO at a p-value of 0.008. For the females, the same test yields a p-value of 0.03 for equality of the CO and SO means (and 0.15 for equality for CO and RSO). Given the multiple testing involved, the former should probably not be treated as statistically significant.

As already noted, the plateau levels in this model are problematic, because the time of determining the plateau level is somewhat arbitrary. While we still contend that differences in plateau levels still should be reflecting real

Table 5
Difference between the pre- and post-plateau slopes in the piecewise log-linear model (IP)

Replicate	Male			Female		
	CO	RSO	SO	CO	RSO	SO
1	0.140	0.0720	0.0556	0.0699	0.0438	0.0252
2	0.171	0.127	0.115	0.0889	0.0587	0.0358
3	0.175	0.135	0.0503	0.0900	0.0693	0.0100
4	0.170	0.141	0.114	0.0643	0.0604	0.0534
5	0.176	0.170	0.0484	0.0615	0.0696	0.0173
Mean	0.166	0.129	0.0768	0.0749	0.0604	0.0283
SD	0.0152	0.0359	0.0347	0.0136	0.0105	0.0170

differences in the plateau behavior, it makes sense to look to other parameters for confirmation. In particular, we consider the strength of the plateau, by looking it the change in slope. The absolute differences are given in Table 5, while the difference in logarithms is given in Table 6.

Applying the ANOVA test to the absolute differences in slope, we reject the equality of means for the males at $p=0.00055$ ($F=15.0$); the females yield $p=0.079$ ($F=3.15$). The Wilcoxon test tells a similar story: $p=0.015$ for the male CO-RSO comparison, $p=0.008$ for the male CO-SO comparison. On the other hand, we see a slightly different picture when we look at the differences in the slope logarithms. There, the ANOVA test rejects equality of means for males ($p=0.0018$; $F=11.2$) and for females ($p=0.00061$; $F=14.6$). The significant difference for the females, it must be noted, is between the SO and the other two strains, and no pair attains statistical significance individually, whether by the t test or the Wilcoxon test. For the males, the t test and the Wilcoxon test give essentially the same result as was found for the absolute slope differences.

3.2. The logistic Gompertz model

As has already been noted, while the piecewise linear model yields fairly clearcut results, it has defects that might make the results seem dubious In particular, the plateau level is not uniquely defined, and fitting the model requires that we drop the early data. While we have argued that these are reasonable procedures for the kind of statistical questions that are at issue here, it would not be inappropriate to try to confirm the results with a different model. This will also support our contention that the original analysis of Drapeau et al. made a particularly unfortunate choice of model, not that our piecewise linear model is the optimal choice.

Table 6
Difference between the logarithms of pre- and post-plateau slopes in the piecewise log-linear model (IP)

Replicate	Male			Female		
	CO	RSO	SO	CO	RSO	SO
1	1.84	0.882	0.998	1.41	0.796	0.614
2	3.62	1.61	1.75	1.08	0.960	0.980
3	3.31	1.52	0.837	1.41	1.30	0.305
4	3.14	1.81	1.77	0.835	1.26	1.20
5	2.97	1.93	0.770	0.831	1.33	0.374
Mean	2.98	1.55	1.22	1.11	1.13	0.694
SD	0.679	0.408	0.494	0.289	0.238	0.386

Table 7
Estimated plateau levels in the logistic Gompertz model (LG)

Replicate	Male			Female		
	CO	RSO	SO	CO	RSO	SO
1	0.113	0.188	0.0950	0.167	0.119	0.0954
2	0.117	0.0845	0.0537	0.165	0.108	0.0845
3	0.144	0.111	0.141	0.164	0.0982	0.104
4	0.133	0.0947	0.0769	0.164	0.119	0.115
5	0.106	0.0817	0.2274	0.162	0.118	0.159
Mean	0.123	0.112	0.118	0.165	0.112	0.112
SD	0.0154	0.0438	0.0686	0.00294	0.00921	0.0289

We present two different measures of the plateaus: the plateau level (Table 7), and the maximum negative second derivative of the log hazard (Table 8), which we take as a proxy for the strength of the plateau.

One noticeable difference to the fitting of the piecewise log-linear model is the incomplete success in avoiding large fluctuations in the characterizing parameters. The male RSO and SO flies each have one extreme outlier in the estimated plateau level, and one extreme outlier in the plateau strength parameter. This means that, despite the conspicuous trend toward lower plateaus in the RSO and SO strains, it is impossible to reject the hypothesis of equal mean plateau levels. The female flies show a much clearer pattern. Where the evidence from the log-linear model was ambiguous, yielding only borderline significant p-values, the female strains show very low variability within strains, and strong differences between strains. The plateau level estimates for the female CO replicates are all larger than the largest of the estimates for the RSO and SO replicates, yielding the minimum p-value of 0.008 for the Wilcoxon test for equality of means in either comparison CO-RSO or CO-SO. The ANOVA F-test gives us a p-value of 0.0005, while t-tests for comparing CO females with RSO and SO females yield p-values 0.00016 and 0.015, respectively.

Table 8
Strength of the plateau: maximum negative second derivative in the logistic Gompertz (LG) model

Replicate	Male			Female		
	CO	RSO	SO	CO	RSO	SO
1	0.0624	0.0118	0.0185	0.0211	0.00907	0.0193
2	0.0738	0.0656	0.0706	0.0295	0.0161	0.0129
3	0.0689	0.0139	0.00720	0.0265	0.0173	0.00937
4	0.0818	0.0636	0.198	0.0241	0.0166	0.0206
5	0.0846	0.754	0.00820	0.0195	0.0187	0.0116
Mean	0.0743	0.182	0.0605	0.0241	0.0156	0.0147
SD	0.00912	0.0.321	0.0813	0.00401	0.00376	0.00490

The same clear difference is seen in the plateau strengths, between the CO and the other females. The Wilcoxon test again yields *p*-value 0.008 for the CO-RSO comparison, and 0.016 for CO against SO. The corresponding *t*-test *p*-values are 0.008 and 0.011.

4. Simulations

In Section 3 we showed that the inferences from the data of Drapeau et al. depend on the choice of model for the plateau. Most notably, the flat-plateau model finds a wide variation in the estimates, and consequently no verifiable difference between the strains, while our alternative models find significant differences for the males or for the females, but not both.

The intuitive arguments of Section 2 are, unfortunately, difficult to make mathematically precise. An alternative is to study the question empirically, on simulated data. We consider five imaginary strains: The NO (Normal) strain has mortality rates given by the empirical mortality rates of the total population in all of the experimental replicates. This is intended to give a baseline of mortality rates that are more or less like those of real flies. We then generate HM (High Mortality) and LM (Low Mortality) hazard rates by increasing or decreasing all age-specific hazard rates by 25%, and VHM and VLM hazard rates by increasing or decreasing them by 67%. The difference between NO mortality and either V strain is quite large, close to the maximum hazard-rate differential that is sustained over any extended period of time between the strains. The difference between the VHM and VLM hazard rates are enough to produce nearly a 50% increase in life expectancy, from 38 to 55 days. Whatever we think the plateau level ought to mean, and whatever statistical estimation procedure we choose, it seems reasonable to expect that it should find a difference between the plateau levels of two strains which have the same shape, and differ by a substantial constant multiplicative factor.

The results of 2000 simulations of each virtual strain are summarized in Table 9, while histograms of all the results are presented in Figs. 4–6. In each trial, we simulated a population of 1350 flies (the average over the real trials), and then estimated the plateau level for the simulated mortality counts with each of the three models. (We have given the means, but the medians are almost the same, and the result would be qualitatively the same if we examined the means of the logarithms of the plateaus.) Ideally, the plateau level estimates should be in the ratio 1.67:1.25:1:0.8:0.6. The realized ratios are

FP	1.64:1.23:1:1.06:0.93
LG	1.63:1.21:1:0.85:0.77
IP	1.93:1.31:1:0.78:0.57

The FP and LG models perform about equally well on the high-mortality strains, coming quite close to the ideal ratios. On the low-mortality strains, though, the FP model fails spectacularly, finding an *increase* in the plateau level from the NO strain to the LM strain, and only a slight decrease even to the VLM strain. This results, as we see in Fig. 4, show a split in the breakpoint estimates exactly like the one we saw in Table 1, with about half the estimates staying down in the 40s, and the other half jumping up into the 70s. The LG model also underestimates the differences in the plateau levels, particularly for the VLM strain, but comes far closer to the truth.

Perhaps the most direct way of testing the models is to carry out the original significance test on the simulated data. What we have done is to take the 2000 simulated plateau levels for high, normal, and low, as representative of the distribution of empirical mortality levels obtained from these strains. We simulated the *Drapeau* et al. experiment 1000 times, by drawing five samples from each strain. When we performed an ANOVA *F*-test for differences among the means of three strains simultaneously, we obtained the results in

Table 9
Mean and standard deviation for 2000 simulated plateau levels and breakpoints from each of the five virtual strains

Parameter		Model	Strain				
			VHM	HM	NO	LM	VLM
Plateau level	Mean	FP	0.115	0.0863	0.0701	0.0745	0.0655
		LG	0.124	0.0923	0.0760	0.0644	0.0586
		IP	0.104	0.0706	0.0538	0.0418	0.0304
	SD	FP	0.0043	0.0062	0.0032	0.0179	0.013
		LG	0.0051	0.00318	0.0025	0.0028	0.0069
		IP	0.0057	0.0039	0.0028	0.0022	0.0015
Break-point	Mean	FP	41.5	43.3	41.6	59.1	66.1
		LG	42.0	42.1	42.5	43.4	47.8
		IP	44.4	44.1	43.9	43.8	43.7
	SD	FP	0.72	4.9	2.3	17	16
		LG	0.36	0.42	0.55	1.1	3.6
		IP	0.27	0.28	0.31	0.33	0.39

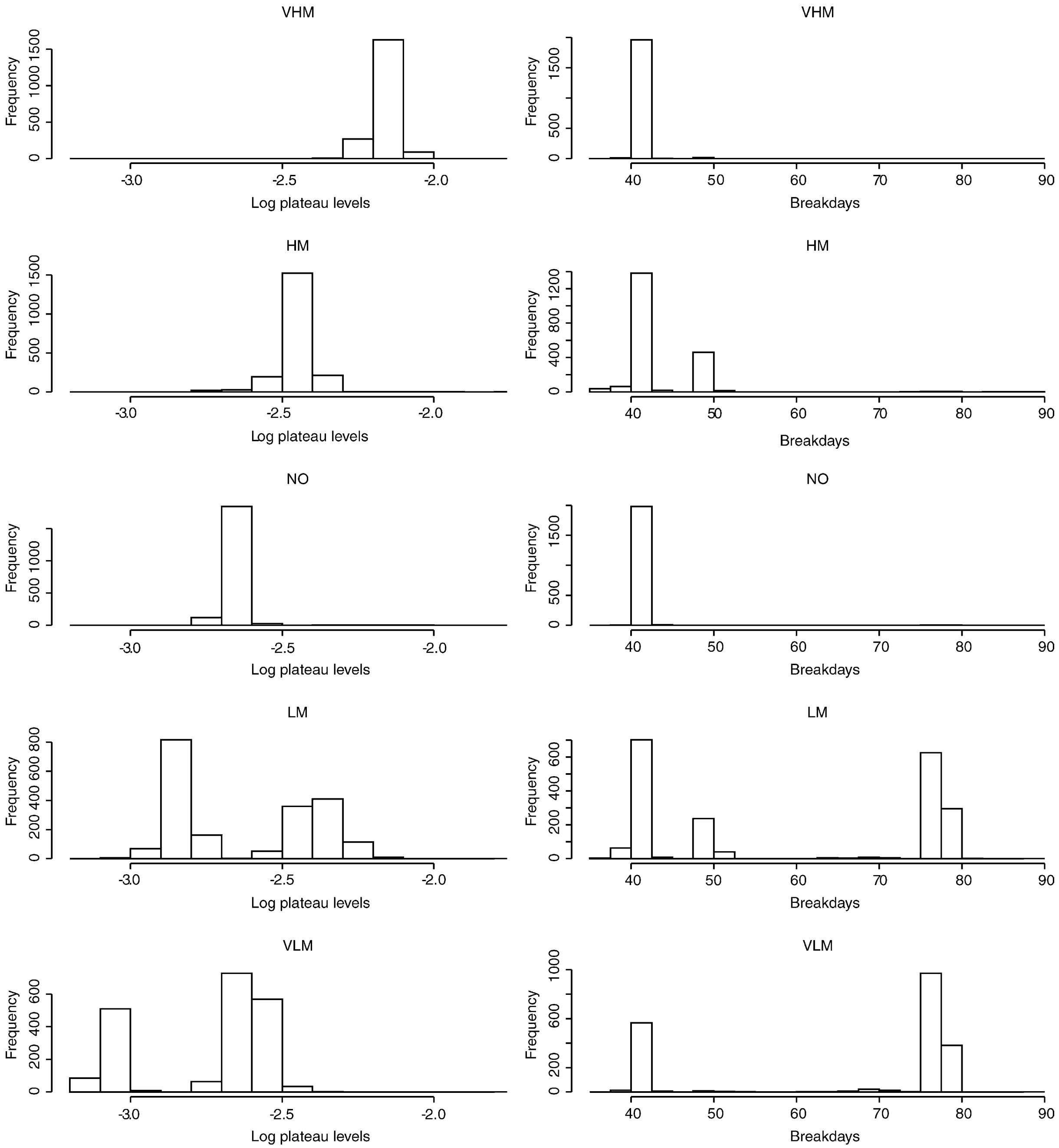

Fig. 4. Histograms of 2000 simulated plateau levels and breakpoints, estimated from the FP model.

Table 10. In column V we show the results when the three strains were VHM, NO, and VLM; in column M we show the results for more moderate differences in mortality rates, represented by the three strains HM, NO, and LM. The conclusion is clear: whereas the LG and IP models both succeed in verifying the differences in the plateau levels even at the 10^{-6} significance level, whether the mortality difference is large or moderate, the FP model is reliable only when the difference is large. When the difference in mortality levels is moderate, the FP-based test will uncover a difference at the 0.05 level in less than half the experiments.

Table 11 gives the results of taking 1000 random samples of five replicates each from the NO strain and one other, and testing (with the *t*-test) the hypothesis that the means differ. When we performed a *t*-test to test the difference in mean plateau levels, we found significant differences at the levels 0.05, 0.01, and 0.001 in a fraction of the simulations given in Table 11. As expected, the FP model fares poorly when the alternative strain has

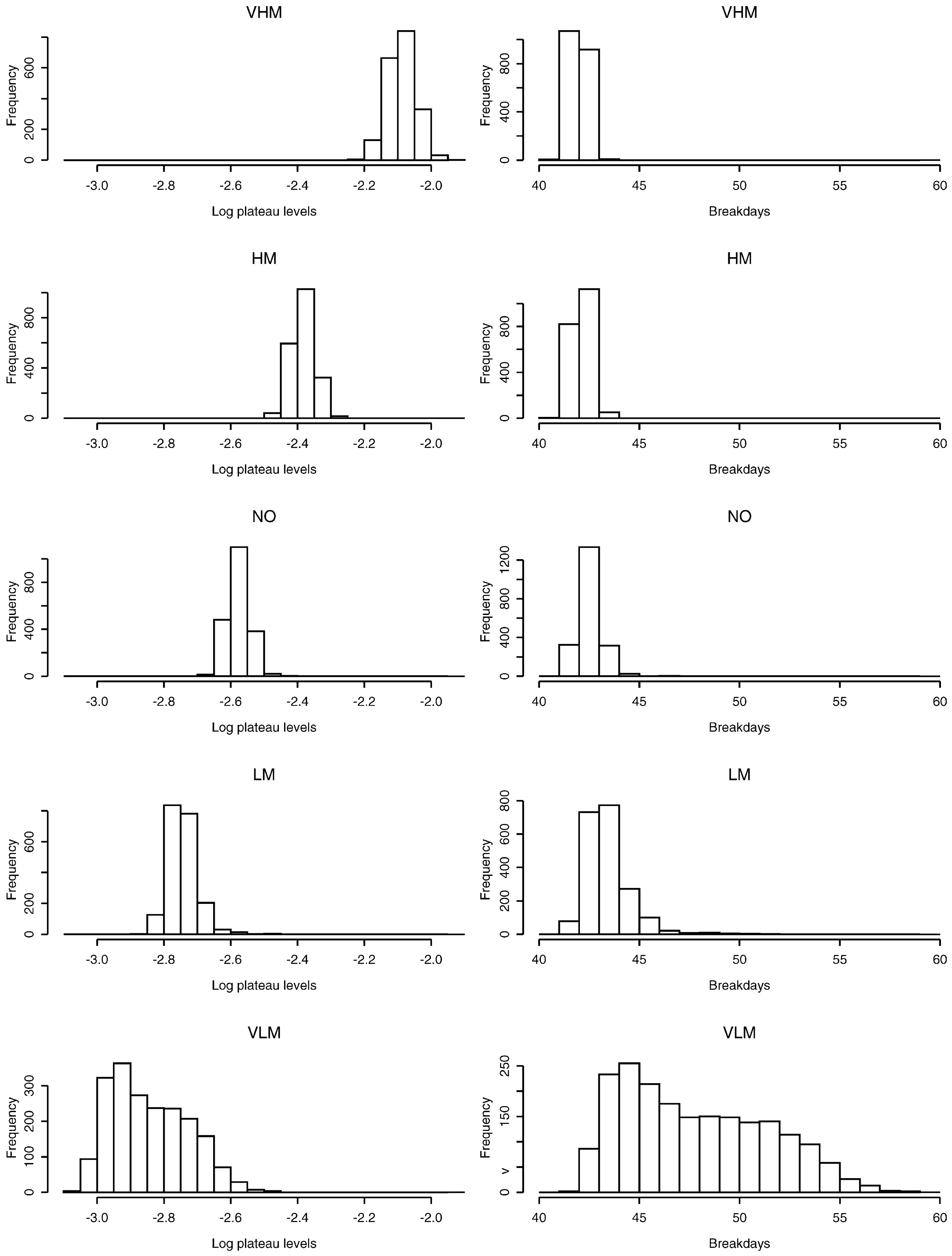

Fig. 5. Histograms of 2000 simulated plateau levels and breakpoints, estimated from the LG model.

lower-than-normal mortality. Interestingly, LM proves slightly easier to distinguish from NO than does VLM, but the difference points in the wrong direction. When the alternative is HM, the FP-based test can usually confirm a difference from the NO plateau level, though by no means always.

The IP model again performs best: nearly always it allows the difference between any of the virtual strains

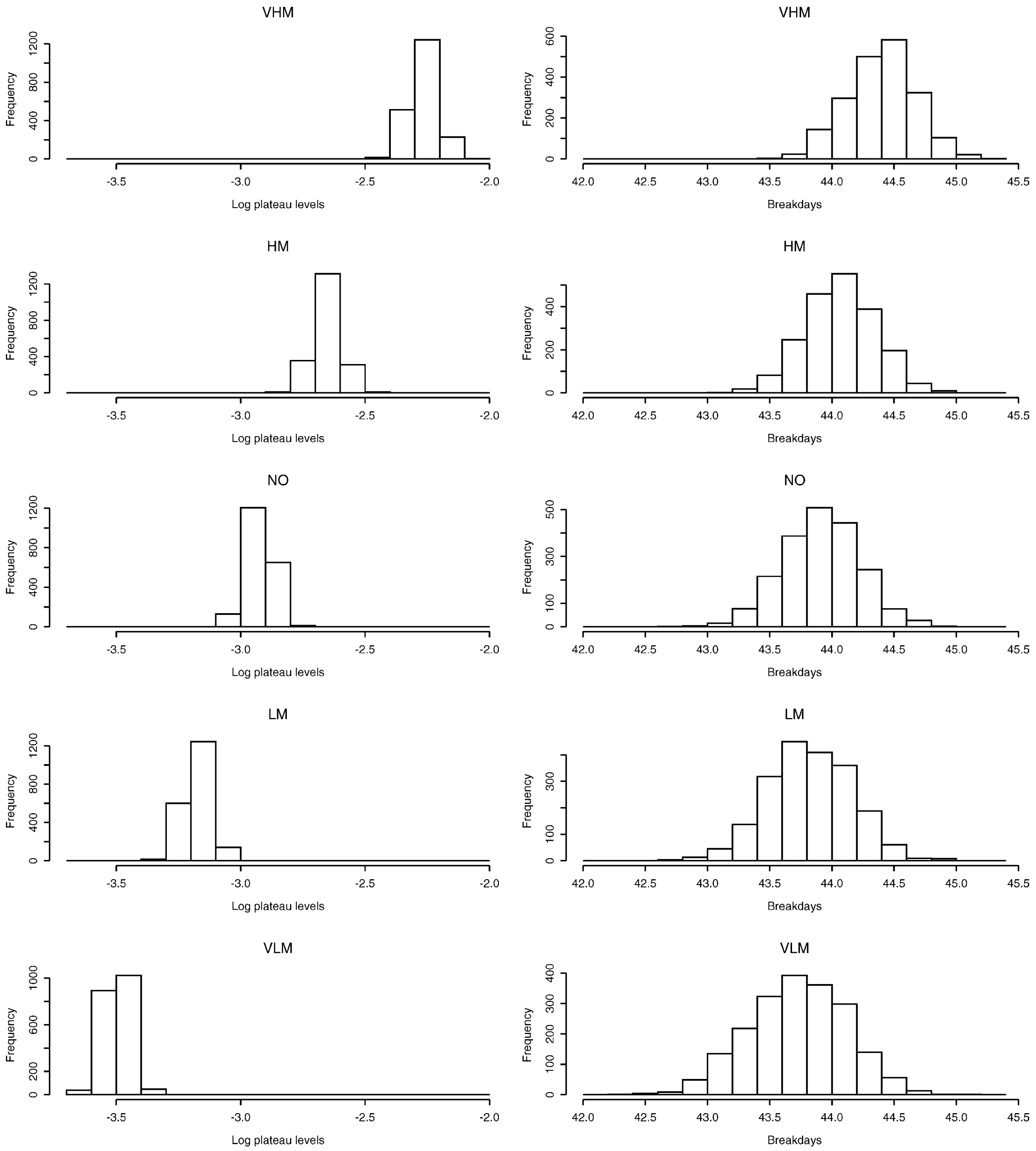

Fig. 6. Histograms of 2000 simulated plateau levels and breakpoints, estimated from the IP model.

and NO to be confirmed at the 0.01 significance level. The LG model turns in a middling performance: for the high-mortality strain, its discrimination is excellent, even somewhat better than that of the FP model. For the low-mortality strains, it nearly always delivers a significant difference at the 0.05 level. As with the FP model, the LM strain proves easier to distinguish from NO than the VLM strain, because of the significantly increased spread of the estimates for the VLM strain.

It might be argued, at this point, that we have not applied exactly the same algorithm as Drapeau et al. Maximum likelihood is a well-defined procedure, though. We have compared our likelihood function to that of Drapeau et al.[1] and confirmed that we are computing the same function, up to a constant. At the same time, any maximum-likelihood fitting procedure stands or falls on the accuracy of its optimization method. The optimization method of Drapeau et al. cannot be reproduced in detail, since it involved (L. Mueller, private communication) a significant degree of searching by hand. At the same time, we would argue that this is not a fault of the present work. We are not testing a ‘private’ algorithm, but, rather, the statistical procedure publicly described. The point of our

[1] The program is available from Laurence Mueller: idmuelle@uci.edu.

Table 10
Fraction of tries in which an F-test found a significant difference (at the given significance level) in mean level in five simulated samples each from the high-mortality, normal, and low-mortality strains

Significance level	Model					
	FP		LG		IP	
	V	M	V	M	V	M
0.05	1.0	0.434	1.0	1.0	1.0	1.0
0.01	1.0	0.164	1.0	1.0	1.0	1.0
0.001	0.998	0.074	1.0	1.0	1.0	1.0
10^{-6}	0.588	0.034	1.0	0.961	1.0	0.970

The plateau level is computed from the stated model. In the columns marked V, the high and low mortalities are represented by VHM and VLM, respectively; in the columns marked M, the high and low mortalities are represented by HM and LM, respectively.

simulations is to confirm our claim that the flat-plateau model is poorly suited to teasing out differences in plateau levels between strains. This argument would only be muddled by attempting to reproduce not only the original model, but also possible errors in the fitting procedure. Instead, we have been at pains to write a reproducible algorithm which reliably finds the true maximum likelihood, as described in Section 3. As already noted, one weakness of the flat-plateau model is that the likelihood function does not have a well-defined maximum. Exploring the likelihood function turned up widely separated local maxima, whose likelihood values differed only slightly.

Our version of the flat-plateau estimation procedure does turn up the same flaws that were so conspicuous in the original analysis: the vast spread of estimates for the plateau levels and breakdays for replicates of the same strain. Fig. 4 shows histograms of the plateau level and breakpoint estimates from the FP model. We see that, as the hazard rates fall, the plateau levels and breakpoints split into bimodal distributions. The corresponding histograms for the LG and IP models, Figs. 5 and 6 do not suffer from this problem. The breakpoint estimates from the IP model are particularly stable against random perturbation.

5. Conclusions

It may seem that the results of the new fitting procedures are ambiguous, or even contradictory. The one procedure yields significant differences for the males, the other yields significant differences for the females. On further reflection, though, it should be recognized that the results reinforce each other, and support the conclusion that there are indeed significant differences in the plateau behaviors between the different strains.

To begin, remember that there is a substantial difference between the quality of evidence provided by a negative outcome to a statistical test, and that provided by a positive outcome. The negative outcome (low p-value) tells us that the observed data would be very unlikely under the null hypothesis—in this case, if the different strains had identical plateaus. The positive outcome simply tells us that the differences observed could plausibly be the result of random fluctuations. Thus, the natural conclusion from the results described in Sections 3.1 and 3.2 is that the CO plateaus differ substantially from the RSO and the SO plateaus, for males and for females.

In addition, recall that the parameter estimates are all prone to significant fluctuations and instabilities. When the data for five replicates of the same strain yield substantially the same parameter estimates, this may be taken as strong evidence that this estimate is truly characteristic of the strain. When, on the other hand, the estimates for replicates of the same strain vary widely, this could be merely an unfortunate failure of this model to tame these data effectively. This forbids drawing any conclusions about that strain; but ought not be seen as contradicting a better result from a different model, which brings the behavior under control. The stability of our proposed models under random fluctuations, which has been confirmed by the simulations of Section 4, should be reassuring.

It could not be claimed that this experiment is decisive, for or against the heterogeneity explanation for mortality plateaus. It does suggest, however, that selection genuinely does shift the plateau behavior. One deficiency is the lack of a concretely defined heterogeneity hypothesis, to which the results may be compared. What precisely should we have expected, if the heterogeneity explanation were correct? To shed more light on these questions will require more careful analysis of heterogeneity's predictions, such as those of Service (2000), to be coupled with more powerful experiments. Some steps in this direction have recently been taken by in Mueller et al. (2003). There are,

Table 11
Fraction of samples for which a T-test found a significant difference in mean plateau level in five simulated samples from the stated virtual strain and the NO virtual strain, with the stated model for the plateau level

Sig. lev.	Model											
	FP				LG				IP			
	HM	LM	VHM	VLM	HM	LM	VHM	VLM	HM	LM	VHM	VLM
0.05	0.926	0.082	0.999	0.076	1.0	0.982	1.0	0.974	1.0	1.0	1.0	1.0
0.01	0.841	0.074	0.988	0.022	1.0	0.952	1.0	0.774	0.996	0.992	1.0	1.0
0.001	0.697	0.061	0.970	0.004	0.971	0.835	1.0	0.346	0.881	0.855	1.0	1.0

however, substantial weaknesses in this work as well, as argued by Service (2004).

Acknowledgements

The author would like to thank David Brillinger, Ken Wachter, Larry Mueller, Brian Charlesworth, and Hans-Georg Müller for helpful discussions concerning this problem. The author would also like to thank Mark Drapeau and Larry Mueller for making their data available.

References

Brooks, A., Lithgow, G.J., Johnson, T.E., 1994. Mortality rates in a genetically heterogeneous population of *Caenorhabditis elegans*. Science 263 (5143), 668–671.

de Grey, A.D.N.J., 2003a. Critique of the demographic evidence for late-life non-senescence. Biochemical Society Transactions 31 (2), 452–454.

de Grey, A.D.N.J., 2003b. Maximum-likelihood fitting falsely convicts heterogeneity hypothesis. Experimental Gerontology 38, 921–923.

Drapeau, M.D., Gass, E.K., Simison, M.D., Mueller, L.D., Rose, M.R., 2000. Testing the heterogeneity theory of late-life mortality plateaus by using cohorts of *Drosophila melanogaster*. Experimental Gerontology 35, 71–84.

Horiuchi, S., Wilmoth, J.R., 1998. Deceleration in the age pattern of mortality at older ages. Demography 35 (4), 391–412.

Miyo, T., Charlesworth, B., Age-specific mortality rates of reproducing and nonreproducing males of *Drosophila melanogaster*, preprint.

Mueller, L.D., Drapeau, M.D., Adams, C.S., Hammerle, C.W., Doyal, K.M., Jazayeri, A.J., Ly, T., Beguwala, S.A., Mamidi, A.R., Rose, M.R., 2003. Statistical tests of demographic heterogeneity theories. Experimental Gerontology 38, 373–386.

Nelder, J.A., Mead, R., 1965. A simplex algorithm for function minimization. Computer Journal 7, 308–313.

Pletcher, S.D., 1999. Model fitting and hypothesis testing for age-specific mortality data. Journal of Evolutionary Biology 12 (3), 430–439.

Pletcher, S.D., Curtsinger, J.W., 1998. Mortality plateaus and the evolution of senescence: why are old-age mortality rates so low?. Evolution 52 (2), 454–464.

Service, P.M., 2000. Heterogeneity in individual mortality risk and its importance for evolutionary studies of senescence. The American Naturalist 156 (1), 1–13.

Service, P.M., 2004. Demographic heterogeneity explains age-specific patterns of genetic variance in mortality rates. Experimental Gerontology 39 (1), 25–30.

Service, P.M., et al., 2000. Stress resistance, heterogeneity, and mortality plateaus: a comment on drapeau. Experimental Gerontology 35, 1085–1087.

Vaupel, J.W., Carey, J.R., 1993. Compositional interpretations of medfly mortality. Science 260 (5114), 1666–1667.

Vaupel, J.W., Manton, K.G., Stallard, E., 1979. The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography 16 (3), 439–454.

Vaupel, J.W., Carey, J.R., Christensen, K., Johnson, T.E., Yashin, A.I., Holm, N.V., Iachine, I.A., Kannisto, V., Khazaeli, A.A., Liedo, P., Longo, V.D., Zeng, Y., Manton, K.G., Curtsinger, J.W., 1998. Biodemographic trajectories of longevity. Science 280 (5365), 855–860.

Wang, J.-L., Müller, H.-G., Capra, W.B., 1998. Analysis of oldest-old mortality: lifetables revisited. Annals of Statistics 26 (1), 126–163.