Extended Response to Weir and Smith

Sven E. Wilson
Brigham Young University

Benjamin L. Howell
Johns Hopkins University

June 20, 2007

Author Contact Information:

Sven E. Wilson
732 SWKT
Brigham Young University

Phone: (801) 422-9018
Email: sven@byu.edu

Abstract

Evidence provided by Weir and Smith, particularly the findings from the National Health and Nutrition Examination Survey (NHANES), leads us to conclude that an increase in arthritis prevalence during the 1990s in the United States is probable, but the trend is likely overstated in the Health and Retirement Study (HRS). However, we show here that a mistake in our earlier method does *not* change substantively our previous conclusion that survey duration effects (meaning that participation in the survey induces a higher reporting rate over time) are occurring in the HRS. This finding is supported by a re-analysis of matched samples from the different HRS cohorts and a variety of regression models (including that of Weir and Smith). In particular, our pooled regression estimates show not only important survey duration effects but also a time effect that is much more reasonable than the highly implausible estimate reported by Weir and Smith. Furthermore, our analysis contradicts two additional claims made by Weir and Smith. First, we show that that very little evidence exists for an upward trend among self-reporters in the National Health Interview Survey (NHIS). And, second, we use Weir and Smith's own evidence (along with other approaches) to show that less than 25% of the increase in the HRS over the 1990s can be attributed to increases in obesity. When weighed together, the different pieces of evidence all suggest that the dramatic increase in arthritis prevalence found in the 1990s significantly overstates the actual trend and that any increase, if real, is due largely to unobserved factors, not obesity.

Introduction

In 2005 we published a short paper in *Social Science and Medicine* that illustrated a trend in arthritis prevalence in the US Health and Retirement Study. The motivation for this analysis was not the increase, but the *magnitude* of the increase: prevalence increased by roughly one percentage point per year starting from 1992 (36.6%) to 2002 (45.3%) for those aged 55-59. Trends for older and younger respondents can be calculated for certain years. For those 60-64, prevalence increased from 49.4% (1996) to 53.7% (2002). For those age 51-54, increases were somewhat smaller, from 28.9% (1992) to 31.7% (1998). Our concern about this increase was amplified when we compared similar age groups to respondents from the National Health Interview Surveys (NHIS). We will show that Weir and Smith's conclusions that there is increasing prevalence of arthritis among those who self-report (as opposed to proxy-report) does not withstand further scrutiny.

Weir and Smith (W&S) argue that the trend in the HRS is real. They attribute this increase to rising obesity, but they do not investigate this claim empirically. We show here, using a variety of methods including the estimated model of Weir and Smith, that changes in the distribution of Body Mass Index (BMI) can explain at most 25% of the increase in prevalence during the 1990s. Rising obesity is probably associated with a whole host of health problems, including arthritis, but the obesity epidemic would have to have been *massive* to account for the magnitude of the increase in the HRS. What other risk factors could explain the trend, especially during a time when age-specific disability rates have been generally falling? Paying close attention to potential survey design effects in explaining an excepted finding is always worthwhile, particularly when there is no ready epidemiological or demographic explanation.

The most compelling evidence presented by Weir and Smith that arthritis prevalence increased in the 1990s is from the National Health and Nutrition Examination Surveys (NHANES), where they report that the magnitude of the increase is roughly the same as in the HRS. In this analysis we do not attempt to replicate and confirm their findings from the NHANES, though that would be a worthwhile exercise. A change in question wording is surely responsible for some of the estimated increase in the NHANES, but neither we nor Weir and Smith really know how much. A largely unresolved puzzle is why the trends in the NHANES and the NHIS—both cross-sectional surveys—are so different from one another. An answer to that question would help us compare those surveys to the longitudinal HRS.

The evidence presented from the NHANES, if valid, definitely causes us to rethink our earlier conclusion that the HRS trend is spurious. However, we do maintain that there is still considerable cause to think that the trend in the HRS is overstated. Weir and Smith focus much of their attention on our conjecture about "panel conditioning"—the idea that participation in a panel survey causes respondents to change their disease reporting tendencies over time. We discussed this conjecture only briefly in our original paper and admit that we had "no ready evidence" for its validity, but it, along with an error we made in using the HRS, become the focus of the Weir and Smith critique. We show here the panel conditioning remains a viable hypothesis consistent with the evidence and that the error we made using the HRS turns out to be relatively benign.

This report is an extended version of a response forthcoming in *Social Science and Medicine* that will accompany Weir and Smith's critique of our earlier work. It provides the detailed evidence that could not be fully presented in the published piece due to space restrictions.

Comparisons with NHIS Results

W&S make quick work of the NHIS. Too quick. Their main criticism is that we should have excluded proxy-reporters from our analysis of the NHIS, since they tend to report lower prevalence than self-reporters. We present a detailed view of the trends in the NHIS between 1992 and 1996 that breaks out the numbers in the NHIS by proxy-reporting and self-reporting status. We do this for both wide and narrow age ranges.

In the wider age intervals at the top of the Table 1, the self-reported prevalence of arthritis in the NHIS is remarkably stable, rising from .334 to .347 for those aged 51-61 in the given year and falling slightly from .301 to .296 for those aged 45-64. Certainly there is no evidence of dramatically increasing prevalence here—nor any increase whatsoever. Interestingly, for both these age groupings, the proxy-reported group experiences significant declines over time, even though we can think of no obvious reason why the tendency of proxies to report conditions—especially among different cross-sectional samples—should be falling over time, other than possible changes in the administration of the questionnaire of which we are unaware. This decline among proxy-reported cases explains the slight fall in arthritis for the sample as a whole. However, despite the inexplicable decline among proxy-reporters, when we restrict the analysis to self-reporters, as promoted by Weir and Smith, we find that the trend is virtually flat.

Where do Weir and Smith, then, get their estimates of increasing prevalence from the NHIS? They report results only for the 55-59 year group, where prevalence among self-reporters rises from .322 to .377, a total of 5.5 percentage points in 5 years. Of course there is nothing wrong with using this age group, since this group is the only 5-year interval that is directly

Table 1: NHIS Arthritis Prevalence: 1992-1996 Cross-Sections

HRS Comparison Age Range: 51-61

Group	1992	95% C.I.	1994	95% C.I.	1996	95% C.I.
All	0.300	(.274 .325)	0.283	(.259 .308)	0.281	(.243 .319)
Proxy-reported	0.241	(.203 .280)	0.190	(.155 .224)	0.154	(.105 .203)
Self-reported	**0.334**	(.300 .369)	**0.333**	(.300 .366)	**0.347**	(.295 .399)

Age Range used by NHIS Publications: 45-64

Group	1992	95% C.I.	1994	95% C.I.	1996	95% C.I.
All	0.260	(.244 .276)	0.239	(.224 .254)	0.240	(.216 .264)
Proxy-reported	0.189	(.165 .213)	0.150	(.129 .171)	0.146	(.114 .177)
Self-reported	**0.301**	(.278 .323)	**0.291**	(.270 .313)	**0.296**	(.262 .330)

Specific Age Intervals:

Age: 45-49

Group	1992	95% C.I.	1994	95% C.I.	1996	95% C.I.
All	0.156	(.132 .179)	0.154	(.132 .176)	0.143	(.111 .176)
Proxy-reported	0.098	(.067 .128)	0.096	(.069 .124)	0.095	(.055 .135)
Self-reported	**0.190**	(.157 .223)	**0.194**	(.162 .227)	**0.181**	(.132 .230)

Age: 50-54

Group	1992	95% C.I.	1994	95% C.I.	1996	95% C.I.
All	0.236	(.204 .268)	0.233	(.202 .264)	0.226	(.179 .273)
Proxy-reported	0.191	(.145 .237)	0.145	(.104 .185)	0.125	(.064 .185)
Self-reported	**0.267**	(.222 .311)	**0.286**	(.242 .329)	**0.279**	(.214 .344)

Age: 55-59

Group	1992	95% C.I.	1994	95% C.I.	1996	95% C.I.
All	0.298	(.259 .337)	0.302	(.263 .340)	0.301	(.240 .362)
Proxy-reported	0.255	(.194 .315)	0.202	(.147 .257)	0.166	(.090 .241)
Self-reported	**0.322**	(.272 .372)	**0.353**	(.301 .404)	**0.377**	(.292 .463)

Age: 60-64

Group	1992	95% C.I.	1994	95% C.I.	1996	95% C.I.
All	0.402	(.356 .447)	0.322	(.280 .363)	0.370	(.299 .441)
Proxy-reported	0.269	(.202 .335)	0.213	(.153 .272)	0.290	(.175 .406)
Self-reported	**0.465**	(.405 .524)	**0.375**	(.320 .430)	**0.404**	(.316 .493)

Age: 65-69

Group	1992	95% C.I.	1994	95% C.I.	1996	95% C.I.
All	0.433	(.384 .483)	0.467	(.416 .519)	0.443	(.363 .522)
Proxy-reported	0.350	(.261 .440)	0.435	(.334 .535)	0.358	(.218 .498)
Self-reported	**0.461**	(.402 .520)	**0.478**	(.418 .538)	**0.473**	(.377 .569)

comparable to the HRS. However, Weir and Smith fail to note that prevalence for this narrow age range cannot be estimated very precisely given the small sample sizes. They also do not mention that a similar upward trend is not found for any other age group. For instance, neighboring age categories show only a 1.3 point increase for those 50-54 and a 6.1 point *decline* for those 60-64. Perhaps there is a some strange combination of period and cohort effects working their way through the population that causes this rise for the 55-59 group during the five year period 1992-1996, but a more likely explanation is the inherent noisiness of the results for narrow age intervals. As noted above, examining prevalence for wider age groups (with much narrower confidence intervals) shows that there is no upward trend in the NHIS.

We also note a discrepancy between the NHIS presented in our Table 1 below and those found in Table 1 of Weir and Smith, though the magnitude of the difference is only a couple of percentage points and does not affect the trend appreciably. Unfortunately, calculating prevalence estimates from the NHIS for this period of time is quite complicated, given the unwieldy data files and procedures released by the NHIS. However, the method we employ for all our results yields estimates that match the published NHIS results[1] for groups that can be compared. We are, therefore, confident that our estimates for the other age intervals are accurate.

Another way to compare the NHIS and the HRS is to compare how prevalence increases over time within the same birth cohort. Since the NHIS is a repeated cross-section, we cannot follow the same individuals over time, but we can create a synthetic cohort of sample respondents who were age 51-61 in 1992 and compare it to the true cohort of similarly aged individuals in the HRS. Prevalence in the HRS cohort increases by 11.0 percentage points (from

[1] See National Center for Health Statistics (1994, 1885, 1999).

.339 to .449) between 1992 and 1996 but by only 5.9 points (from .334 to .393) among self-reporters in the NHIS synthetic cohort.

Finally, the NHIS has undergone significant revisions since 1992 which make tracking longer term trends impossible. However, we can examine changes during short periods between the revisions. Our analysis of the NHIS (still using self-reporters) in later periods shows no upward trend whatsoever after 1998. For those age 51-61, prevalence declined from .290 to .289 between 1998 and 2000 and declined again from .332 to .323 between 2002 and 2004 (a significant change in question wording occurred between 2000 and 2002). Thus, among self-reporters in the NHIS, we find *no* increase from 1992-1996, *no* increase from 1998-2000 and *no* increase from 2002-2004. Simply put, the stable prevalence levels among self-reporters in the NHIS are simply not consistent with the sharp upward trend in the HRS, and the issue of proxy reporting has nothing to do with this inconsistency.

Comparisons With New HRS Cohorts

As noted by W&S, our original analysis failed to account for the non-representative nature of the new HRS cohorts (taken in 1998 and 2004). As they note, the new respondents added to the sample in 1998 were not representative of the population because only those cases who did not meet the sample eligibility requirements in 1992 were added in the 1998 cohort. More specifically, respondents added in 1998 were aged 51-56 (birth years 1942-1947) who did *not* have a spouse born between 1931-1941, since having a spouse in the 1931-1941 birth cohort would make the individual part of the original sampling frame. Thus the new cohort of 51-56 in 1998 was demographically quite different from the original sample of those aged 51-56 in 1992.

The subsequent addition in 2004 (which we did not exploit in our original analysis) was added according to similar criteria.

What we showed in our 2005 paper is that 1) when the new cohort entered the sample in 1998, arthritis prevalence plummeted to a level markedly lower than what would be predicted based on the sharp trend in the original cohort and 2) in subsequent years (2000 and 2002), the prevalence in the new cohort rose sharply again, mimicking the original cohort. This repeating pattern strongly suggested to us (and still does) that the upward trend is related in some way to survey duration. Does the error pointed out by Weir and Smith undermine these basic results? Fortunately, it does not.

In our 2005 paper, we conduct a direct comparison between those aged 55-56 in 1992-1996 and those aged 55-56 in 1998-2002 ignoring the sample differences discussed above. Weir and Smith were right to point out our mistake, but their Table 2 seriously misrepresents our analysis. In the second column of Table 2 they compare the prevalence in 1998 between the original cohort (42.8%) and the war babies cohort (32.6%) for those aged 53-56.[2] They claim that "Wilson and Howell contrasted rates obtained in 1998 from the incremental 'war babies' sample with persons the same age from the original HRS sample." This statement suggests that we are making a comparison in 1998 such as they one they make in their Table 2, but this is false. We made no such comparison. We did not use the original cohort in 1998 or beyond because, in 1998, the original cohort is only representative of the population for those aged 57-67. We were careful to restrict our analysis to only those cases aged 55-56, since that is the only age-range where one of the HRS cohorts belongs to the age-eligible, representative sample in

[2] We remain perplexed as to why Weir and Smith chose this unusual age interval.

each survey year, and we were careful not to make any comparison between cohorts within a given survey year.[3]

Can the mistake we made earlier be repaired? Yes, and we show that the consequences of our error were, contrary to what W&S claim, relatively benign. We do this by drawing a subset of the 1992 original cohort that is directly comparable to later cohorts. Recall that the new cases in 1998 were people born in 1942-1947 who do not have a spouse in 1998 who was age-eligible (and, hence, already in the sample) in 1992. We can replicate this selection criteria in the 1992 cohort by picking those cases who were born from 1936-1941 but do not have a spouse born between 1925 and 1935. We refer to the 1992 sub-sample with this restriction and the new cases from the 1998 and 2004 samples as the "matched samples." Table 2 shows that the matched samples are almost identical to each other in terms of age, sex, marital status, race, and percent Hispanic. As expected, education and obesity are rising slightly with successive cohorts, and smoking is falling). We stress that the matched samples are *not* representative of the population; but they *are* representative of each other. Furthermore, each matched sample consists only of observations that are age-eligible within the year they are used.[4]

[3] Technically, all the analysis we present here is based on birth years, not ages, so by 55-56 we mean the birth years associated with those ages in each year ('36-'37 in 1992, '38-'39 in 1994, etc.). The original cohort was born between 1931-1941, the war babies cohort between 1942-1947, and the early baby boomers between 1948-1953.

[4] We note that the 2004 sample does not satisfy strictly the criteria used for the 1992 and 1998 matched samples. This is because the 2004 incremental cohort (the Early Baby Boomers) are born between 1948-1953 who do not have a spouse who is born between 1931-1947; in other words, married respondents have a spouse who is either younger than the respondent or much older (at least 17 years older). The 1992 and 1998 cohorts, on the other hand, include some people with a spouse who are slightly younger, (they include those who have spouses 11-16 years older than the respondent, depending on birth year). We have conducted a sensitivity analysis of this issue by creating matched samples for 1992 and 1998 for the spousal age range criteria used in 2004 and found that the results are unchanged. Essentially, because so few individuals have spouses that are more than 10 years older than them, using the 2004 criteria instead of the 1998 criteria leads to only trivial differences. As Table 2 illustrates, the matched samples used in this analysis are demographically very similar to one another.

TABLE 2: Descriptive Statistics for Matched Samples

	Mean or Percent		
Variable	*1992*	*1998*	*2004*
Age	53.46	53.34	53.37
Percent Female	0.44	0.45	0.45
Percent Black	0.11	0.11	0.12
Percent Hispanic	0.07	0.07	0.10
Percent Married	0.70	0.69	0.65
Years of Schooling	12.57	13.18	13.56
Percent Overweight	0.41	0.41	0.41
Percent Obese	0.23	0.30	0.30
Percent Ever Smoked	0.66	0.62	0.55
Percent Smoke Now	0.29	0.25	0.23

Figure 1 shows prevalence for ages 55-56 in the matched samples. In the absence of survey design effects, we would expect only minor discontinuities between the cohorts; in other words, connecting the dots between 1996 and 1998 and between 2002 and 2004 should give a relatively smooth line. Clearly this is not the case. The prevalence between 1996 and 1998 not only doesn't continue the 1992-1996 trend, but it falls sharply and then rises again in the next two waves. We confirm this same pattern again with the entry of the new cohort in 2004: the 55-56 year group switches from the war babies cohort to the early baby boomers cohort and the prevalence drops precipitously. Thus the general pattern we identified in our original paper—rapidly increasing prevalence followed by a sharp drop when a new cohort enters the sample—is clearly present in our new analysis and is further confirmed with the sharp drop in 2004, a data point not available when we did our original work.

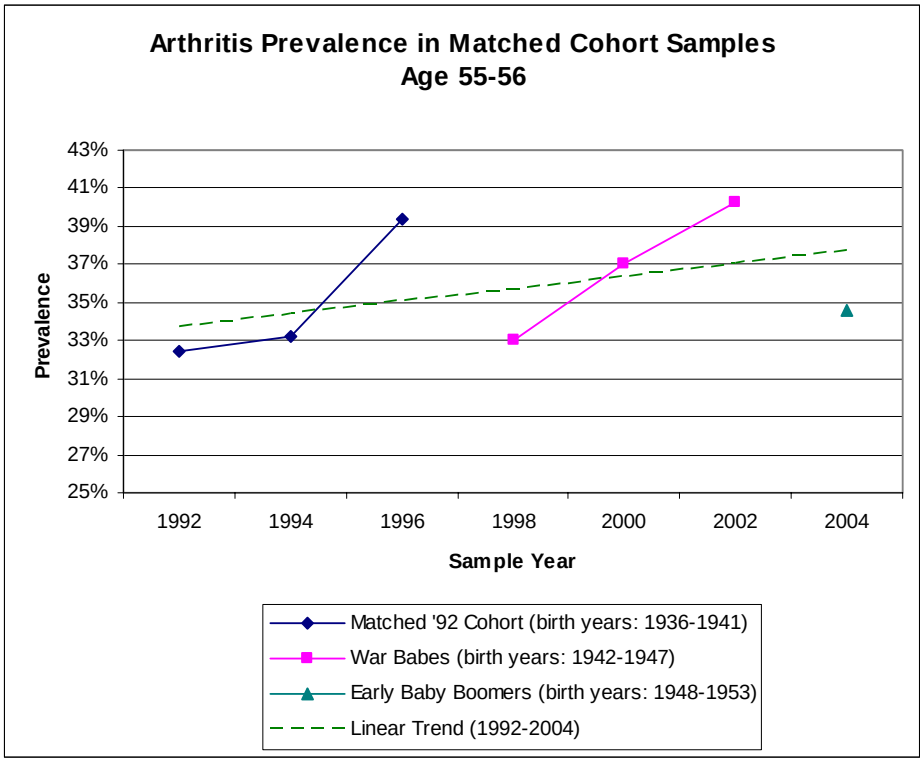

Figure 1

The repeating cohort patterns shown in Figure 1 are, we should note, associated with quite a bit of noise due to the narrow age range, though the drop between 1996 and 1998 is still statistically significant. It is quite conceivable that the data are consistent with an upward trend in prevalence over the course of the study. Indeed, if we look at the prevalences at baseline for each matched sample (1992, 1998, and 2004), we see a very modest increase over time. And if we fit a trend line through all seven points (R^2=.19), it shows an increase of about .32 percentage points per year—which is not large or statistically significant (t=1.1), but neither is it a trivial with respect to the trend in the overall HRS. However, the sharp drops that occur when the cohorts enter in 1998 and 2004 are strongly suggestive of survey design effects.

The effects of new cohorts can also be seen in the full HRS sample, though they are much less dramatic because the full sample mixes cases from different cohorts. This is illustrated in Figure 2, which shows arthritis prevalence for ages 55-59 using the full sample, where the large (blue) squares indicate the HRS values. The upper (green) dashed line indicates the linear trend fit to the first three waves of the HRS, 1992-1996, and the lower (maroon) dashed line indicates the trend implied by the next three waves, 1998-2002. Each time a new cohort enters (1998 and 2004), there is a clear deviation from the previous trend, consistent with what we found with the matched samples. But since the full sample is a mixture of the different cohorts, trends in the full sample do not reveal the effects of individual cohorts, which is why we performed the analysis with the matched samples above. An alternative approach is regression analysis with the full sample, a topic to which we now turn.

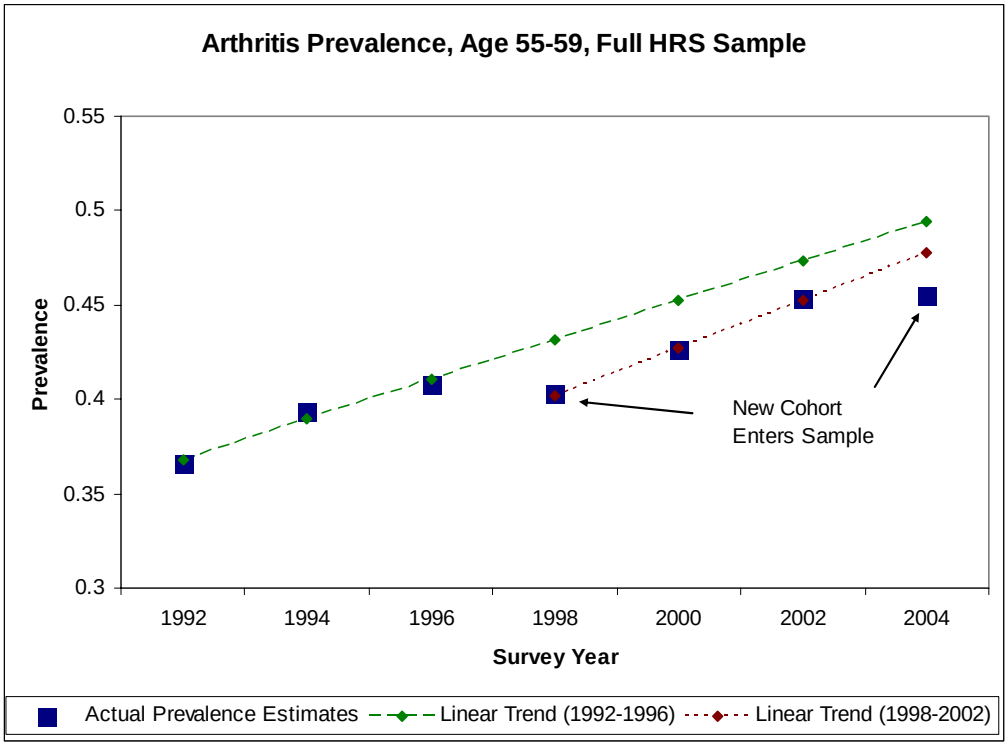

Figure 2

Alternative Regression Results

We are very supportive of the general idea of using regression analysis to identify survey design effects. Weir and Smith use a probit model to test for the presence of a survey duration effect. Controlling for both age and calendar time (as well as other controls, such as education and obesity), they incorporate a variable that measures years as a participant in the survey. Their estimate is that each year as a participant in the survey raises prevalence by .44 percentage points (z=1.2).

Unfortunately, Weir and Smith commit the common mistake of confusing statistical and substantive significance. They quickly dismiss the survey duration effect because it is not statistically significant. However, this effect is clearly *substantively* significant. A marginal effect of .0044 actually represents a 4.5 percentage point increase[5] over a 10 year period—half the actual increase in the HRS from 1992-2002.[6] Furthermore, they fail to discuss the magnitude of any other coefficient estimates. In particular, their estimate of the effect of time is .194, meaning that their model predicts a rise of 18.6 percentage points over a decade, even after controlling for obesity and other factors. What could possibly account for a trend that is this large, even over a shorter period of time such as the sample years they used in their analysis?[7] This estimate strikes us as wildly implausible and casts some doubt on their overall specification.

[5] In the probit model, the increase is calculated as Prob(arthritis|duration=10) - Prob(arthritis|duration=0), holding other variables constant at their means. The exact changes in probability tend to be very close to the linear estimates obtained from multiplying the marginal effect by the change in the variable of interest.

[6] This value is given only as a reference for gauging the value of the coefficient estimate; we do not mean to say that survey duration accounts for nearly half of the increase over a decade. This is because the average increase in duration within the sample over 10 years (say from 1992-2002) does not increase by 10 because a portion of the sample at any point in time after 1996 consists of newer cohorts. Thus in 2002, the HRS consists of cases who have been in the sample 10 years and cases who have been in for 4 years (plus a few others who entered in at different times).

[7] This range of years extrapolates beyond the range of data used by Weir and Smith. But if we estimate the effect of moving from 1998-2002 (the range of years used by Weir and Smith), we still find a whopping 7.9 percentage point increase over those four years, holding other variables constant. Indeed, the entire increase in prevalence for those aged 55-59 in the HRS is only 5.1 percentage points.

But if we take their specification as valid, we find their estimate to be an important (if imprecisely measured) indication of a survey duration effects. One sensitivity check that can be performed is to more fully account for the cohort selection criteria. We thus expand the marriage category into six different groups (being single is still the reference group) depending on the age difference between the respondent and his/her spouse. With these additional controls, the effect of survey duration rises to .0074 (z=1.74).[8]

The sample selection used in W&S—respondents born in 1942-1947 and pooled from 1998 to 2002—is not unreasonable, but it is somewhat arbitrary. Further exploration, summarized in Table 3, shows that a variety of alternative models yield marginal effects of survey duration that are both substantively and statistically more significant than the Weir and Smith estimate.[9] We explore analyses that have both less pooling and more. For starters, if their pooled model is legitimate, then it is also appropriate to un-pool the data and estimate the model for a single year. The key year is 1998, the year the first new cohort entered and the year where, if the story we tell in Figure 1 is correct, we expect to find the greatest impact of survey duration. In this case the marginal effect rises to .0077 (z=2.06).[10] Weir and Smith use only individuals from the 1942-1947 birth interval in their pooled approach. However, in 1998, the HRS is representative of the population born between 1931 and 1947. Using this wider interval gives us

[8] Most of the action in the marital status categories comes from those who have a spouse more than 10 years older than them. We suspect that this has little to do with marriage market effects and much more to do with the cohort selection criteria. In this age range, those who in 1998 have a spouse more than 10 years older than them are almost all (about 91%) from the original cohort. Thus the large negative effect on this variable just reinforces the survey duration effect by shifting the initial level for the 1992 cohort downard. In this specification, we need to interpret the overall survey duration effect using both the years in survey variable and the marital status variables.

[9] Following Weir and Smith, we use robust standard errors using the cluster option in STATA.

[10] For ease of comparison to the original Weir and Smith specification, we do not exploit the expanded marital status categories used above; in each case, the survey duration effect with the expanded categories is at least as large as the effect without the simple marital status category.

Table 3: Alternative Regression Specifications

	Model							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variable	Marginal Effects (dF/Dx)							
Years (Survey Duration)	0.0044	0.0040	0.0074	0.0077	0.0127	0.0085	0.0073	0.0092
	(1.20)	*(1.13)*	*(1.74)*	*(2.06)*	*(4.50)*	*(2.23)*	*(3.04)*	*(5.39)*
Calendar Year	0.0194	0.0168	0.0146					0.00341
	(3.20)	*(2.81)*	*(2.40)*					*(3.09)*
Age at Interview	0.0129	0.0136	0.0125	0.00949	0.0136	0.0193	0.0142	0.0139
	(2.40)	*(2.56)*	*(2.33)*	*(1.76)*	*(9.14)*	*(3.34)*	*(7.74)*	*(11.4)*
Sex (female=1)	0.1495	0.146	0.153	0.122	0.132	0.108	0.134	0.144
	(7.21)	*(7.10)*	*(6.92)*	*(5.81)*	*(11.2)*	*(5.14)*	*(9.74)*	*(15.8)*
Black (0/1)	-0.0146	-0.0282	-0.0272	-0.0490	-0.0307	0.0102	-0.00166	-0.0184
	(-0.53)	*(-1.05)*	*(-1.01)*	*(-1.80)*	*(-1.88)*	*(0.39)*	*(-0.088)*	*(-1.50)*
Hispanic (0/1)	-0.1093	-0.0948	-0.0946	-0.105	-0.124	-0.0614	-0.0776	-0.112
	(-2.97)	*(-2.61)*	*(-2.61)*	*(-2.77)*	*(-5.50)*	*(-1.96)*	*(-3.13)*	*(-6.81)*
Married (0/1)	-0.0088	-0.0102		-0.0235	-0.0249	-0.0407	-0.0170	-0.0241
	(0.39)	*(-0.46)*		*(-1.01)*	*(-1.86)*	*(-1.90)*	*(-1.14)*	*(-2.48)*
Years of Education	-0.0283	-0.0269	-0.0269	-0.0260	-0.0230	-0.0127	-0.0165	-0.0205
	(-7.78)	*(-7.38)*	*(-7.37)*	*(-6.91)*	*(-11.2)*	*(-3.51)*	*(-6.83)*	*(-13.1)*
Overweight (0/1)	0.0569	0.0559	0.0569	0.0572	0.0647	0.0535	0.0563	0.0635
	(2.67)	*(2.65)*	*(2.70)*	*(2.42)*	*(4.82)*	*(2.14)*	*(3.48)*	*(6.95)*
Obese (0/1)	0.1938	0.192	0.193	0.174	0.177	0.144	0.171	0.183
	'(8.07)	*(8.05)*	*(8.10)*	*(6.71)*	*(11.8)*	*(5.51)*	*(9.97)*	*(17.1)*
Ever Smoked (0/1)	0.0622	0.0675	0.0650	0.0748	0.0625	0.0722	0.111	0.0602
	(2.92)	*(3.20)*	*(3.09)*	*(3.43)*	*(4.93)*	*(3.11)*	*(4.96)*	*(6.15)*
Current Smoker (0/1)	0.0086	0.00764	0.00770	-0.00515	-0.000790	0.0331	0.00654	0.00159
	(.36)	*(0.33)*	*(0.33)*	*(-0.21)*	*(-0.051)*	*(1.28)*	*(0.41)*	*(0.15)*
Married: 0-5 years older than spouse			-0.0130					
			(-0.48)					
Married: 5-10 years older than spouse			0.0158					
			(0.38)					
Married: 10+ years older than spouse			0.0185					
			(0.33)					
Married: 0-5 years younger than spouse			-0.00448					
			(-0.17)					
Married: 5-10 years younger than spouse			-0.0394					
			(-1.11)					
Married: 10+ years younger than spouse			-0.150					
			(-2.86)					
Observations	8,723	8,714	8,714	3,008	11,109	2,947	8,881	67,366

Robust z statistics in parentheses

Models
(1) Weir and Smith estimates (survey years: 1998-2002; birth years 1942-1947)
(2) Replication of Weir and Smith model
(3) Same as above but with expanded marital status category
(4) 1998 only, birth years 1942-1947
(5) 1998 only, birth years 1931-1947
(6) 2004 only, birth years 1948-1953
(7) 2004 only, birth years 1931-1953
(8) Fully pooled model (survey years: 1992-2004 with population-representative birth years (see text))

a very large marginal effect for years in the survey: .0127 (z=4.50).[11] Repeating this exercise for 2004, yields an effect of .0085 (z=2.23) using the narrow birth interval 1948-1953 and .0073 (z=3.04) using birth years 1931-1953. These regression results are essentially a confirmation of the pattern of abrupt discontinuities shown in Figure 1.[12]

The above alternatives are essentially robustness checks for the criteria used by Weir and Smith. Another obvious alternative not pursued or discussed by Weir and Smith is to use as much data as possible. We can use data from all seven survey waves as long as we use only observations for age ranges that are representative of the population in the year of observation. Eligible birth years, therefore, are 1931-1941 for the 1992-1996 waves; 1931-1947 for the 1998-2002 waves and 1931-1953 for the 2004 wave. We then pool observations from all seven waves of data and estimate the Weir and Smith model. This approach shows a large and statistically significant marginal effect for survey duration of .0092 (z=5.39), while the marginal effect of calendar time (.0034) is much smaller than the unreasonable value W&S estimate.

In a sense, this last pooled approach is less conservative than the Weir and Smith model because more waves are pooled into the same model. But *any* regression analysis involves pooling observations with different characteristics (age, gender, race, etc.) and assuming that the coefficient values are constant across all subgroups of the data. The Weir and Smith estimates pool individuals across six birth years, two survey cohorts, and three waves of data. Our pooled model is more extensive—23 birth years, three survey cohorts, and seven waves of data. The

[11] The older cases in this analysis are from the original cohort; hence they will be both older and have a longer survey duration. To the extent that the model does not fully capture the age effect, part of the estimated duration effect may be reflecting older age.

[12] There are some slight differences in the coefficients estimated by Weir and Smith and our replication of their model. We use Version F of the RAND data, while they appear to use Version D. They also calculated survey duration as the time between interview dates, represented as fractions of years, while we use the method they actually describe in their paper, where durations are integers. Typically, the differences in estimated results are very small. Our observation-by-observation comparisons of the data set we construct from the RAND files and the data used by Weir and Smith (graciously provided to us by the authors) has only a small handful of discrepancies.

key is whether the different characteristics are adequately controlled for in the model. We have done sensitivity checks for non-linearities in the key variables of time, survey duration, and age by using dummy variables for each year rather than a linear trend. The qualitative results do not change. As we might expect, the effect of survey duration is highest in the first year and then subsequent increases become smaller and eventually flatten out. The pattern of survey duration estimated in the model with dummy variables is given in Figure 3.

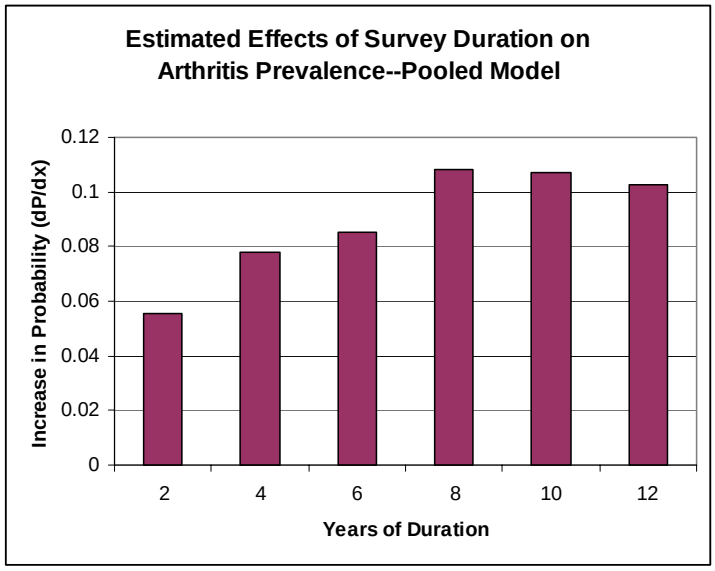

Figure 3

In our pooled model, the estimates for survey duration and calendar year are quite different than those reported in the Weir and Smith model, but the other coefficients in the model are remarkably similar. Indeed the important variables of age and gender have almost identical estimates across the two models, and all the other coefficient estimates are similar as well. Furthermore, the coefficients of control variables are relatively constant across all the alternative specifications we examine, as seen in Table 3. At least in terms of the variables in the model,

introducing the more extensive pooling does not change the implications of the model except for the survey duration and calendar year variables.

We argue that the estimates of our pooled model are reasonable. Ceteris paribus, there is an increasing trend in arthritis over the period covered by the HRS that is not explained by other factors, but the increase is modest: only 3.4 percentage points in a decade. As noted above, the Weir and Smith estimate of the time coefficient is not reasonable. The age profile implied by our model (and by Weir and Smith's) is also believable, increasing about 14 percentage points for each 10-year increase. This is broadly consistent with the average 10-year increase in the NHIS in the 1990s as well and with the age profile within the HRS baseline.[13]

Over the course of the study from 1992-2004, the prevalence among those age 55-59 rises 8.8 percentage points. The estimated increase due to survey duration in our pooled model is 4.5 points; the time effect[14] is 4.1 points; increased obesity raises the estimate by 1.9 points (more on that below), but increased education lowers it by 2.3 points; other factors explain the rest. Weir and Smith use their model only for the purpose of dismissing the hypothesis of panel conditioning. It is useful to examine its other implications. Their model has a huge time effect (with a magnitude of more than more than twice the actual increase). And, more important, they offer no explanation for what could cause such an alarming time trend. The cause cannot be obesity because obesity is already controlled in the model. Our pooled model, on the other hand, shows that the increase in prevalence is due partly to survey duration and partly to the unexplained time trend.

[13] If we fit a simple regression line to the age-specific prevalences in the 1992 HRS sample, we get a 10-year increase of prevalence of 12.9 percentage points between ages 51 and 61

[14] The time effect is not really the effect of time, but, instead, captures the impact of factors not included in the model.

Other Chronic Conditions

W&S make much of our brief discussion of panel conditioning. They apply their pooled model to estimate the determinants of hypertension, diabetes and heart disease. We think it is an excellent idea to make these comparisons since we might learn something about survey design effects by comparing different conditions. However, we do not agree with Weir and Smith regarding which conditions are more likely to exhibit panel conditioning.

The Weir and Smith assumption is that the likelihood of panel conditioning will increase with the probability of conditions being underdiagnosed. They argue that hypertension and diabetes would are often underdiagnosed and are often asymptomatic in the early stages. Our thinking on this issue, however, is that the conditions that will be most likely to increase due to panel conditioning are those that are either informally or easily diagnosed. Blood pressure is taken as a part of virtually every visit to the doctor and individuals can even test themselves with the simple little machines set up in supermarkets and other places. Arthritis is not as easily diagnosed, but, we argue, it is likely to be informally diagnosed frequently in clinical settings. For instance, a patient who mentions to a health care provider that his or her joints are frequently sore will very likely be told that this is probably arthritis (and the provider will usually be correct). Diabetes, on the other hand, requires specific testing. Typically, only patients who have a specific concern about diabetes or who are undergoing a complete physical exam will be screened for diabetes (reasons why it is underdiagnosed). Those who are visiting the physician for treatment of an acute condition (say cold or flu symptoms) or treatment of another chronic condition will not usually be screened for diabetes. Hypertension and arthritis, however, might come up in a variety of clinical settings. The heart disease category in the HRS is so broad that it

is hard to say much about it, other than many types of heart disease, such as coronary artery disease, require relatively extensive testing to diagnose.

Given our conception of panel conditioning, then, we would expect to find positive survey duration with respect to arthritis and hypertension, but not necessarily diabetes or heart disease, though if panel conditioning were strong, we would expect to see it in diabetes as well and possibly heart disease. The Weir and Smith model bears out this conception since they estimate a survey duration effect for hypertension very close to the effect for arthritis. The estimates for diabetes and heart disease, on the other hand, are relatively close to zero. We thus find the Weir and Smith results broadly consistent with our expectations about survey duration for the four conditions examined. We also estimate our fully pooled model for the four conditions, which are shown in Table 4. The survey duration variable is about half as large for hypertension as it is for arthritis, but it is non-trivial and statistically significant. The estimates for diabetes and heart disease are near zero. Furthermore, the highly unrealistic time effect in the Weir and Smith estimates is also found for hypertension and, to a lesser extent for diabetes in their Table 3. We find much more reasonable time effects for all conditions in our pooled estimates.

The purpose of our earlier paper was to point out a highly suspect increase in arthritis prevalence in the HRS. It was certainly not to confirm a theory about panel conditioning. We view our ideas about panel conditioning discussed above as highly preliminary and not well-developed. We believe the evidence presented in this paper is very suggestive of important survey duration effects, but we know very little about the mechanisms by which duration might influence disease reporting behavior and for which conditions. Better theory and evidence are certainly warranted in this case.

Table 4: Comparison of Chronic Conditions, Pooled Model

Dependent Variable:	Arthritis	Hypertension	Diabetes	Heart Disease
Coefficient		Marginal Effects (dF/Dx)		
Years (Survey Duration)	0.0092	0.0047	-0.0004	0.0012
	(5.39)	*(2.78)*	*(-0.39)*	*(1.00)*
Calendar Year	0.0034	0.0041	0.0035	0.0022
	(3.09)	*(3.75)*	*(5.32)*	*(2.85)*
Age at Interview	0.0139	0.0109	0.00470	0.00725
	(11.4)	*(9.05)*	*(6.73)*	*(9.11)*
Sex (female=1)	0.144	-0.0129	-0.0200	-0.0455
	(15.8)	*(-1.43)*	*(-3.69)*	*(-7.39)*
Black (0/1)	-0.0184	0.183	0.0858	-0.00126
	(-1.50)	*(14.4)*	*(10.9)*	*(-0.15)*
Hispanic (0/1)	-0.112	-0.0291	0.0408	-0.0579
	(-6.81)	*(-1.80)*	*(3.94)*	*(-5.33)*
Married (0/1)	-0.0241	-0.0258	-0.0136	-0.0175
	(-2.48)	*(-2.70)*	*(-2.36)*	*(-2.71)*
Years of Education	-0.0205	-0.0085	-0.0057	-0.0078
	(-13.1)	*(-5.50)*	*(-6.44)*	*(-7.62)*
Overweight (0/1)	0.0635	0.145	0.0588	0.0104
	(6.95)	(16.0)	(9.71)	(1.68)
Obese (0/1)	0.183	0.294	0.179	0.0551
	(17.1)	(27.6)	(22.7)	(7.34)
Ever Smoked (0/1)	0.0602	0.0232	0.0219	0.0421
	(6.15)	(2.39)	(3.92)	(6.42)
Current Smoker (0/1)	0.0016	-0.0299	-0.0198	-0.0137
	(0.15)	(-2.83)	(-3.28)	(-2.01)
Observations	67336	67364	67345	67361

Robust z statistics in parentheses

Pooled Model includes HRS survey years 1992-2004 and age ranges that are representive for the given year (see text)

The NHANES Estimates

The estimates from waves 3 and 4 of the National Health and Nutrition Examination Survey (NHANES) are the strongest evidence presented by Weir and Smith for an upward trend in arthritis prevalence. Over roughly the same time period, they find increases in the NHANES that are similar to increases in the HRS. Of course the levels of arthritis are much different between the two surveys as Weir and Smith's Table 1 shows. They claim that the differences in levels are due to question wording. But given that the levels are so much different (indicating they are measuring somewhat different things), it is natural to wonder how much we can make of comparisons between the trends.

The most important issue in the NHANES comparisons is the change in question wording. In wave 3 diagnoses were to come from a "doctor," and in wave 4, from a "doctor or other health professional." W&S claim that the consequences of this claim are likely minimal,[15] which seems reasonable to us, though this is just a guess. In any case, the direction of the bias seems clear, in that expanding the wording should result in higher rate of reported professional diagnoses (and possibly self-diagnoses) in wave 4. But how much higher? Enough to account for both an upward trend in prevalence and a 3-5 point increase due to survey duration in the HRS? Neither we nor W&S have any evidence to answer these questions, but it seems very reasonable for the differences to be large enough to reconcile our findings with the reported NHANES estimates.

The Role of Obesity

W&S's estimates in their Table 3 suggest a powerful role for obesity, and they fault us for not recognizing this "clear signal." However, it would have been helpful had they paid more attention to their own regression results. We show that at most 25% of the increase in arthritis can be explained by changes in the BMI distribution—which is consistent with their estimated parameters. Our findings hold if we use the same categories as Weir and Smith, but we can do a more robust test by expanding the BMI categories, since it is not just moving into the obese category that might be important, but shifts within the obese category as well. Thus in addition to the overweight category (BMI:25-30), we identify class 1 obesity (BMI:30-35), class 2 obesity (BMI: 35-40), and morbidly obese (BMI: 40+). These expanded categories increase slightly, but only slightly, the estimated impact of increasing obesity on arthritis.

Table 5 estimates the obesity-related part of the trend in arthritis prevalence using alternative assumptions that are noted at the bottom of the table. Notice first the odds ratios associated with each BMI category. Clearly, increasing BMI is highly associated with higher arthritis prevalence. The decomposition consists of estimating the trend in arthritis that would have occurred over time assuming a constant odds-ratio and the actual change in the BMI distribution (the first block of columns in Table 5). In method 1 we use the odds ratios from 1992; in method 2, we use an average odds ratio for 1992-2004; in method 3 we hold odds ratios constant at the 2004 level; in method 4 we used Weir and Smith's coefficient estimates[16] to calculate odds ratios based on the 1992 prevalence of normal BMI; and in method 5, we apply those coefficients to an average prevalence of normal BMI over the survey years. The trend in arthritis prevalence that would have resulted assuming only changes in the BMI distribution is

[16]We use their data and specification except we refine the disability categories to give more detail.

TABLE 5: Proportion of Arthritis Trend Explained by Changes in the BMI Distribution, Ages 55-59

	BMI Distribution					*Arthritis Prevalence*						*Odds Ratio*				
Year	BMI<25	25<BMI<30	30<BMI<35	35<BMI<40	BMI>40	*All*	BMI<25	25<BMI<30	30<BMI<35	35<BMI<40	BMI>40	BMI<25	25<BMI<30	30<BMI<35	35<BMI<40	BMI>40
1992	0.367	0.414	0.155	0.043	0.021	*0.366*	0.305	0.359	0.422	0.603	0.672	1.00	1.18	1.38	1.98	2.20
1994	0.357	0.407	0.170	0.045	0.021	*0.394*	0.341	0.382	0.452	0.538	0.731	1.00	1.12	1.33	1.58	2.15
1996	0.330	0.418	0.179	0.049	0.024	*0.408*	0.359	0.382	0.474	0.588	0.697	1.00	1.07	1.32	1.64	1.94
1998	0.307	0.410	0.192	0.061	0.030	*0.403*	0.343	0.389	0.442	0.532	0.653	1.00	1.13	1.29	1.55	1.90
2000	0.283	0.411	0.198	0.068	0.041	*0.426*	0.370	0.408	0.427	0.614	0.659	1.00	1.10	1.15	1.66	1.78
2002	0.280	0.396	0.195	0.083	0.046	*0.453*	0.395	0.397	0.507	0.667	0.685	1.00	1.01	1.28	1.69	1.74
2004	0.282	0.388	0.209	0.073	0.049	*0.455*	0.361	0.411	0.532	0.650	0.731	1.00	1.14	1.47	1.80	2.02

Distributional Component of Prevalence Trend

Year	(1)	(2)	(3)	(4)	(5)
1992	0.366	0.349	0.362	0.368	0.359
1994	0.368	0.350	0.364	0.370	0.361
1996	0.372	0.353	0.368	0.374	0.364
1998	0.379	0.359	0.374	0.380	0.370
2000	0.386	0.364	0.380	0.386	0.375
2002	0.391	0.368	0.384	0.391	0.379
2004	0.390	0.368	0.384	0.390	0.379
Percent Explained by Distributional Changes ('92-'04):	**26.8%**	**21.2%**	**25.3%**	**25.8%**	**22.2%**

Decomposation Method:

(1): Hold odds ratios constant at 1992 Levels

(2): Hold odds ratios constant at average levels from 1992-2004

(3): Hold odds ratios constant at 2004 level

(4): Use Weir-Smith model (2007, Table 3)--with additional BMI categories-- to calculate odds ratios based on 1992 prevalence for normal BMI group'

(5): Use Weir-Smith model (2007, Table 3)--with additional BMI categories-- to calculate odds ratios based on average (1992-2004) prevalence for normal BMI group'

given in the bottom block of columns on Table 5. The methods give similar results: 21.2% to 26.8% of the overall trend in arthritis can be linked to the change in the BMI distribution, depending on method. This percentage is even lower (17.1% to 22.6%) if we use the simpler normal/overweight/obese categorization of the Weir and Smith model (full results not shown).

In short, within-category increases dominate compositional changes. For instance, for those with BMI <25, the increase in arthritis prevalence rises from 30.5% (1992) to 39.5% (2002). Increases in obesity are a good place to start looking for why reported levels of arthritis have increased in the HRS, but they do not get us very far. Furthermore, the marginal effects of survey duration and time discussed above already control for obesity.

Conclusions

From our analysis we conclude that:

1) In general, arthritis prevalence in the NHIS, even among self-reporters, does *not* increase from 1992-1996 among those in late middle age. Only for one narrow age band (where confidence bands are wide) do we see an increase.

2) Using appropriate matched samples from each HRS sample cohort, we see abrupt declines in arthritis prevalence (among those age 55-56) when the new cohorts enters the sample. The impact of these new cohorts can even be seen when looking at the full HRS. Thus the impact of our "serious flaw" turns out not to be very serious. This pattern is also visible, though less distinct, using the full HRS for ages 55-59

3) Regression models for duration effects generally show evidence for important duration effects. Even the estimated effect from W&S is quite substantial. However, the Weir and Smith estimate for time (changes in missing factors) is wildly implausible. Our

extended pooled analysis provides reasonable estimates for survey duration and time. Survey duration accounts for about half the increase and unexplained factors the other half. Increases in obesity cause the trend to increase, but increases in education cause it to fall.

4) The analysis of other chronic conditions is not inconsistent with the idea of panel conditioning. We find, as expected, significant effects of survey duration for both arthritis and hypertension, but not for diabetes and heart disease. In our conception of panel conditioning, survey duration effects are more likely to occur in conditions that are either easily (hypertension) or informally (arthritis) diagnosed.

5) The NHANES results show a significant increase in prevalence, though the extent of the increase due to change in question wording is not known.

6) Shifts in the BMI distribution can explain at most 25% of the trend in arthritis prevalence in the HRS.

We were much too hasty in rejecting the possibility for an increasing trend in arthritis during the 1990s. Weir and Smith, as well as our subsequent work, have convinced us that an upward trend is likely. We remain, convinced, however, that the magnitude of the trend is overstated in the HRS, though we are not sure by how much. We can say, though, that all the evidence we discuss above are generally consistent with an important role for survey duration effects. We are not wedded to any particular explanation of why this trend is present, but participation in the survey, holding other factors constant, increases the tendency to report the presence of arthritis (as well as hypertension).

Weir and Smith present a useful regression approach to testing for survey duration effects. We think that a more "aggressive" pooling approach using all available data is superior and shows strong evidence of survey duration and much more reasonable estimates of the time effect. However, even the more limited pooling done by Weir and Smith still find non-trivial duration effects. They adopt a classical hypothesis test with a null hypothesis of no duration effect, and they fail to reject that null, so they accept it. This is a convenient position to take if you have a high personal stake in the integrity of the HRS. But many users of this data are also concerned with making type 2 errors (rejecting an effect that is actually there). In the end, the evidence in support of a positive survey duration effect is much stronger than the evidence that the effect is less than or equal to zero. The positive effects are not precisely estimated, but Weir and Smith's *best estimate* points to a positive effect that is non-trivial in magnitude. The additional regression models we present plus the analysis of the matched samples further support this claim.

Of course this issue would be much less confusing if we had a better understanding of factors affecting the propensity to report health problems in survey data. The most striking feature of Table 1 from W&S is the tremendous range in estimates across the surveys. Clearly study design issues are profound. And given that there is so much variation in estimation of levels, we ought to be skeptical about observed trends, especially when the trends are in conflict across surveys. More than anything, Table 1 is an intriguing puzzle.

It seems entirely plausible to us that long-term participation in a health survey could (a) increase the comfort level respondents have with respect to reporting conditions and (b) induce respondents to pay more attention to symptoms and conditions between survey waves (especially after they get the HRS participant newsletter). Furthermore, HRS design features such as moving

from face-to-face to telephone interviews after the initial interview may play a significant role in explaining the patterns in Figure 1. Given the evidence, survey duration effects are very likely present and are troubling.

References

National Center for Health Statistics (1994). Current Estimates from the National Health Interview Survey, 1992. *Vital and Health Statistics, Series 10, No. 189.*

National Center for Health Statistics (1995). Current Estimates from the National Health Interview Survey, 1994. *Vital and Health Statistics, Series 10, No. 193.*

National Center for Health Statistics (1999). Current Estimates from the National Health Interview Survey, 1996. *Vital and Health Statistics, Series 10, No. 200.*

Weir, D. & Smith, J. (2007) Do panel surveys really make people sick? US arthritis trends and the Health and Retirement Study." *Social Science and Medicine*. Forthcoming.

Wilson, S. & Howell B. (2005). Do panel surveys make people sick? US arthritis trends in the Health and Retirement Survey. *Social Science and Medicine*, 60(11), 2623-2627.

Wilson, S. & Howell B. (2007). Disease Prevalence and Survey Design Effects: A Response to Weir and Smith. *Social Science and Medicine*, Forthcoming.

.