

An Unbiased Estimate of the Relative Crash Risk of Cell Phone Conversation while Driving an Automobile

Richard Young
Wayne State Univ.

ABSTRACT

A key aim of research into cell phone tasks is to obtain an unbiased estimate of their relative risk (RR) for crashes. This paper re-examines five RR estimates of cell phone conversation in automobiles. The Toronto and Australian studies estimated an RR near 4, but used subjective estimates of driving and crash times. The OnStar, 100-Car, and a recent naturalistic study used objective measures of driving and crash times and estimated an RR near 1, not 4 - a major discrepancy. Analysis of data from GPS trip studies shows that people were in the car only 20% of the time on any given prior day at the same clock time they were in the car on a later day. Hence, the Toronto estimate of driving time during control windows must be reduced from 10 to 2 min. Given a cell phone call rate about 7 times higher when in-car than out-of-car, and correcting for misclassification of some post-crash calls as pre-crash, the final required downward adjustment of the Toronto and Australian RR estimates is about 7 times. The Toronto adjusted RR is 0.61 and the Australian adjusted RR is 0.64, which now agree with the OnStar RR estimate of 0.62. All five adjusted RR estimates for cellular conversation while driving in automobiles are near 1, with a pooled RR of 0.61 (95% confidence interval 0.51 to 0.74). Talking on a cell phone while driving an automobile does not increase crash risk relative to not talking.

CITATION: Young, R., "An Unbiased Estimate of the Relative Crash Risk of Cell Phone Conversation while Driving an Automobile," *SAE Int. J. Trans. Safety* 2(1):2014, doi:10.4271/2014-01-0446.

INTRODUCTION

Part 1 reviews the data and biases in five major epidemiological studies of the effect of cell phone conversation on the relative risk (RR)¹ of crashing while driving an automobile. These reviews inform Part 2, which develops new methods to remove bias in these studies, and re-analyzes their data to arrive at an unbiased estimate of the RR of cell phone conversation while driving an automobile. Two of the studies used a "case-crossover" analysis, one a "cohort" analysis, and two a "case-control" analysis.² For present purposes, it is only necessary to understand that, in the absence of bias, all study designs should yield RR estimates close to the same value when risks are small.¹ RR estimates that are substantially different for the same risk factor indicate that bias or error must be present in one or more of the studies, which must be removed to estimate a valid RR.

PART 1. REVIEW OF EPIDEMIOLOGICAL STUDIES OF CALLS, CRASHES, AND THEIR BIASES

1.1. Toronto Study Review

The Toronto study [3] analyzed the cell phone billing records of 699 study subjects who crashed while driving. It found that 170 of them talked on the phone³ in the 10-min *case window* (just before the estimated time of their crash), and 37 of them talked in the 10-min *control window* (during the same clock time the day before the estimated time of their crash).⁴ This study used a case-crossover analysis method that estimates the RR via the *discordant pair*.⁵ Table 1A shows the tabulated data from the Toronto study [4, Table 4].

3. Studies such as [3] that use billing records as its basic data source can make conclusions only about talking and no other aspect of cell phone use. Billing records record outbound time-stamps only after a caller finishes dialing, hits "send," the call is successfully routed, and the called party answers. There is no billing record if there is a crash while reaching for a phone, looking up phone numbers, dialing, or any other activity before the called party answers. If talking is interrupted by a crash, then talking is the proximate cause, not dialing or those other prior activities. Therefore, the term "Talk" will be used in the rest of this paper to refer to "cell phone conversation."

4. It is possible that more than one Talk occurred in a given 10-min window. If so, these would have been tabulated as a single count by the Toronto methods. Thus, technically the counts in Table 1 are not counts of Talk, but are counts of subjects who engaged in at least one Talk in a given window.

1. Under the "rare-disease assumption" [1, p. 207], the risk ratio, rate ratio, and odds ratio all approximate one another; hence, the term "relative risk" will be used here as a synonym to represent any of them, as in common epidemiological practice [2, p. 96].

2. These epidemiological terms may be unfamiliar to many in traffic safety. See Rothman [2, Chapter 5] for an introductory description.

Table 1. Toronto study Talk counts. A. Crude RR estimate. B. RR adjustment for driving intermittency (as done in [3]).

		Control Window ^b			Total
		Talk (in-car)	Not Talk (in-car)	(out-of- car)	
A. Crude	Case Window ^a Talk	13	157	0	170
	Not Talk	24	505	0	529
	Total	37	662	0	699
	RR (95% CI)	6.5	(4.5 to 10)		
B. Corrected for Driving Intermittency					
Case Window ^a	Talk	13	102	55	170
	Not Talk	24	505	0	529
	Total	37	607	55	699
	RR (95% CI)	4.3	(3.0 to 6.5)		

^a 10-min duration near crash time

^b 10-min duration, day before crash

The first two columns and rows in Table 1A are the standard 2×2 matrix used for a case-crossover analysis. The *discordant pair* is the cells containing 157 and 24. Dividing those numbers yields a “crude” (i.e., unadjusted for bias) risk ratio of 6.5, shown with the 95% confidence interval (CI) from 4.5 to 10 as reported in the Toronto study [3].

1.1.1. Driving Intermittency Bias

The Toronto investigators correctly recognized that they had to adjust for bias from *driving intermittency* (i.e., drivers who crashed, but who did not drive at all in a control window on a previous day). The investigators had the complete billing records making it possible for them to know the call rates during a control window vs. not during a control window. For example, subjects who did not drive during a control window would have a lower call rate than subjects who did drive there [5], causing a smaller RR denominator and incorrectly elevating the RR.⁶ The Toronto investigators' main correction method for intermittency was to discard the subjects who did not recall driving in a control window. However, they had no objective data as to whether a subject actually drove or not in a control window. Therefore, they surveyed another group of 100 respondents (not in the original study, thereby losing the benefits of the case-crossover method) and found that 35 percent of them remembered not driving during a “selected

period” the day before.⁷ The Toronto investigators then multiplied all their RR estimates by 0.65 to correct for this estimated driving intermittency bias.

Table 1B illustrates their intermittency correction, which is the equivalent of discarding 55 of the 170 subjects because they were out-of-car in the control window [4]. The RR estimate declined from 6.5 to 4.3 (102 divided by 24 in Table 1B), with 4.3 being the final RR as reported in the Toronto study [3].

1.1.2. Unadjusted Biases in the Toronto Study

Part-Time Driving Bias

The Toronto study neither recognized nor controlled for *part-time* driving bias, which arose from those who drove in a control window, but during only part of the time there. This part-time driving reduced driving exposure in the control windows, and gives rise to a lower number of calls in the RR denominator and erroneously inflates the RR by about 3 times [5, 6, 7, 8, 9, 10]. Part-time driving bias is not controlled for by the driving intermittency correction in Section 1.1.1. Discarding those with no driving at all in a control window does nothing to correct for those who drove there, but for only part of the time in that control window. Part-time driving bias is thus an additional bias to driving intermittency bias. Part-time driving consistently biased *all* RR estimates in the Toronto study upward by about 3 times.⁸

Misclassification Bias

Another bias arises from misclassification of calls as occurring before the crash, when in fact they occurred after the crash. These call misclassifications arise from errors in estimating crash times. Redelmeier and Tibshirani [3] recognized this misclassification bias and attempted to correct for it by various means. They did not have objective information on crash times, and so had to rely in large part on police reports. However, these tend to be rounded to the nearest 5-min mark [11], causing misclassification errors when comparing the call times in cell phone bills to the erroneous crash times. The Toronto study also used times of calls to emergency services to estimate the time of the crash, if such calls appeared in the billing records. However, the billing records in [3] showed a sharp increase in calls of all kinds after the crash, not just to emergency services. Some of those calls could have been personal calls that occurred after the crash but before the call to emergency services, which would have led to the calls being misclassified as pre-crash when, in fact, they were post-crash.

Usually in epidemiological studies, misclassification bias occurs equally in both directions, thereby causing the RR to be biased towards the null.⁹ In call and crash association studies, however, the imbalance in calls after the crash compared to

5. The *discordant pair* represents subjects who talked in the case window (but not in the control window), and subjects who did not talk in the case window (but did during the control window). In a matched case-control (or case-crossover) study, *discordant* describes a pair whose members had different exposures to the risk factor under study. Under conventional analytical methods, only the discordant pair is informative about the association between exposure and disease [1, p. 68].

6. Note that any biasing effect on RR from driving intermittency is unrelated to Talk risk; it simply means that subjects drove in the case window but did not in the control window. The same Talk imbalance between any driving period vs. any non-driving period would be observed, even with no crashes.

7. Another method used in [3] to estimate the intermittency factor was a survey of the actual subjects a year after the study, asking each subject to remember his or her driving pattern on both the day of the crash and the day before the crash the year before. Among those who responded (the number was not reported in the paper), the investigators limited their analysis to only those subjects who were confident that they had driven during both windows on both days, and obtained a similar RR.

8. The apparently consistent RR estimate near 4 irrespective of demographic factors, hand-held or hands-free phones, weekday or weekend, etc. [3, Table 2] was not because of any uniform elevated risk from cell phone use in all these conditions, but because part-time driving bias consistently caused about a 3x upward bias in all the RR estimates in the entire Toronto study.

just before the crash leads to a differential misclassification bias which biases the RR upward [12, 13, 14, 15, 16]. Bellevance [15,16] developed a model for such misclassification bias in the Toronto study, and estimated it inflated the RR by a factor of 3.

1.2. Australian Study Review

McEvoy et al. [17] interviewed 744 subjects who gave permission to access their cell phone billing records in a hospital emergency room after a crash in which they were driving and had been injured. They were asked to remember how long they had been driving at the time of the crash up to a maximum time window of 10 min before the crash. The time of the crash was estimated from emergency response records, medical records, and subjects' self-reporting. The subjects were also asked if they remembered driving during control windows of the same duration and clock time as the crash window, but at 1, 3, or 7 days before the estimated date and time of their crash. Of the 744 subjects, 456 (61%) remembered driving during one or more of the three control windows.

From the cellular billing records of those 456 subjects, the investigators found that 40 of them (9%) had a cell phone conversation sometime up to 10 min before the crash during the period in which they said they were driving. If the remembered driving time was less than 10 min, the call was counted only if it occurred during the remembered driving time, relative to the estimated time of the crash. The billing records were then examined for those control periods during which a subject remembered driving, for the same time duration that the subject remembered as having driven before the crash. The correction for driving intermittency was hence built into the Australian study design and the RR did not need to be corrected for it after the fact as in the Toronto study.

The results showed that 25 subjects had engaged in a call in at least one of the control windows at 1, 3, or 7 days before the estimated time of the crash. Using a conditional logistic regression method, the Australian study estimated an RR of 4.1 (95% CI 2.2 to 7.7), consistent with the Toronto RR final published estimate of 4.3 (Table 1B).

The current study could not replicate the conditional logistic regression analysis used in the Australian study with the data provided, because it requires knowing, for each driver, whether he or she called or not in the case window and during each control window. In addition, the data for the discordant pairs⁹ were not provided, and these are required for a case-crossover analysis. Therefore, the discordant pairs for each control period were solved for algebraically by using the Talk counts and RR estimates in [17, Table 3].

Table 2 gives the 1-, 3-, and 7-day solutions and the combined solution. It can be seen that these solutions adequately replicated the conditional logistic regression analysis results of McEvoy et al. [17]. It would have been desirable to have the 2x2 cell entries from McEvoy et al. directly, but the marginal totals are the same as in McEvoy et al. [17, Table 3]. In addition, the RR estimates from the solved tables are the same to 2 significant digits for the 1-, 3-, and 7-day estimates in McEvoy et al. [17, Table 3].

Table 2. Australian study Talk counts, no driving.

		Control Window ^b		Total ^c
		Talk	Not Talk	
A. 1 Day Before				
Case Window ^a	Talk	4	22	26
	Not Talk	6	216	222
	Total ^c	10	238	248
	RR (95% CI) ^c	3.7	(1.5 to 9.0)	
B. 3 Days Before				
Case Window ^a	Talk	1	14	15
	Not Talk	3	209	212
	Total ^c	4	223	227
	RR (95% CI) ^c	4.7	(1.3 to 16.2)	
C. 7 Days Before				
Case Window ^a	Talk	5	27	32
	Not Talk	6	288	294
	Total ^c	11	315	326
	RR (95% CI) ^c	4.5	(1.9 to 10.9)	
D. Combined				
Case Window ^a	Talk	20	20	40
	Not Talk	5	411	416
	Total	25	431	456
	RR (95% CI) ^c	4.1	(2.2 to 7.7)	
	Pooled RR	4.2	(2.1 to 8.4)	

^a up to 10-min duration before crash, day of crash

^b up to 10-min duration 1, 3, or 7 days before day of crash

^c McEvoy et al. [17, Table 3]

1.2.1. Unadjusted Biases in the Australian Study

The unadjusted biases in the Australian study were the same as those in the Toronto study, as it neither controlled for part-time driving in control windows, nor did it have objective crash times (and so was also subject to call misclassification bias).

1.3. OnStar Study Review

Young and Schreiner [12] conducted a cohort study of OnStar Hands-Free Calling (HFC) personal calls, which use the same embedded cellular device as OnStar advisor calls.¹⁰ This study employed objective data on crash, call, and driving times, so it could calculate the rate of an airbag crash during an HFC personal call as well as the rate when there was no HFC personal call while driving. A cohort study can also estimate absolute as well as relative risk, and risk differences, while a case-control or case-crossover study can only estimate relative risk. The start and end times of the phone calls for both the

9. That is, if the true RR is above the null value of 1 (i.e., in the causal direction), the RR estimate will be biased low (towards 1). If the true RR is below 1 (i.e., in the preventive direction), the RR estimate will be biased high (towards 1).

10. OnStar advisor calls had been previously studied by Young [18], who found 1 airbag deployment crash per 4 million advisor calls, a rare event.

automatic emergency calls (which went to known numbers reserved for that purpose), and the HFC personal calls were obtained from de-identified cell phone billing records from the cellular provider. The data represent a total of 30 months of driving by 3 million individual OnStar personal calling subscribers in the U.S. and Canada. There were 91 million HFC personal calls made by an average of 323,994 drivers per month. For the period studied, there were 14 crashes severe enough to deploy an airbag¹¹ during the 524.02 person-years¹² of HFC personal conversation ("Talk") while driving. There were 2,023 airbag-deployment crashes during an estimated 47,125 person-years¹³ of driving with no HFC personal conversation ("Not Talk").

Table 3 shows the resultant airbag-deployment crash rates per person-year for Talk and Not Talk. Their ratio is 0.62, indicating that personal conversations using OnStar HFC do not increase the risk of a crash severe enough to deploy an airbag relative to the combined effect of all other activities in normal driving.¹⁴

Table 3. OnStar Hands-Free Calling (HFC) study data and rate ratio analysis (based on data in Young and Schreiner [12, Table III]).

	Talk (in-car)	Not Talk (in-car)	(out-of-car)
Airbag-deployment crashes	14	2,023	0
Person-years	524.02	47,125	762,542
Rate	0.02672	0.04293	0
RR (95% CI)	0.62	(0.34 to 1.05), $p = 0.06$	
PPF%	0.42%		
Rate Difference (95% CI)	-0.0162	(-0.0303 to -0.0021)	

1.3.1. Unadjusted Biases in the OnStar Study

Portable Cell Phone Usage in Unexposed Condition

Braver et al. [19] argued against the validity of the Young and Schreiner [12] OnStar result with two hypotheses: (1) OnStar users may have used portable cell phones while driving when they were not using OnStar, and (2) that such behavior made normal driving periods more risky and hence made OnStar use look relatively less risky (that is, it biased the OnStar RR estimate low). The Braver et al. hypotheses require that portable cell phone usage increases crash risk relative to the OnStar device, and that subscribers to the OnStar device use their portable phones while driving in lieu of the OnStar device.

Young and Schreiner [13] agreed in principle, noting that their original article pointed out [12, p. 187] that embedded hands-free devices have better human factors properties than

portable cell phones (e.g., higher quality in-vehicle speakers and microphone and improved signal-to-noise ratio because of an antenna outside the vehicle). Hence, they hypothesized that embedded devices such as OnStar should reduce crash risk relative to portable devices. Indeed, Young and Schreiner [12] showed that OnStar use does not increase crash risk relative to baseline driving, which includes the combined effect of all forms of inattention (and attention) during the 47,125 person-years of driving in the "Not Talk" baseline periods in the OnStar study (Table 3).

It would be desirable to test the Braver et al. [19] hypotheses with direct data from the OnStar cohort. However, unlike a naturalistic driving study (NDS), the OnStar databases contain no information on other secondary tasks that may have been performed in the vehicle,¹⁵ and so the Braver et al. hypotheses cannot be directly tested with the available OnStar data.

However, Fitch et al. [20] (in a recent NDS of cell phone usage while driving) found direct evidence refuting both Braver et al. [19] hypotheses. Hypothesis 1 is refuted because drivers who were users of one type of cellular device (portable hands-free, portable hand-held, or embedded hands-free) rarely used another. Hypothesis 2 is refuted because conversation on hand-held or hands-free portable cell phones had about the same RR near 1 as did embedded hands-free cellular devices. Therefore, the Braver et al. [19] hypotheses can be rejected based on recent NDS evidence, which indicates that including portable cell phone usage or not in the unexposed condition would lead to negligible change in the RR estimate of 0.62 for personal conversations using the OnStar device.

It might still be conjectured that some secondary tasks (such as manual dialing of a hand-held cell phone) do increase the odds of a crash, and these could have theoretically been present in the OnStar baseline period. However, many secondary tasks reduce the odds of a crash just as others increase them. In fact, the pooled net RR of all secondary tasks as a whole may be near 1, the same as baseline driving without performing those tasks. Indeed, Young [21, Figure 2] found that the pooled RR was near 1 for eight commonly performed secondary tasks¹⁶ in the 100-Car NDS [22], after removing biases in the original 100-Car RR estimates [22].

Demographic Variables Limitation

A valid limitation noted by Young and Schreiner [12, p. 200] about their own study is that, although it was a U.S.-Canada total cohort of 3 million drivers with OnStar HFC, it did not control for possible bias from confounding by demographic variables.¹⁷ Stratifying the data by at least age and sex is required in the analysis of any data collected in a cohort or

11. The severity criterion for the Toronto study was property-damage-only, while for the Australian study it was being seen in a hospital emergency room. However, neither of these two studies indicated how many of the crashes had airbag deployments, nor did the OnStar study have any data on injuries or property damage, so it is not known to what extent the OnStar airbag crash severity level overlapped with the severity levels in other studies.

12. These were converted to person-years from person-minutes in [12, Table III].

13. For comparison, the 47,609 driving-years in the OnStar study [12] is 12.5x the expected 3,800 driving-years of the Strategic Highway Research Program 2 NDS at its scheduled completion date in 2014.

14. The RR CI (0.34 to 1.05) overlapped 1, meaning that the ratio vs. baseline was not statistically significant at the 95% confidence level ($p = 0.06$). However, the risk difference CI was below 0, indicating a protective effect.

15. Young and Schreiner [12] did not have access to portable cell phone billing records and so could not compare embedded cell phone usage to portable cell phone usage.

16. Read, Makeup, Dial, Eat, Reach, Talk, Drink, Radio Tune.

case-control study design in order to control for possible bias from those variables.¹⁸ Bias from confounding by demographic variables can raise or lower crude RR estimates by unknown amounts in any cohort or case-control study that does not control for such bias.

1.4. 100-Car Study Review

Naturalistic driving studies are based on non-experimental real-world driving and include real-time vehicle data (e.g., speed, braking, and accelerator inputs) as well as video recordings of the driver and roadway. They are particularly useful for cell phone studies because they provide direct, objective driving and calling data associated with safety-relevant events for each study subject. This information was available in the OnStar and NDS studies, but not in the Toronto or Australian studies.

Specifically, the 100-Car study [22] examined video data from 241 drivers in about 100 vehicles over a period of one year. It reviewed data on 830 crash and near-crash events, and compared video clips of a driver's behavior at the time of these events to that in randomly-selected video clips from baseline periods driving without a crash, near-crash, or incident (the latter are safety-related events of less severity than crashes and near-crashes).¹⁹ The 100-Car study used a case-control analysis to compare the odds of exposure of drivers to cell phone conversation before a crash/near-crash to the odds of exposure to that task in baseline periods. The crash/near-crash RR estimate was reported as 1.29 for a hand-held cell phone conversation, as shown by the 100-Car data in Table 4.

Table 4. RR estimate by Klauer et al. [22] for hand-held cell phone conversation comparing Talk vs. No Task for 100-Car NDS with unmatched baseline. Data from Hankey [23, Table 6].

	Talk	No Task	Total	Odds	Prevalence
Crash/Near-Crash	44	237	281	0.186	N/A
Unmatched Baseline	1,299	9,059	10,358	0.143	N/A
RR (95% CI)	1.29	(0.93 to 1.80)			

1.4.1. Unadjusted Biases in the 100-Car Study

1. The 100-Car Study used a different driver fault condition for the "Talk" and "No Task" crashes and near-crashes, which biased its RR estimates high [24].²⁰

17. Young and Schreiner [12] recognized this limitation prior to publication but were unable to obtain the de-identified demographic data about OnStar subscribers in time to meet the publication date.

18. Such stratification is not required to control for demographic bias in case-crossover designs, which inherently control for all demographic variables by using each subject as his or her own control; stratification is sometimes used in case-crossover studies [3,17] to look at possible differences between population subgroups, rather than to control for possible confounding factors.

19. Klauer et al [22] included only crashes/near-crashes in their analysis of cases, and not incidents, so incidents were censored from both cases and controls. The number of control clips selected for a particular vehicle was, however, based on the total number of crashes, near-crashes, and incidents for that vehicle, introducing possible selection bias.

20. The 44 cases in the "Talk" group in Table 4 were all crash/near-crash cases where "Talk" occurred, regardless of subject driver fault. The 237 cases in the "No Task" group were restricted to cases with no observed secondary task where the subject driver was at fault, introducing upward bias [21].

2. It used a non-standard "No Task" condition (meaning "no secondary task") rather than the standard "Not Talk" for the unexposed category,²¹ which biased the prevalence and Population Attributable Risk % (PAR%) estimates high;²² for this reason, the prevalence estimates are listed as not available (N/A) in Table 4, since they are incorrect [24].
3. It collapsed crashes and near-crashes into a single group. Although near-crashes are also a safety concern, the RR estimates are uniformly lower for secondary task involvement in near-crashes than for crashes [25].²³ Hence, combining crashes and near-crashes lowers the RR estimate vs. crashes alone, giving the false impression that a task may be safer than it really is. For example, combining crashes and near-crashes into one group severely underestimates the RR for drowsiness as a cause of crashes [24].
4. It did not adjust its RR estimates for demographic variables, even though it presented data showing that crash/near-crash prevalence varied by age and sex. The RR estimate of 1.29 in Table 4 could be higher or lower after adjustment for demographic variables.

1.5. Cell Phone NDS Review

Fitch et al. [20] used instrumented passenger vehicles to record driving behavior for 204 volunteer subjects for an average of 31 days, to investigate their calling, driving, and safety-critical event behavior. It is the first known study to publish data obtained from cell phone billing records for both in-car and out-of-car calls by the same drivers. Of the 204 drivers, 187 provided their cell phone records for analysis, which were matched to driving times as determined by both GPS and video recordings. Subjects were stratified into those who primarily used portable hand-held, portable hands-free, or embedded hands-free cell phones.

Fitch et al.'s [20] study was also the first NDS to use a new method for selecting which video clips to analyze for the baseline comparison data. Fitch et al. [20, p. xxvi] used a 20-s baseline video clip sample "selected 30 s prior to the start of each sampled cell phone interaction time period."²⁴ (All other

21. The "No Task" category does not include other secondary tasks, but does include drowsiness, and so does not represent "attentive."

22. Notice that this particular bias does not affect the RR, only the prevalence and PAR% estimates. The RR is not biased because the pooled RR is near 1 for all secondary tasks combined in the 100-Car study [6], so including secondary tasks or not in the baseline sample makes little difference in the RR estimate.

23. That is, drivers who crashed had higher odds of being exposed to the secondary tasks examined by Guo et al. [25] than those who had near-crashes. This may seem contradictory, because there are 10 times more near-crashes than crashes. However, the exposure odds are the probability of the occurrence of an exposure to that of non-occurrence of the exposure. Thus, the RR is estimated as a ratio of those odds, rather than a count, and as such is independent of the absolute number of crashes or near-crashes.

24. Fitch et al. [20] were concerned that their results might not be representative of the general public, because only drivers who reported talking on a cell phone while driving at least once per day were recruited, which the investigators indicated is a relatively high cell phone use rate. However, their baseline sampling method used each subject as his or her own control, so the absolute rate of cell phone usage per subject (or any other individual difference for that matter), has no effect on the RR estimate.

NDS analysis methods to date have used baselines from previous days.) This idea was first proposed by Maclure and Mittleman [26] in an editorial review of the Toronto study:

Another way to adjust for the intermittency of driving and at least partially for telephone use outside cars is to compare the risk 1 to 5 minutes after a call with the risk 16 to 20 minutes after it. This adjustment also partially corrects for additional confounding factors, such as weather or traffic conditions that may increase both telephone use and the frequency of collisions.

Young [27] was the first to recommend this improved case-crossover baseline selection method for use in NDS data analyses. He pointed out that selecting the baseline video clips from the driving time that was before the case video clips during the same trip on the same day of the crash, controls for driver, vehicle, weather, and traffic conditions, as well as all other factors that are generally constant for a given trip, such as presence or absence of passengers and time of day.²⁵

This closely coupled baseline method follows a case-crossover design in its sampling of baseline video clips. However, Fitch et al. [20] and Klauer et al. [28] used a case-control method to analyze their data, rather than a case-crossover analysis, which considers only the discordant pairs.⁵ This matched baseline method nonetheless retains the advantages of reducing bias from many of the confounding variables listed above. However, had a case-crossover analysis method been used, it would have shown greater sensitivity to small effect sizes than the case-control analysis method used in [20,28].

Table 5 presents the Fitch et al. [20, Table 2] "Talk" data for vehicles moving more than 8 km/h, using this new method for matching baselines. This data estimates an RR of 0.75 for Talk combined across all phones (hand-held, portable hands-free, and embedded hands-free wireless devices), and all safety-critical events (crashes, curb strikes, near-crashes, and crash-relevant conflicts^{26,27}). In Table 5, Fitch et al. [20] compared "Talk" to "No Task," a non-standard epidemiological method that causes errors in prevalence and population percent estimates, as shown by Young [21]. Therefore, prevalence estimates are listed as "N/A" in Table 5.

Table 5. RR estimate for NDS study [20, Table 2] of cell phone conversation comparing Talk vs. No Task with matched baseline.

	Talk ^b	No Task	Total	Odds	Prevalence
Safety-Critical Event ^a	28	154	182	0.1818	N/A
Matched Baseline	259	1,068	1,327	0.2425	N/A
RR (95% CI)	0.75 (0.49 to 1.15)				

^aCrashes, curb strikes, near-crashes, and crash-relevant conflicts.

^bCombined hand-held, portable hands-free, and embedded hands-free

1.5.1. Unadjusted Biases in the Cell Phone NDS

Heterogeneity between Severity Levels

Fitch et al. [20] did not test for heterogeneity between crash-relevant conflicts, near-crashes, and crashes before combining them into one group. If there were heterogeneity, then combining those strata into the single group of "safety-critical events" would be misleading at best or invalid at worst [29].

As a hypothetical example, assume the Fitch et al. [20] data are heterogeneous between the different crash severity levels: e.g., assume a lower RR for near-crashes than crashes as found by Guo et al. [25]. Alternatively, assume crash-relevant conflicts have a different effect size or direction than near-crashes or crashes. When such levels are summed into one large group as in [20], their different effect sizes and directions, which may be of importance and interest for safety issues, would not be seen, leading to invalid conclusions [29]. Thus, heterogeneous subgroups should not be combined [29].

Even if these multiple severity levels were homogeneous, Guo et al. [25] showed that crashes have higher RR estimates than near-crashes for every secondary task examined. Thus combining near-crashes with crashes would lower the RR estimate, and adding crash-relevant conflicts would lower it even further, possibly giving rise to false negative errors.²⁸

False Negative Errors from Low Number of Crashes

Data in [20] were collected for only one month rather than a year or more as in other NDS datasets, and resulted in only 4 crashes. Such a small number of crashes results in insufficient sensitivity (i.e. statistical power) to properly assess the risk effects of cell phone tasks on crashes. Fitch et al. [20] should have clearly stated this limitation, as it could lead to false negative errors.²⁸

This limitation can be overcome if severity levels are homogeneous and can be merged to create a dataset with a larger number of safety-related events, which is a widely adopted strategy in NDS analyses. However, such grouping is commonly done without a test for heterogeneity. Thus, there could be an issue with a risk factor that has an elevated RR for crashes but not near-crashes or lower severity levels. If the

25. This case-crossover baseline video clip sampling method is far less resource-intensive than the Klauer et al. [28] case-crossover method that attempts to match baseline video clips on previous days to case video clips by driver, time of day, GPS location, number of lanes, city vs. highway driving, traffic conditions, weekday/weekend, etc. It is a difficult task, and impossible for a trip that was never repeated.

26. The "intersection" event data, although novel for an NDS, were "too noisy" and Fitch et al. did not include them in many of their RR estimates.

27. Note that Fitch et al. [20] included these "conflicts" of lower severity than near-crashes and crashes in their RR estimates, whereas the 100-Car study [22] did not, with an unknown effect on heterogeneity as described in the next section.

28. A false negative error means that the crash risk for a cell phone task is estimated to be no different from baseline risk when in fact it is higher than the baseline risk. That is, saying a task is benign when it actually increases crash risk.

crash events are simply merged with the lower severity events, the crash effect could be “washed out” by the larger number of events of low-severity, leading to false-negative errors.

Other Biases in the Cell Phone NDS [20]

1. It did not test for heterogeneity between cell phone subtasks (reaching, dialing, talk, etc.) before grouping those subtasks into the larger category of “cell phone use” for each cell phone type.
2. It used “No Task” in its odds ratio estimates, which biases prevalence and population attributable risk estimates upwards [21].
3. It presented results for a rate ratio analysis that correctly used the complement of the exposure (e.g., “Not Talk” instead of “No Task”), but from the information given in the report, it was not possible to replicate those rate ratios [20, Table 22]. There is concern of an unknown bias or error in the study's rate ratio estimates because after recalculating the odds ratio estimates to use “Not Talk” (e.g. Section 2.7), they did not agree with the rate ratios [20, Table 22], as they should have.¹
4. Placing the control window only 30 s in advance of the case window may cause the Talk counts in the control window to be upwardly biased, because of the natural tendency of people to make several calls in a row. This tendency could bias the RR low if such cases had been included.
5. Similarly, there are multiple subtasks involved before a Talk subtask begins, which could “spill over” into a baseline that was too close in time to the Talk subtask. These include possible cognitive processes involved in planning the call (e.g. retrieving the phone number from memory), the possible eyes-off-road time involved in searching for a portable phone, looking up a phone number in an address book, and dialing by any method. Although some of these subtasks were separately analyzed in the report, it is not clear from the methods what baselines were used each subtask, and how these baseline comparisons were made, given that subtasks may have been performed in close succession. It is not clear that a 30 s gap would prevent these preparatory tasks from being included in the baseline for the Talk portion of the call, thus biasing the RR low.

1.6. Epidemiological Study Heterogeneity

Table 6 summarizes the original RR estimates for Talk while driving. Studies 1 and 2 estimate an RR near 4, while studies 3-5 estimate an RR near 1. Attempting to pool across these data using meta-analysis²⁹ shows that they are heterogeneous at $P < 2E-11$. The heterogeneity is caused by the inconsistency between the RR of 4 estimated by the first 2 studies, and the RR near 1 in the more recent 3 studies. Therefore, the pooled result is invalid and is shown as N/A.

29. Calculated with the meta-analysis worksheet in Episheet [34].

Table 6. Original RR estimates as published for cell phone conversation while driving for 5 epidemiological studies and pooled meta-analysis estimate.

#	Study	Metric	Design	RR	95% CI	
					Low	Upper
1	Toronto [3]	Risk Ratio	Case-Crossover	4.3	3.0	6.5
2	Australia [17]	Risk Ratio	Case-Crossover	4.1	2.2	7.7
3	OnStar [12]	Rate Ratio	Cohort	0.62	0.37	1.05
4	100-Car [22]	Odds Ratio	Case-Control	1.29	0.93	1.80
5	Cell Phone [20]	Odds Ratio	Case-Crossover	0.75	0.49	1.15
Pooled		Relative Risk		N/A†		
P homogeneity				2E-11		

†heterogeneous

PART 2. NEW ANALYSES OF CALL-CRASH DATA FROM PRIOR EPIDEMIOLOGICAL STUDIES

2.1. Objective

The aim of Part 2 is to calculate unbiased estimates of the RR for cell phone conversation for automobile crashes. In lieu of conducting a new NDS, which is expensive and time-consuming, it is more efficient to eliminate as much bias as possible from previous estimates through improved analyses.

Evidence of bias in one or more of the studies in Table 6 is apparent. The Toronto and Australian studies found an RR near 4, while the OnStar, 100-Car, and Fitch et al. [20] studies found an RR near 1, a major discrepancy.¹ A minor discrepancy is that the Klauer et al. [22] 100-Car study estimated a Talk RR above 1 (1.29), in a causal direction, while the Young and Schreiner [12] OnStar study and the Fitch et al. [20] cell phone study estimated Talk RRs below 1, in a preventive³⁰ direction.

The current study therefore investigates biases at the root cause of these discrepancies, and adjusts the estimates in Table 6 in order to arrive at unbiased RR estimates of cell phone conversation for automobile crashes. Part 2 further refines previous adjustments in the literature [5, 6, 7, 8, 9, 10, 29].

2.2. Methods to Estimate and Adjust Bias in Call-Crash Studies

Adjustments and corrections were made to 4 of the 5 RR estimates in Part 1, in an effort to remove the remaining biases that were identified regarding their estimates of the cell phone conversation RR while driving. These adjustment methods have been previously described [5,7, 8, 9, 10,21,29]. The

30. The term *protective* could be used here as a synonym for *preventive*, but *preventive* is used here as the correct technical term in epidemiology, namely, *Prevented Fraction (Population)*, or *Population Prevented Fraction % (PPF%)*. In a situation in which exposure to a given factor is believed to protect against an outcome (such as a crash), the *prevented fraction* is “the proportion of the total load” of the crashes in the population that has been prevented by exposure to the factor. Porta [1, p. 192] cautions that, “this measure must be interpreted with caution, as part or all of the apparent protective effect may be due to other factors associated with the apparent protective factor.” [1, p. 192].