Contents

1. **The Study Context Determined the Study Methodology**

a. **A Randomized Controlled Trial is Not Ethically Feasible in this Industry Context**

As noted in the main text, a randomized prospective controlled trial of how commercial driver crash risk varies with OSA-related status is neither ethically nor legally feasible. A gold-standard treatment, provision of an auto-adjusting positive airway pressure (APAP) mechanism, is available and clinically well established.[1] It is therefore not ethically appropriate to randomly assign some drivers diagnosed with OSA to receive no treatment, out of concern for the rights of drivers as OSA patients. An important additional consideration is that if untreated OSA increases the crash risk among commercial drivers as it does in the general driver population, deliberately putting commercial drivers who have untreated OSA who have not been offered treatment in command of tractor-trailers with an 80,000 lb. gross vehicle weight on the public highways would put both the truck drivers and the motorists sharing the highway with them at higher risk of death or serious injury from an OSA-induced large truck crash.

b. **Retrospective Analysis of an Existing Employer-based Program is Appropriate**

Because data on OSA programs among commercial drivers is extremely rare, and due to the potential public health and safety policy importance of the question of whether OSA programs are successful, it was appropriate to perform a retrospective observational study of the employer-based program analyzed herein.

The data elements available for the study are human resource records and operational records of the job performance of the study firm's employee drivers, including such items as demographic characteristics and week-by-week observations of miles driven and trip segments completed. The operational records also include data on each crash event recorded by the firm, including whether or not the crash was Department of Transportation-reportable (i.e. was it serious) and whether or not the crash was preventable by actions of the driver (see Section 2.b). OSA screening, diagnosis, and treatment information was compiled for each relevant driver by Precision Pulmonary Diagnosistics (PPD) in the course of the provision of OSA-related services to the study carrier. The Truckers & Turnover Project research team at the University of Minnesota, Morris, integrated data from these sources into combined files that were cleaned, formatted, and analyzed.

It should be noted that the study firm's OSA program was not designed as a scientific test of the value of OSA screening, diagnosis, and treatment, but for application under the existing business conditions. In the next several sections these business conditions are examined, and the methodological implications are presented.

c. Background: Turnover Rates in this Part of the Industry Are High

The study firm operates in a part of the for-hire trucking industry (the long distance "full truckload" (TL) segment) in which driver turnover rates are high [2,3] The American Trucking Associations surveys its member TL firms quarterly on their turnover rates, and during the study period reported numbers generally in the range of 100% per year.[3,4] Based on exits from the reference population, a similar analysis estimates the annual turnover at the study firm to range from 34% to 76% over the course of the study. (See also Figure S1 Panel A, which shows the hazard of separating from the firm by week of job tenure.) Thus, a large fraction of the firm's workforce turns over each year. This affects the study design because, as discussed next, crash rates for commercial drivers decrease with tenure.

d. Background: Crash Risk is Lower with Experience at Hire and Falls with Job Tenure

Crash rates for commercial drivers generally decrease as their experience increases.[5,6] Drivers who are experienced at the time of hire have lower initial crash rates than drivers who are inexperienced at the time of hire. And, for all drivers, crash rates drop with job tenure, as we see in Figure S1 Panel B, which shows the hazard of having a preventable crash by the week of job tenure. The drop in crash risk with tenure is due to two factors.

i. Drivers Learn with Experience

One is that drivers become safer with more experience. This is especially true for drivers who are inexperienced at hire and learning the basics of their profession. But it is true to a lesser extent for experienced drivers who are new to the study firm, as they are gaining familiarity with the tractor which has been assigned to them and to the driving conditions associated with their new home base.

ii. Unsafe Drivers Are Discharged, Creating "Safety Selection"

A second reason that crash risk drops with job tenure is a process of "safety selection." The study firm has an active safety management program. It monitors the performance of its drivers using multiple metrics. One of the important ones is the preventable crash record accumulated by the driver. (See the discussion in Section 2.b of how "preventability" is determined.) As a result, drivers who accumulate an unacceptable preventable crash history while on the job are either discharged, or in some cases, quit because they anticipate the likelihood of being discharged.

How strong is the safety selection effect? A multivariate Cox proportional hazards model of discharges was created, using a similar format to the Andersen-Gill multivariate crash risk model described in Section 2.a.i. This model was run on data consisting of weekly observations on all the drivers in the reference population (more than 41,000 subjects observed on more than 2,750,000 driver-weeks). It controlled for all the variables, such as those identifying demographic characteristics, work type, and miles in each week, which are also in the Andersen-Gill crash-risk model. In this model the baseline hazard of being discharged in a specific week, given that one was employed up to that week, is estimated to be raised by approximately 30-fold if the driver has had a DOT-reportable preventable crash during the prior or current week (HR=29.94, 95% CI: 26.32, 34.05; details available from the authors upon request).

e. **Diagnosed Drivers have More Tenure than the Reference Population**

As established above, the outcome variable of the study is preventable crashes, and the risk of these varies with job tenure. Thus if there is any time delay after being hired involved in being screened and in being diagnosed, job tenure at diagnosis has the potential to affect who gets a diagnostic PSG test for OSA and who does not. A greater degree of safety selection associated with higher job tenure at diagnosis will increase the safety of those who are diagnosed when there is some lag between being hired and being diagnosed, because many potentially relevant drivers (those with undiagnosed OSA who had an unacceptable preventable crash record) will not be present in the workforce to be diagnosed

Table S1 shows that this possibility is likely to be realized in the study data, because the tenure of drivers who are ever diagnosed follows a different job tenure distribution on the date of diagnosis than does the tenure of drivers in the reference population. Specifically, at every point in the job tenure distribution, the tenure with the study firm of drivers who are diagnosed is greater than that of the population as a whole.

The tenure distribution of the reference population reflects the fact established in Section 1.c, that turnover rates at the study firm, while lower than the average levels recorded by the American Trucking Associations for carriers of this type,[3,4,7] are nonetheless quite significant. At study midpoint, twenty-five percent of the workforce had been on the job with the study firm less than eight months (31.7 weeks of tenure / 4.3 weeks per month = 7.4 months). This reflects the rate of turnover documented above in Section 1.c.

This difference in tenure between the diagnosed group and the reference population is driven in part by the fact that the firm covers the cost of diagnosis and, if required, APAP treatment, as preventive medicine under its voluntary employee medical insurance program, which does not become effective until some time period after the date of hire. The study firm permits new-to-the-industry employee drivers to join the program after 90 days of tenure, and there is also a time lag to qualify for experienced hires (usually 30 days). In addition, since the study firm's OSA program takes medical care in this area beyond the current accepted medical standard of care (since, as noted in the main text there is currently no mandate for screening and diagnosis of commercial drivers in the absence of self-referral), the study firm developed its program over the study period, while it broadened its scope from the initial pilot work in 2005 (Table S2).

f. **Implication: Due to the Tenure-Safety Relationship Construct Retrospective Cohorts by Matching Cases to Controls on Experience at Hire and Job Tenure at the Case's Diagnosis Date**

An implication of the evidence presented in the prior sections is that—if one compares the diagnosed driver group with the reference population as a whole—not only will OSA status be different across the two groups, but also the underlying crash risk associated with learning and safety selection will be different. Directly comparing diagnosed drivers with the entire non-diagnosed driver population thus offers no possibility of drawing statistically sound conclusions about the relationship between OSA status and crash risk.

Instead, the appropriate course is to match cases (drivers who are diagnosed to have OSA) with controls who will have the same underlying or background crash risk profile. Based on the evidence presented in the prior sections, the most important factors affecting the underlying crash risk are the experience level at the time of hire and the amount of time on the job at the study firm, or job tenure (see the discussion above and Figure S1).

Doing this creates what is in effect a retrospective cohort study—a comparison of crash outcomes for individuals with and without the disease—through a case-control matching process. The matching creates a cohort of controls that are unlikely to have OSA to compare to the cases, who are those with diagnosed OSA. It is important to understand that the usual relationship between cohorts is modified here because the outcome variable is not whether or not a subject contracts OSA, but rather whether a subject has a serious preventable heavy truck crash. OSA is then one of many risk factors (including particularly varying degrees of exposure on the road) which can affect the likelihood of this outcome.

Thus, as described in the main text, each driver diagnosed positive for OSA is matched with a driver drawn randomly (with replacement) from the group of drivers who were screened as being at low priority for an OSA diagnosis, under the constraint that the selected control driver should have the same experience level when hired, and the same job tenure (plus or minus a week) as the case, during the week of the case's PSG diagnostic test. The corresponding date for the control driver is denoted the "matching date."

g. **Comparing Cases to Controls *before* the Diagnosis/Matching Date is Not Useful in these Data Due to Safety Selection**

The comparison between cases and their matched controls during the period before the PSG/matching date would theoretically provide a retrospective comparison of two groups that have similar background crash risk levels, but differ only in whether or not they have untreated obstructive sleep apnea. However, as documented above in Section 1.d.ii, the problem with making this comparison is safety selection. Drivers who had unacceptable preventable crash histories during the "before" period were discharged, and therefore, generally did not get diagnosed with OSA, and thus, they are not present in the study population. This washes out the difference in serious preventable crash risk we would have otherwise expected to observe between cases and controls in the "before" period. A statistical analysis of the "before" data confirms this fact (details available from the authors upon request), and further evidence may be observed in lines 1-5 of the Andersen-Gill model results in Table S3.

h. **Comparing Cases to Controls *after* the Diagnosis/Matching Date Is the Best Feasible Methodology**

Given that the comparison between cases and controls before the PSG/matching date is not informative, the next step is to examine the crash risk of cases and controls in the period after the PSG/matching date. This is analogous, in a retrospective setting, to a clinical trial which measures differences in outcomes after some intervention. In this case, the "intervention" for OSA, APAP, was provided after the date of the diagnostic PSG.

Even though drivers who are *ex ant*e riskier (either due to untreated OSA or any other reason) are now a smaller proportion of the study groups (i.e. of both cases and controls) than they were of the pre-safety-selection driver population, we can still observe the full range of exposures and all the crash outcomes for the drivers who are in the study during the period after the PSG/matching date.

There is, however, a limitation in that treatment was not assigned randomly. Since APAP is provided and APAP treatment adherence is mandated after a positive OSA diagnosis, drivers with OSA self-select into a specific level of treatment adherence during the period after PSG/matching date. The level of treatment adherence can be calculated from the data as Full Adherence, Partial Adherence, and No Adherence (see main text), and the crash risk in the "after" period of each group so-identified can be compared to that of its control drivers.

However, if the No Adherence drivers have a higher crash risk than their controls (as is reported in the main text), it is not possible to conclude that this difference was caused by the effect of untreated OSA. A higher crash rate for the No Adherence drivers is consistent with there being such an effect. But other factors affecting crash risk may also vary in a non-random manner between the No Adherence drivers and other drivers, since these groups are self-selected.

Data is available on some of these potential differences. It is thus possible to control for all observable differences (i.e. differences recorded in the rich data available about drivers' demographic characteristics and work lives) using a multivariate model (as is done in the following section). However, the self-selection process means that unobserved characteristics could also differ between No Adherence drivers, Full Adherence drivers, and controls.

Specifically, it is a reasonable conjecture that drivers who are informed that they have OSA and that treatment adherence is mandated, but who never in fact achieve any degree of adherence, also have a tendency to violate other safety rules. Thus the higher crash risk observed for No Adherence drivers could be due to the clinical effects of untreated OSA and it could also be due to unmeasured characteristics of this group, such as their tendency to ignore safety rules. There is no way with the available data to assign relative importance to these causal factors.

However, given the discussion in the preceding sections, it is clear that the evidence and conclusions presented in the main text represent results of the best analysis that can be performed with the first data ever available on a large-scale employer-based and employer-mandated OSA program for commercial vehicle drivers. It is also clear that our results support the efficacy of the program in removing less safe drivers from the company's driver pool and therefore, increasing road safety for the employer.

2. Robustness Checks on the Main Results

a. Controlling for Potential Confounding Factors with a Multivariate Analysis

The comparison of crash rates across study sub-groups provides the primary evidence presented in the paper. The strength of this approach is that it utilizes the case-control matching procedure to make cases and controls similar in underlying or background crash risk, and accounts for the most important exposure factor in the risk of crashes, miles driven, but requires no statistical assumptions about the distributions of the variables involved or the appropriateness of model specifications. However, a corresponding limitation is that study sub-groups defined according to treatment adherence are not randomly assigned, so it is possible that crash-risk-relevant exposure covariates are not evenly distributed across the sub-groups. Table 1 in the main text, on the demographics and operational characteristics of the study sub-groups, provides suggestive evidence that this may be the case.

Therefore, as a robustness check we use a multivariate technique to examine the association of study sub-group status with crash risk, controlling for a number of factors that may be associated with crash risk and which may vary across drivers, but which may not be evenly distributed across study sub-groups. These factors broadly fall into two categories, demographic and other personal characteristics (such as sex, ethnic/racial identity, or experience level at hire), and operational exposure covariates (such as miles operated, number of trip segments, or job type).

i. Factors that Could Affect Crash Risk

The set of control variables included in our multivariate model are next described. Personal characteristics that may be associated with variation in crash risk include:

- Sex: One binary indicator variable was defined to describe driver sex as "Female", the base (omitted) category is "Male or missing data".
- Age at PSG/matching: Age at PSG (cases) or at comparison date (controls) was specified as set of indicator variables. The age ranges are "21 to 40", and "51 and higher"; the range "41 to 50" is the base (omitted) category. 21 is the minimum age for a commercial driver's license. Three discrete categories are used to allow for non-linear effects without the complexity of a polynomial age specification.
- Race: Both race and ethnicity are captured in the model in the form of indicator variables for "African-American", and "Other." Other drivers are those not African-American nor White, which included drivers with missing race data. The base (omitted) category is "White".
- Experience at Hire: Experience level when the driver was hired is captured as two indicator variables. "Inexperienced at Hire" was included in the model, and was defined to mean that the driver was determined to be inexperienced by the firm during at least one spell of employment that is observed. The base (omitted) category is therefore "Experienced at Hire", meaning the driver had worked at the study firm or another trucking firm prior to being hired (on all spells of employment if more than one was observed). The criterion to designate a driver as "experienced at hire" was the determination by the study firm that the driver was not required to go through the firm's basic driver training program, which was required of all drivers not meeting the firm's standards for prior experience.

The control variables that capture differences in the exposure to crash risk associated with different operational conditions include:

- Trip Segments per Week: This varies week-by-week and indicates how many times the driver starts and stops trips, which is frequently associated with changing trailers, maneuvering in tight spaces, and travelling in urban areas, each of which could be associated with variations in crash risk. This information is captured as a set of indicators for 0 – 5 segments per week and more than 11 segments per week, with 6-10 segments in a week as the base (omitted) category. Three discrete categories are used to allow for non-linear effects without the complexity of a polynomial trip segments specification.
- Miles per Week: This varies week-by-week, and is an important measure of risk exposure. It is specified as a set of indicator variables for different levels of miles per

week, with 1,500 to 2,500 miles (the most common value) as the base (omitted) category. Three discrete categories are used to allow for non-linear effects without the complexity of a polynomial miles per week specification.

- Job Type: The type of job performed by a driver may vary from week to week, though it is generally stable. A set of indicator variables identify the following specific job types: drivers assigned to the service of one (usually quite large) customer, called "dedicated" drivers; long-haul drivers who have other kinds of specialized work, referred to as "other". The base (omitted) category, is "solo system" drivers, who work alone on long-haul trips with random routes; prior work has shown crash risk to vary by job type in this population.[8]
- Season: Since week-by-week data on individual drivers is used and drivers are observed in the study firm's workforce at different calendar dates, control variables are added for general background changes in the risk of a crash associated with changes in the seasons. These are captured as spring, summer, and fall, with winter as the base (omitted) category.
- Calendar Time: There may also be variations in crash risk over time with changes in economic activity. These are captured with time indicators for a) the 1st half of 2005 through 2nd half of 2006 (i.e. all of 2005 and 2006), b) all of 2008, and c) all of 2009. The base (omitted) category is d) 2007. It should be noted that 2005 and 2006 are grouped together because the study firm's OSA program started during the final three quarters of 2006, so there is little variation in crash outcomes for drivers who are cases or controls until later in the calendar period covered by the data (recall that the study period is Jan, 2005 through December, 2009).

ii. Andersen-Gill Time-to-Crash Multivariate Model

The Andersen-Gill time-to-crash multivariate model is a generalization of the perhaps more familiar Cox proportional hazards time-to-event model.[9] The analysis time is job tenure, or weeks employed by the study firm. In this approach the time until a crash (job tenure in weeks) is assumed to have a hazard, or risk function, with two components: an underlying risk function (the instantaneous probability of a crash event—the "hazard"—which varies with analysis time) that is shared by all drivers, and a risk adjustment function that depends on the characteristics of individual drivers and their operational settings week by week. This is a particularly useful approach in the context of CMV driver data, in which it is known that the risk of a crash is initially high and then declines with job tenure (see Section 1.d).

Three amendments to this basic approach are employed in our robustness check. First, the model is stratified by the experience level of the driver at the time of hire because inexperienced-at-hire drivers have a higher initial crash risk than do ones who are experienced at hire (see Figure S1 Panel B). In other words, the model permits these two distinct driver groups to have different underlying risk processes (baseline hazards) over time. Second, a version of the model is used that applies to week-by-week data on each individual driver (variously known as panel data or repeated measures data), which allows the value of independent variables (e.g. miles per week or job type) to vary for a given driver week-by-week. Third, as mentioned above, the Andersen-Gill variant of the Cox model used allows each subject to experience more than one event (crash) in their event history.[9,10]

The strength of this model is its sophisticated accommodation of variations in exposure to risk across both drivers and weeks, and the fact that it lets the data determine the baseline hazard function(s) without distributional assumptions. Its limitation is that it imposes the assumption that changes in independent variables are associated with proportional shifts in the hazard functions. In order to check the validity of the specification, statistical tests of the proportional shift assumption were made and appropriate residual plots were examined. The proportional shift assumption is not rejected for any of the key study or control variables, nor for the model as a whole. To confirm these test results, the relevant residual plots were examined. Both the test results and residual analysis lead to the conclusion that the implementation of the Andersen-Gill approach used is an appropriate complement to the crash rates approach.

iii. Statistical Details: Results Derivation in this Model for the Period *after* the PSG/Matching Date using Interaction Terms

The treatment group effects in our Andersen-Gill model are specified as a series of indicators for each treatment adherence level relative to an implied comparison group of controls. To distinguish between weekly observations that are after versus before the PSG/matching date, an indicator variable is created to indicate observations of driver work-weeks that take place after the PSG/matching date for that driver. To allow crash risk differences after the PSG/matching date, as compared to that before, we create interaction terms between the OSA adherence levels and the After- vs Before-PSG/matching date indicator variable. The estimates of these terms are presented in lines 1 to 10 of Table S3. (Interaction terms are denoted with an “x” between the terms being interacted, in lines 7-10.)

The analysis of crash risk differences after the PSG/matching date presented in Figure 1 Panel B of the main text are the results of *tests of the statistical significance of linear combinations* of the parameter estimates which are displayed in hazard ratio form in the lines of Table S3. The model was specified with time indicator variables that allow the crash risk relationships in the period before the PSG/matching date to be read directly from lines 1-4 of Table S3. However, the crash risk comparisons after the PSG/matching date exhibited in Figure 1 of the main text require linear combinations of several of the values on separate rows in Table S3 to be formed.

For example, to compare the crash risk for the No Adherence sub-group versus the controls during the “after” period in a hazard ratio, the components of risk for a driver in the No Adherence group are collected in the numerator (OSA Never Adherent (Before PSG) + After vs Before PSG + OSA Never Adherent x After vs Before PSG) and the risk components for a Control driver are collected in the denominator (After vs Before PSG). The (After- vs Before-PSG) component is common to both groups and therefore cancels, producing the hazard ratio: $\exp(\log(1.560) + \log(2.430)) = 3.791$ as presented in Figure 1 Panel B in the main text. (Note that this hazard ratio is statistically significant, even though some of its component parts are not when considered individually.)

The other crash risk comparisons are produced similarly, with p-values and 95% confidence intervals generated by Stata. As Figure 1 Panel B in the main text shows, we find No Adherence drivers have estimated hazard ratios that exceed 1.0 with p-values less than 0.05 compared to all other adherence levels. No other adherence groups have significantly different risks from each other.

b. **Why "Preventable" Crashes are the Best Available Outcome Measure**

One may think of every commercial vehicle crash event as being the combined result of two factors, the crash risk exposure and the crash-relevant behavioral performance of the driver. There is a legal standard for the contribution of driver behavior to the causation of a crash event in the US system of legal liability—being "at fault." Crashes in which a commercial driver is legally at fault are practically hard to identify, unless there has been a determination by law enforcement, or even more indicative, by the courts, both of which are statistically much rarer events than are serious truck crashes.

More specifically, a recent study by the US Federal Motor Carrier Safety Administration (FMCSA) showed that it is very difficult to identify crashes in which the driver's behavior was a large causal component (i.e. was the "critical reason" in the lexicon of the report) from government data, because the police accident reporting system (PARS) does not contain sufficient accurate and relevant information about enough crashes to make this statistically feasible.[11] The study was commissioned because of a debate about whether crashes in which poor CMV driver behavior was not a cause could be removed from the FMCSA safety rating system for motor carriers, where their listing may make it appear a carrier has unsafe drivers when this is not the case. However, the only crash data presently available for the FMCSA's rating system is data generated by PARS. Since the FMCSA's study showed that PARS-based data does not permit CMV driver behavioral responsibility to be identified, all crashes recorded via PARS are used in the FMCSA's safety rating system, whether or not there is evidence of poor commercial driver behavior.

However, the present study does not use PARS-based data, but instead uses administrative data internal to the study firm. While a disadvantage of this approach is that we cannot follow drivers when they depart from the study firm, there are several important advantages. First, the crash count is more accurate, due to the same limitations in the reporting process across the thousands of state and local legal jurisdictions who originate PARS-based data that the FMCSA study found make it impossible to identify poor commercial driver behavior in that data. For instance, the study firm compared its internal record of DOT-reportable crashes to that collected by FMCSA via PARS for 2013, and found the internal records recorded approximately 15% more US DOT-reportable crashes than did the federal data. Second, the internal administrative data on each driver and on each crash are much richer than the available Federal data, as described in the main text and in sections above. Specifically, many factors that could potentially confound the primary results can be controlled for in a multivariate model to check robustness when internal administrative crash data is used.

In addition, each crash is categorized as "preventable" or "not preventable" in internal administrative crash data. All firms in the industry standardly categorize all crashes in which poor driver behavior was an important causal factor as "preventable," which roughly means that the commercial driver could have and should have taken actions that would have prevented the crash (whether or not any judgment was made as to who was "at fault" in the crash event), or alternatively, as "not preventable." Preventability is determined using long-established guidelines promulgated by the National Safety Council.[12]

Carriers have an economic incentive to get this categorization right for two reasons. One is that these determinations are subject to review by the FMCSA during motor carrier safety audits (and very large carriers such as the study firm can expect to be audited periodically). A second is that disciplinary actions against employees, including discharges for cause, are governed by employment law, and firms prefer to have reasons that are legally defensible when they discipline or discharge drivers with unacceptable crash records. Using an unacceptable preventable crash record, when that categorization is based on regulatory and industry practice, is more legally defensible than referring to all crashes, since "all crashes" would include those in which the driver may not have behaved poorly.

Further, motor carriers face the potential for wrongful death or injury lawsuits due to crashes involving trucks operated by the carrier, and the liability cost is likely to be smaller if the carrier can show it disciplines and remediates its workforce based on preventable crash histories, independently of whether the current crash was judged preventable or not. Indeed, in the study firm the categorization of crashes is done by individuals in a captive liability insurance company, whose personnel are not in the reporting structure of the operational managers, who could conceivably have biased perceptions of driver behavior. One may think of using the study firm's crash data, including the preventability designation, as analogous to the use of nosocomial infection rates from a hospital's own infection control committee, rather than relying on data from a third party, such as a state health department.

There is clear statistical evidence that this is appropriate. A Cox proportional hazard model of discharges run on the reference population (more than 41,000 subjects observed on more than 2,750,000 driver-weeks) was described above in Section 1.d.ii, in which "safety selection" affecting who entered the study subject pool was analyzed. It was reported above that a driver who had a DOT-reportable preventable crash during the prior or current week had approximately a 30-fold increase in the baseline hazard of being fired (HR=29.94, 95% CI: 26.32, 34.05; $p<.0001$). That same model also contained an independent variable identifying weeks in which the driver had experienced a DOT-reportable crash during the present or prior week which was judged <u>not</u> preventable. This was estimated to raise the risk of discharge by less than two-fold (HR=1.68; 95% CI: 1.15, 2.47; $p<.01$). The baseline risk of discharge varies with tenure in this model, but is generally modest in absolute size (recall that only 25% of all exits are discharges). Thus, the stark difference between a 68% increase in the baseline hazard for a DOT-reportable crash that is <u>not</u> preventable, and a 2,900% increase in the baseline hazard for one that <u>is</u> preventable, shows the study firm actually uses the "preventable" designation in the manner intended by the FMCSA. One need not believe this categorization is always made perfectly to observe that using preventable crashes as an outcome variable will provide results that have clearer relevance for safety and health policy.

c. Neither Small Ns nor the Preventability Designation Spuriously Drives Results: Considering "All DOT-reportable Crashes"

Most other studies of crash risk do not make use of the preventable/non-preventable distinction, because it is not available in data on the general driver population. Further, many non-OSA-related studies of commercial drivers use governmental records, which also do not contain a preventability categorization. Therefore, for comparability to other studies that do not use preventable crashes, and to establish that this designation does not spuriously drive the results, in this section we consider <u>all</u> DOT-reportable crashes, irrespective of their preventability designation by the study carrier, as a dependent variable.

Table S4 presents the hazard ratios estimated from the Andersen-Gill multivariate time-to-crash model presented in Section 2.a.ii when that model is re-run on the same data using all DOT-reportable crashes as the outcome variable, instead of just the preventable ones. Several points should be noted.

First, the event count has more than doubled for all study sub-groups (and the exposure levels in weeks and miles presented are the same as those in the main text, Table 2). This reduces the chance that the model is fitting noise in the sample as opposed to finding a valid statistical difference across the study sub-groups. Second, the qualitative pattern of the results is the same as those presented in the main text when the outcome variable is preventable DOT-reportable crashes: while the point estimates suggest a somewhat increased risk for all sub-groups when compared to Controls, Full Adherence drivers have the smallest estimated risk, and only the No Adherence drivers are statistically different from Controls (and the p-value is low). Third, the No Adherence drivers have a differentially smaller increase in the number of crash events when adding in the non-preventable crashes than do the other study groups, but this is to be expected because—given their higher preventable crash rate—they are exiting (through discharge or quit-in-lieu-of-discharge because of safety behavior) more rapidly, which affects their likelihood of exposure to a non-preventable crash. Fourth, in this model No Adherence drivers still have more than twice the risk of controls, which is a large enough difference to be practically important, shown in a manner consistent with studies that do not rely upon a preventability categorization. Taken together, this evidence suggests that the results found using preventable crashes as an outcome variable are qualitatively robust.

d. Using a Higher Diagnosis Threshold: Considering AHI≥15 as the Criterion for Positive Diagnosis

In this section the sensitivity of results is checked to using a more stringent AHI value for the definition of a positive clinical diagnosis of OSA than was the case in the actual study protocol. The hazard ratio (HR) values, 95% CIs, and p-values presented in Table S5 are generated from the same Andersen-Gill model presented in Section 2.a.ii, and reprised in the section immediately above, except that the criterion of AHI $\geq$ 15 is used as the definition of OSA (and as in the main text, only preventable DOT-reportable crashes are considered).

Several points should be noted. First, the qualitative pattern of the results is the same as those presented in the main text when the clinical criterion for a positive OSA diagnosis is AHI $\geq$ 5: while the point estimates suggest a somewhat increased risk for all sub-groups when compared to Controls, Full Adherence drivers have the smallest estimated risk, and only the No Adherence drivers are statistically different from Controls (and the p-value is low). Second, the crash risk point estimate for No Adherence drivers is larger (4.54) than the risk estimated in the same model when the criterion was AHI $\geq$ 5 (3.79), although the confidence intervals overlap. An increase is consistent with the expectation that drivers with OSA that is clinically more severe will experience greater degradation of on-the-job performance and thus greater crash risk. The fact that Partial Adherence drivers are estimated to have a somewhat elevated crash risk which is not far from being statistically different from Controls (HR = 1.78, 95% CI: 0.933,3.239;

p=.082) is also consistent with the clinical expectation that partial adherence would produce partial remission of symptoms.

As was observed using the AHI ≥ 5 threshold, the No Adherence drivers have significantly higher risk than all other sub-groups, and no other sub-groups are significantly different from their matched controls. Finally, while the results of this change in criterion are consistent with the interpretation of all the results that has been offered in the main text, the crash counts are lower here, so results with the OSA diagnosis criterion that was actually used (AHI≥ 5) are presented as the main results of the paper.

3. **Results related to Drivers Diagnosed as Not Having OSA ("Negatives")**

   a. **Results for Drivers Screened as Likely to Have OSA whose Diagnostic PSG was Negative Compared to Matched Controls**

When we consider the 403 subjects diagnosed as "negative" (AHI<5) we find a preventable DOT-reportable crash rate of 0.019 per 100,000 miles, which forms an incidence rate ratio (IRR) of 1.34 compared to matched controls. However, the p-value for this estimate is 0.458 (95% CI: 0.533, 2.948). We find a similar result when we examine the hazard ratio estimated for "negative" drivers in the Anderson-Gill multivariate time-to-event model used as a robustness check (HR=1.34, p=0.461; 95% CI: 0.614, 2.931). In addition, Table S4, which records Andersen-Gill hazard ratio estimates for all DOT-reportable crashes (as opposed to only the preventable ones), and Table S5, which records the Andersen-Gill hazard ratio estimates for DOT-preventable crashes when the diagnostic criterion for having OSA is AHI≥15, both include the Negative study group and both show a similar pattern. Thus, despite point estimates suggesting a small increase in crash risk, there is no statistically significant evidence that drivers screened as at High Priority for OSA diagnosis, but who did not have OSA (Negatives), have higher crash rates than do the controls after the PSG/matching date.

   b. **Overall Results when Study Sub-groups are Changed by Including among Controls Drivers Screened as Likely to Have OSA whose Diagnostic PSG was Negative**

As a further robustness check, in this section the construction of the study groups described in the main text is re-done, but this time including drivers sent for a PSG whose diagnosis was negative (AHI<5) in the set of drivers from which controls for cases may be drawn. The original pool from which control drivers were drawn was drivers that were designated as "low priority", and who are thus likely to be relatively healthy. This adds a group of drivers who definitively do not have OSA, but who are likely to be obese or have other indicators of potential illness that led them to be designated as "high priority" for a PSG by the Somni-Sage® screening questionnaire.

As may be observed in Table S6, this raises the DOT-reportable preventable crash rate for control drivers rather noticeably, from 0.14 crashes per 100,000 miles to 0.24. As a result, the incidence rate ratio for Partial Adherence and Full Adherence now drop from a little above 1.0 to a little below. But as before, only No Adherence drivers have a crash rate that is statistically different from Controls, and it is three times as high. The pattern is similar for the Andersen-Gill model used as a robustness check: the relative risk of Partial and Full Adherence drivers falls but is not statistically different from that of controls, and only the No Adherence sub-group has a statistically higher risk, though not quite as high as in the results when Negatives are not included in the controls (complete details available upon request from the authors). Thus, while the crash rate of the reference group of control drivers is raised by this approach, the pattern of relative crash risk across study sub-groups is the same as that presented in the main text.

4. **Further Details on Exits: Reasons for Separation by Study Sub-group**

As would be expected from the study firm's policy that drivers diagnosed to have OSA are required to maintain adherence with APAP treatment in order to continue in employment, drivers who never demonstrate adherence have significantly shorter job tenures after the PSG/matching date than other study sub-groups. The mean number of weeks each study sub-group is observed after the PSG or comparison date is recorded in Table S7. Drivers in the No Adherence sub-group are observed in the data for approximately 1/3 as long as any other sub-group. By contrast, Fully Adherent drivers are observed for a length of time that is comparable to that of controls, and Partially Adherent drivers are actually observed longer than controls.

Since the data set ends on December 31, 2009, some drivers from all sub-groups are censored, in that they remain as employees of the study firm on that date. A second way to examine the relative performance of drivers in the No Adherence sub-group is to tabulate the number who are censored (remain on December 31, 2009), and also the number who exit for each type of reason.

Table S7 shows that only 17.2% of drivers in the No Adherence sub-group were censored (remained as employees on December 31, 2009), as compared to 68% or more for controls and the other treatment-adherence sub-groups. This reflects that fact that some drivers received a PSG late in the study period, and the process of treatment adherence monitoring and remediation, followed discharge for continuing non-adherence, and not yet been completed for those exhibiting no adherence. In addition, the rate of total discharges among No Adherence drivers, at 23.9%, was approximately four times that of drivers in the Full Adherence sub-group (5.7%), reflecting discharges for all causes including preventable crashes and failing to comply with treatment.

It is perhaps most noteworthy that the largest proportion of all exits by drivers in the No Adherence sub-group is made up of voluntary quits (57.5%). Thus, most of these drivers choose to leave before the process of treatment-adherence monitoring led to their discharge.

5. **Regulatory Issue Details: Drivers with OSA Quitting the Study Firm Can Drive Elsewhere**

There is a further point that is raised by the findings of the present study about the regulations governing the medical qualifications for commercial drivers. Although a commercial driver's license medical examiner (CDME) may request that a driver undergo an OSA diagnostic test based on the examiner's clinical judgment, there are currently no formal standards which require such a test to be ordered for all commercial drivers meeting some particular set of criteria, and none may be established in the absence of a full rulemaking process due to recent legislation.[13-16] Due to medical privacy rules, and despite the newly established Registry for CDMEs,[17] there is also no required procedure—other than the subject's self-report—by which a positive OSA diagnosis received by a given subject outside the biennial CDL medical examination process would necessarily be brought to the attention of the medical examiner who gives that subject his or her next CDL medical examination.[18] And, while the FMCSA Form 649 CDL medical examination document asks drivers to self-report any sleep disorder (among other qualificationrelevant symptoms or conditions),[19,20] there is substantial evidence that commercial drivers strongly under-report such symptoms and conditions, due to the potential for losing a medical certification and becoming unemployable as a CMV driver.[21-23] Further, there is considerable resistance among some elements of the trucking industry to being proactive in screening for and diagnosing OSA.[24]

This raises the question of how many of drivers who have diagnosed OSA but no adherence with treatment might choose to quit from a motor carrier at which they were diagnosed and at which treatment was required, such as the study carrier, and instead seek employment—without revealing their OSA diagnosis—with a different firm that does not have an OSA program. Nearly 58% of the drivers diagnosed with OSA but never adherent with treatment at the study firm were observed to quit in the post-PSG interval (Table S7). This suggests that drivers found to have a five-fold increase in the risk of a serious preventable crash by the present study could choose to further expose themselves and the motoring public to that risk while working at a carrier that does not have an OSA program.

6. **References**

1. Epstein LJ, Kristo D, Strollo PJ, Jr., et al. Clinical guideline for the evaluation, management and long-term care of obstructive sleep apnea in adults. *J Clin Sleep Med* 2009; 5(3): 263-76.
2. Burks SV, Carpenter, J., Götte, L., Rustichini, A. Cognitive Skills Affect Economic Preferences, Social Awareness, and Job Attachment. *Proceedings of the National Academy of Science* 2009; 106(19): 7745-50.
3. Burks SV, Belzer M, Kwan Q, Pratt S, Shackelford S. Trucking 101: An Industry Primer. *Transportation Research Board Research Circular* 2010; (Number E-C146).
4. Driver Turnover Rate at Large Truckload Carriers Rises to 136%. *Transport Topics Online*, 2006. (accessed August 6, 2013).
5. Guest M, Boggess MM, Duke JM. Age Related Annual Crash Incidence Rate Ratios in Professional Drivers of Heavy Goods Vehicles. *Trans Res: Pt A: Pol & Prac* 2014; 65: 1-8.
6. Rodriguez DA, Targa F, Belzer MH. Pay Incentives and Truck Driver Safety: A Case Study. *Industrial and Labor Relations Review* 2006; 59(2): 205-25.
7. Transport Topics. December 16: Driver Turnover Rose in Third Quarter, ATA Reports. 2006b. http://www.ttnews.com/articles/basetemplate.aspx?storyid=16694&t=Driver-Turnover-Rose-in-Third-Quarter-ATA-Reports (accessed August 8 2013).
8. Anderson JE, Govada M, Steffen TK, et al. Obesity is associated with the future risk of heavy truck crashes among newly recruited commercial drivers. *Acc Anal Prev* 2012; 49: 378-84.
9. Cleves M, Gutierrez RG, Gould W, Marchenko YV. An Introduction to Survival Analysis Using Stata, Third Edition. Third ed. College Station, TX: Stata Press; 2010.
10. Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox Model. New York, NY: Springer; 2000.
11. FMCSA. Crash Weighting Analysis. Washington, DC: US Department of Transportation, 2015.

12. NSC Staff. Guide to Determine Motor Vehicle Collision Preventability. Itasca, IL: National Safety Council; 2011.
13. FMCSA. Federal Motor Carrier Safety Administratin Medical Examiner Handbook. Washington, DC: US Department of Transportation; 2014. p. 1-260.
14. A bill to ensure that any new or revised requirement providing for the screening, testing, or treatment of individuals operating commercial motor vehicles for sleep disorders is adopted pursuant to a rulemaking proceeding. 49 USCA. 113th Congress of the United States of America, 1st Session ed; 2013.
15. Parker Poe staff. New Bill Will Prohibit DOT From Addressing Sleep Apnea Without Rulemaking. 2013.
16. Jaillet J. President signs bill forbidding FMCSA to use sleep apnea 'guidance'. *Commercial Carrier Journal - Fleet Management*, 2013. (accessed.
17. FMCSA. National Registry of Certified Medical Examiners. In: FMCSA, editor. 49 CFR Parts 350, 383, 390, and 391. Washington, DC: Federal Register; April 20, 2012. p. pp. 24104-35.
18. Health Insurance Portability and Accountability Act; Public Law 104-191. HIPAA-1996.
19. FMCSA. Form 649-F (6045): Medical Examination Report for Commercial Driver Fitness Determination. Washington, DC: US Department of Transportation. p. 1-8.
20. FMCSA. Driver Medical Fitness for Duty. March 19, 2014 2014. http://www.fmcsa.dot.gov/medical/driver-medical-requirements/driver-medical-fitness-duty (accessed June 3 2014).
21. Dagan Y, Doljansky JT, Green A, Weiner A. Body Mass Index (BMI) as a first-line screening criterion for detection of excessive daytime sleepiness among professional drivers. *Traffic Inj Prev* 2006; 7(1): 44-8.
22. Parks P, Durand G, Tsismenakis AJ, Vela-Bueno A, Kales S. Screening for obstructive sleep apnea during commercial driver medical examinations. *J Occup Environ Med* 2009; 51(3): 275-82.
23. Talmage JB, Hudson TB, Hegmann KT, Thiese MS. Consensus criteria for screening commercial drivers for obstructive sleep apnea: evidence of efficacy. *J Occup Environ Med* 2008; 50(3): 324-9.
24. Supplement on Sleep Disorders. *ACOEM Com Drvr Med Exmr Rev* 2014; (Fall): 1-8.

7. **Tables and Figure**

Table S1—Tenure in weeks for diagnosed drivers at their PSG dates versus a cross section of the reference population taken in the middle of the study period.

Percentile of Driver Group	**A.** **Reference population at study midpoint (weeks of tenure)**	**B.** **Diagnosed drivers at PSG date (weeks of tenure)**
10%	9.7	18.7
25%	31.7	40.1
50%	98.8	105.7
75%	324.9	359.0
90%	715.0	751.0
Avg	229.3	246.6

The probability that Column B values are greater than or equal to Column A values by chance variation is less than 0.001 on every line (Rank Sum Test). Thus, this table shows that entire group of drivers who received a PSG has significantly higher job tenure at their PSG date than the reference population overall as of a date in the midpoint of the study period, July 3, 2007. Selection of other dates to take the population cross section in the study period produces similar findings.

Table S2—Screenings and PSGs by year

	2005	**2006**	**2007**	**2008**	**2009**
Screenings	0	6,424	3,701	2,983	5,188
PSGs	5	493	370	632	662

Table S3—Andersen-Gill time-to-crash multivariate model estimated results.

		Hazard Ratio	**P value**
1	OSA Negative (Before PSG)	1.233	0.487
2	OSA Fully Adherent (Before PSG)	0.820	0.493
3	OSA Partially Adherent (Before PSG)	1.301	0.302
4	OSA Never Adherent (Before PSG)	1.560	0.155
5	OSA Control (Before PSG)	BASE	
6	After PSG vs Before PSG	1.065	0.790
7	OSA Negative x After PSG vs Before PSG	1.089	0.865
8	OSA Fully Adherent x After PSG vs Before PSG	1.368	0.466
9	OSA Partially Adherent x After PSG vs Before PSG	1.065	0.870
10	OSA Never Adherent x After PSG vs Before PSG	2.430	0.073
11	Age 21 to 40 at PSG/Comp	1.549*	0.017
12	Age 51+ at PSG/Comp	1.564*	0.018
13	Age 41 to 50 at PSG/Comp	BASE	
14	Race African American	1.356	0.123
15	Race Other	1.138	0.574
16	Race Caucasian	BASE	
17	Sex Female	0.622	0.158
18	Sex Male & Other	BASE	
19	0-5 segments per week	0.884	0.489
20	More than 11 Segments per week	0.526*	0.014
21	5-10 segments per week	BASE	
22	Job Type Dedicated	0.778	0.132
23	Job Type Other	0.585	0.103
24	Job Type Solo System	BASE	
25	Year 2005-2006	0.824	0.319
26	Year 2008	0.638*	0.038
27	Year 2009	0.755	0.216
28	Year 2007	BASE	
29	Spring	1.010	0.962
30	Summer	1.000	0.999
31	Fall	1.309	0.171
32	Winter	BASE	
33	0-1500 miles this week	1.734**	0.001
34	More than 2500 miles this week	0.484**	0.002
35	1500 - 2500 miles this week	BASE	
	N (driver-week observations)	602,697	

Key: *$P \leq 0.05$; **$P \leq 0.01$. This model covers both the before-PSG/matching date period and the after-PSG/matching date period. As detailed in the paper, driver sub-groups are defined based on driver diagnosis (or control) status and, as applicable, driver treatment adherence behavior after the PSG (even though the model estimates crash rate differences for both the before- and after-PSG periods). The HRs on lines 1-4 show that no study sub-group has a crash risk that is statistically different from that of controls in the "before PSG/matching date" period. As noted in prior sections, this is due to the effects of safety-selection during this period, which washes out crash risk differences by removing drivers with bad preventable crashes from the study pool. The results for the after-PSG/matching date period are generated using tests of the statistical significance of linear combinations of some of the model coefficients, specifically, those that produce the hazard ratios shown on rows 6-10. The base level for comparison is a driver-week with these characteristics: Control (drawn from those screened as at Low-Priority for OSA diagnosis); Male or missing sex; Caucasian; Age 41-50 at PSG/comparison Date; Solo system driver; Year 2007; Season Winter; 5-10 trip segments, and 1,500-2,500 miles.

Table S4—Andersen-Gill model results for all DOT-reportable crashes, irrespective of preventability status.

Study Group	Drivers	Avg Weeks per Driver	Crashes	HR	95% CI	P value
Controls	2,016	64	163	1.00	BASE	BASE
Negatives	403	57	37	1.28	0.865, 1.884	0.22
Full Adherence	682	65	60	1.12	0.830, 1.513	0.46
Partial Adherence	571	73	63	1.24	0.936, 1.633	0.14
No Adherence	360	21	20	2.21***	1.406, 3.476	< 0.001

Key: ***$P \leq 0.001$. Controls are drivers screened as "Low-Priority" and are unlikely to have OSA. OSA cases are defined as those with AHI $\geq$ 5. Data is in one-observation-per-driver-week format. Hazard ratio estimates are generated as linear combinations of coefficients from an underlying Andersen-Gill model for the period after the matching date. Crash counts are higher now that all DOT-reportable crashes are included, but exposure elements (drivers, average weeks per driver) are the same as in main text, Table 2.

Table S5—Andersen-Gill model results for DOT-reportable preventable crashes using AHI $\geq$ 15 as the criterion for a positive OSA diagnosis.

Study Group	Drivers	Avg Weeks per Driver	Crashes	HR	95% CI	P value
Controls	2,016	64	33	1.00	BASE	BASE
OSA Negative	855	57	15	1.20	0.652, 2.217	0.555
Full Adherence	523	66	9	1.06	0.507, 2.209	0.879
Partial Adherence	397	71	14	1.74	0.933, 3.239	0.082
No Adherence	241	19	7	4.54***	2.542, 10.318	0.000

Key: ***$P \leq 0.001$. Controls are drivers screened as "Low-Priority" and are unlikely to have OSA. OSA cases are defined as those with AHI $\geq$ 15. Data is in one-observation-per-driver-week format. Hazard ratio estimates are generated as linear combinations of coefficients from an underlying Andersen-Gill model. Drivers who were considered to have OSA but had AHI $\geq$ 5 and AHI < 15 are now shifted into the Negative sub-group. The numbers of drivers, average weeks per driver, and crash counts are all adjusted accordingly.

Table S6—Incidence rate ratios; DOT-reportable preventable crashes per 100,000 miles by study sub-group when drivers diagnosed as negative for OSA are included among controls.

Study Sub-Group	Mean Weeks Observed	Crash Rate	IRR	95% CI	P value
Controls (including Negatives)	65.3	0.024	1.00	n/a	Base
Partial Adherence	65.0	0.015	0.63	0.305, 1.225	0.159
Full Adherence	72.5	0.022	0.95	0.500, 1.720	0.876
No Adherence	21.1	0.074	3.14	1.344, 6.522	0.006

Table S7—Exit Reasons by study sub-group.

Exit Reason	Study Sub-Groups										
	Controls		OSA Negative		Full Adherence		Partial Adherence		No Adherence		Total
	n	%	n	%	n	%	n	%	n	%	n
Discharge	129	6.4%	46	11.4%	39	5.7%	58	10.2%	86	23.9%	358
Quit	473	23.5%	112	27.8%	132	19.4%	116	20.3%	207	57.5%	1,040
Missing Data	27	1.3%	1	0.3%	9	1.3%	4	0.7%	5	1.4%	46
Still Employed	1,387	68.8%	244	60.6%	502	73.6%	393	68.8%	62	17.2%	2,588
Total	2,016		403		682		571		360		4,032

A

B

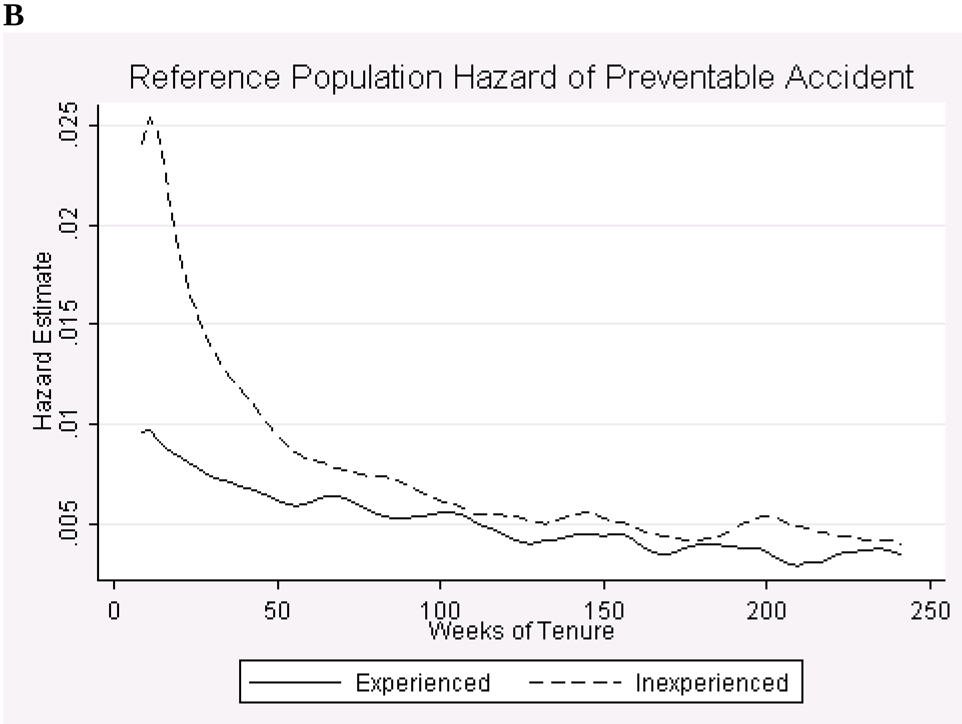

Figure S1—Exit and crash rates for the reference population, by job tenure and by experience level at hire. The hazard of both exits **(A)** and crash rates **(B)** decline as tenure increases. About 75% of exits are quits and 25% are discharges. (The second spike in exits for those who are inexperienced at hire is primarily due to completion of a 12-month contract which canceled the driver's debt for initial training.) The crash graph displays all preventable crashes, and the decline reflects both the growing experience of those who remain and the discharge of those who accumulate an unacceptable record of preventable crashes.