Projekt: Weiterentwicklung Methode und Tool Sicherheitsmonitoring öV Schweiz

Simulationen zur Empfindlichkeit unterschiedlicher methodischer Ansätze im Sicherheitsmonitoring öV Schweiz

Michael Vock,
Institut für mathematische Statistik und Versicherungslehre, Universität Bern
michael.vock@unibe.ch

Auftraggeber:
Bundesamt für Verkehr, Sektion Sicherheitsrisiko-Management

Überarbeitete Version vom 1. Juli 2024

Zusammenfassung

Die hier beschriebenen Untersuchungen ergänzen die 2015 durchgeführten Simulationen zur Validität unterschiedlicher methodischer Ansätze. Während damals lediglich überprüft wurde, ob die Methoden die vorgegebene Wahrscheinlichkeit für einen „falschen Alarm“ (im statistischen Jargon: die Fehlerwahrscheinlichkeit 1. Art) einhalten, wurde nun zusätzlich geprüft, mit welcher Wahrscheinlichkeit gewisse Veränderungen in den Verteilungen der betrachteten Daten entdeckt werden (also mit welcher Empfindlichkeit die Methoden auf ein gewisse Veränderung reagieren; im statistischen Jargon: Güte/Power bzw. 1 minus die Fehlerwahrscheinlichkeit zweiter Art). Ausserdem wurden zusätzliche, in der Zwischenzeit entwickelte Verfahren betrachtet. Neben den ursprünglich verwendeten zweiseitigen Tests wurde auch der Fall einseitiger Tests miteinbezogen, welche für die vorgesehene Anwendung adäquater erscheinen, da primär Verschlechterungen des Sicherheitsniveaus entdeckt werden sollen.

In der vorliegenden überarbeiteten Fassung des Berichts wurden gegenüber der Version von 2018 Anpassungen am Prior für den bayesianischen Ansatz gemäss Andrášik (2019) berücksichtigt und ein weiteres Verfahren („rate ratio test“ in zwei Varianten) in den Vergleich miteinbezogen.

Für den Vergleich der Anzahl Ereignisse scheinen zwei der betrachteten Verfahren besonders geeignet (und in der Praxis liefern fast immer beide die gleiche Schlussfolgerung), drei weitere liefern fast so gute Ergebnisse und kämen ebenfalls in Frage. Zwei weitere Verfahren liefern typischerweise eine etwas tiefere Power, und nochmals zwei zusätzliche Verfahren sind für den einseitigen Fall denkbare Alternativen. Der Vergleich der FWSI- und FWI-Werte scheint wesentlich schwieriger zu sein; als Alternative zum Vergleich der FWSI-Werte über Wilcoxon-Tests kommt ein bayesianischer Ansatz in Frage, der direkt auf den Häufigkeiten tödlicher bzw. schwerer Verletzungen beruht.

Das genaue Vorgehen und die Resultate werden im Folgenden auf Englisch beschrieben.

Aim and Basic Approach of the Study

Several methods have been proposed for the comparison of safety indicator data from a current/ target period to corresponding data from a reference period in the past. A selection of such methods ("tests") was compared in a previous simulation study (Vock, 2015) with respect to their performance under the statistical null hypothesis of no change. In that study, their actual type I error rate ("size") was estimated, which is the probability that a certain test indicates that there is a change when there actually was no change in the distribution of the safety indicator ("false alert"). While a type I error rate close to the specified value (i. e., close to the nominal significance level) is an important property of a statistical test, a good test should also provide a sufficiently small type II error rate, which corresponds to the probability of not detecting a change that has actually occurred. Equivalently, a test should have a sufficiently high power, which is the probability of detecting such a change and can also be called the sensitivity of a test.

In the present study, the type I error rate as well as the power of several approaches are estimated by simulation. For that purpose, a parametric model was fitted to available data for a certain indicator (i. e., type of event). This model was then used to simulate data from the reference period, while different scenarios (corresponding to modifications of the model for the reference period) were used to simulate data from the current period. By considering different indicators and different scenarios, it should be possible to provide an adequate overall assessment of the performance of the different tests in situations that are relevant for practical application. For each combination of an indicator and a scenario, 100,000 complete sets of data for the reference and the current period were simulated, which should provide sufficiently precise estimates of the probabilities of interest.

Models for the Distribution of the Data

Real data from the years 2009 to 2016 of five types of events ("indicators") were used to fit a parametric model for the number of events and the resulting injuries for each indicator. This model consists of three independent parts:

- The occurrence of events is modelled by a homogeneous Poisson point process, i. e., the time between two events has an exponential distribution, and the events occur independently from each other. The only estimated parameter for this part of the model is the rate/intensity with which the events occur.
- Given that an event has happened, the number of injuries is modelled by one of the following four distributions (using formal goodness-of-fit tests as well as graphical methods and selecting the simplest model that provides an adequate fit):
    - Poisson: This is a classical model for count data, which should be adequate if the individual injuries could be assumed to occur independently of each other. The only parameter is the expected number of injuries per event (lambda).
    - Negative binomial: This is traditionally used in cases where the independence assumption of the Poisson distribution is implausible; it may provide a reasonable fit even if the variability of the numbers of injuries is too high for a Poisson model.

Two parameters have to be estimated; in the parametrization used here, these are the expected number of injuries (mu) and a "size" (or dispersion) parameter.

  - Bernoulli and positive Poisson: A Bernoulli (0/1) random variable is first used to determine whether there are any injuries. If this is the case, a "positive Poisson" distribution is used for the number of injuries, which is just the positive part of the Poisson distribution, scaled in order to sum to 1 again. This provides more flexibility compared to the Poisson model, in the sense that events without any injuries could be more or less probable than under the corresponding Poisson distribution. The unknown parameters are the probability of a positive number of injuries (p.pos) as well as the expected value of the underlying Poisson distribution (lambda).
  - Bernoulli and positive negative binomial: This modifies the negative binomial distribution in the analogous way, using a Bernoulli random variable and the positive part of the negative binomial distribution, scaled in order to sum to 1 again. The unknown parameters are the probability of a positive number of injuries (p.pos) as well as the expected value (mu) and the size parameter of the underlying negative binomial distribution.

- Finally, given the total number of injuries in a certain event, a multinomial distribution was used for modelling the number of fatal, severe, and light injuries. The unknown parameters are the three multinomial probabilities (p.f, p.s, p.l, which have to sum to 1).

Based on this fitted parametric model for a certain indicator for the reference period, different scenarios were considered for the current period:

- The same scenario as for the reference,
- the reference scenario, but with the event rate multiplied by a certain factor,
- the reference scenario, but with the mean number of injuries per event multiplied by a certain factor,
- the reference scenario, but with increased probabilities of fatal and severe injuries.

For each of the three "directions" of modifying the reference scenario, two steps were used – e. g., the event rate was multiplied by a factor of 1.2 or 1.5 in the case of the indicator "Zusammenstoesse". The step sizes were chosen in such a way that some of the tests reached a reasonable power to detect the changes, but that the powers did not become too high to be interesting.

When increasing the mean number of injuries in the Bernoulli and positive Poisson/negative binomial cases, the expected value of the underlying Poisson or negative binomial distribution was multiplied by the desired factor and the probability of a positive number adapted accordingly to obtain the same factor for the overall expectation, as long as this was possible. Otherwise, the probability of a positive number was set to 1 and the expected value of the underlying distribution was increased by a larger factor in order to obtain the desired change for the overall expectation. The size parameter of the negative binomial distribution remained unchanged in all scenarios.

In the simulations, the reference period was assumed to consist of four consecutive years (2009 to 2012), and a single year (2013) was used as the current period. The simulated event data were aggregated into monthly counts of events, numbers of injuries of each type, as well as FWSI and FWI values.

Two-Sided Tests Considered

Two-sided tests are designed to detect any change in the corresponding variable, which can represent an improvement or a deterioration with respect to safety. The following two-sided tests were considered in the simulation study (at the 5% significance level, unless otherwise indicated):

- For the number of events:
    - Overlap of bootstrap ABC CI. For each period, a 95% Bootstrap ABC confidence interval (CI) is calculated; these confidence intervals are adjusted in order to more closely approximate the desired type I error rate of 5% in the following step, which indicates a change if the two intervals do not overlap.
    - Wilcoxon test, R default. This uses Wilcoxon's rank sum test; for small samples and no ties, the exact distribution is used, while for larger samples, an asymptotic approximation is used.
    - Wilcoxon test, exact. Wilcoxon's rank sum test using the exact distribution in all cases.
    - CI for odds ratio. This checks whether an approximate 95% confidence interval for the odds ratio covers the value 1 (Andrášik, 2015).
    - Bootstrap ABC CI for ratio of means. This checks whether a 95% Bootstrap ABC confidence interval for the ratio of the expected values in the two periods covers the value 1.
    - Wald test for Poisson GLM. Wald test for the effect of the binary predictor corresponding to the period in a generalized linear model (GLM) with a Poisson response and a log link used to model the number of events.
    - LR test for Poisson GLM. Likelihood ratio (LR) test for the same effect in this model.
    - Bayesian approach, two-sided. Checks whether the posterior probability from a Bayesian model for the number of events (Andrášik, 2019) is either below the specified threshold or above 1 minus the threshold. The threshold is chosen in such a way that the resulting probability of indicating a change is approximately 0.05 when the same distribution is used for both periods.
    - Exact rate ratio test for the comparison of two Poisson counts, taking into account the time at risk associated with each of the counts.
    - A modification of the exact rate ratio test using the mid-p-value.
- For the FWSI (calculated from the numbers of fatal and severe injuries):
    - Overlap of bootstrap ABC CI. As above.
    - Wilcoxon test, R default. As above.
    - Wilcoxon test, exact. As above.
    - Bootstrap ABC CI for ratio of means. As above.
- For the FWI (calculated from the numbers of all three types of injuries):
    - Overlap of bootstrap ABC CI. As above.
    - Wilcoxon test, R default. As above.
    - Wilcoxon test, exact. As above.
    - Bootstrap ABC CI for ratio of means. As above.

- For the numbers of fatal and severe injuries used separately:
    - Bayesian approach, two-sided. Checks whether the posterior probability from a Bayesian model for the FWSI (Andrášik, 2019) is either below the specified threshold or above 1 minus the threshold. The threshold is chosen in such a way that the resulting probability of indicating a change is approximately 0.05 when the same distribution is used for both periods.

Results for the Two-Sided Tests

The estimated probabilities of indicating a change are given in the following tables, for each of the five indicators considered. Colors were used to facilitate the interpretation:

In the "reference" column, everything below 0.05 is in dark green. In that scenario, both periods have data from the same distribution, and one would like to see a "false alert" in at most 5% of the cases. Slightly higher values (in the color range of light green and yellow) may still be acceptable – this corresponds to slightly more false alarms than specified.

In the other columns, large probabilities are good since they indicate a high power to detect a certain change. Everything below 0.05 is dark red – these are tests that are not able to detect the corresponding change with a probability that is larger than the specified error probability under the reference scenario. No general rule can be given about what a good power is here, but one should primarily compare between different tests for the same summary measure. It is clear that for all five indicators considered, all tests for the number of events are unable to meaningfully detect changes in the number of injuries per event and the distribution of injury severities, since these tests do not use any information affected by these changes.

In addition to the results, the model fitted to the real data (used for the reference period) as well as the scenarios for the current period are specified in detail for each indicator.

Personenunfall (accidents to persons):

	Scenario						
	reference	more events		more injuries per event		more severe injuries	
Summary measure / Test applied	**ref**	**rate*1.1**	**rate*1.3**	**mu*1.2**	**mu*1.5**	**p.fsl=0.025, 0.17,0.805**	**p.fsl=0.04, 0.2,0.76**
Number of events							
Overlap of bootstrap ABC CI	0.07870	0.20021	0.82151	0.07892	0.07881	0.07851	0.07813
Wilcoxon test, R default	0.04623	0.15534	0.77480	0.04693	0.04685	0.04681	0.04720
Wilcoxon test, exact	0.04779	0.15842	0.77919	0.04845	0.04826	0.04831	0.04868
CI for odds ratio	0.04893	0.18019	0.83276	0.04893	0.04893	0.04893	0.04893
Bootstrap ABC CI for ratio of means	0.07510	0.20217	0.82489	0.07474	0.07552	0.07423	0.07406
Wald test for Poisson GLM	0.04893	0.18019	0.83276	0.04893	0.04893	0.04893	0.04893
LR test for Poisson GLM	0.04923	0.17169	0.82431	0.04923	0.04923	0.04923	0.04923
Bayesian approach, two-sided (threshold: 0.026)	0.04996	0.17341	0.82598	0.04996	0.04996	0.04996	0.04996
Exact rate ratio test assuming Poisson counts	0.04401	0.16245	0.81425	0.04401	0.04401	0.04401	0.04401
Rate ratio test assuming Poisson counts, mid-p	0.04894	0.17355	0.82599	0.04894	0.04894	0.04894	0.04894
FWSI							
Overlap of bootstrap ABC CI	0.11319	0.09770	0.12475	0.10398	0.20790	0.13431	0.43377
Wilcoxon test, R default	0.04809	0.05950	0.15015	0.09021	0.27892	0.12957	0.44837
Wilcoxon test, exact	0.04953	0.06127	0.15330	0.09229	0.28340	0.13258	0.45392
Bootstrap ABC CI for ratio of means	0.10397	0.09711	0.12498	0.10315	0.20288	0.13207	0.42564
FWI							
Overlap of bootstrap ABC CI	0.11333	0.09554	0.15595	0.11301	0.30004	0.13119	0.40939
Wilcoxon test, R default	0.04850	0.06938	0.25105	0.12857	0.45205	0.11519	0.38825
Wilcoxon test, exact	0.05016	0.07157	0.25636	0.13173	0.45893	0.11808	0.39380
Bootstrap ABC CI for ratio of means	0.10494	0.10013	0.15947	0.11633	0.29606	0.13068	0.40636
Numbers of fatal and severe injuries							
Bayesian approach, two-sided (threshold: 0.062)	0.04968	0.06643	0.15620	0.10557	0.31078	0.15895	0.53461
Model fitted to the original data ("ref" scenario)							
Poisson point process intensity:		0.365					
number of injuries per event:		Bernoulli/positive Poisson with p.pos=0.994, lambda=0.0255					
multinomial probabilities for fatal/severe/light injuries:		0.017, 0.143, 0.840					

Scenario	ppp.intensity	size.generator	p.f	p.s	p.l
ref	0.3650	rbernpois(p.pos=0.994, lambda=0.0255)	0.0170	0.1430	0.8400
rate*1.1	0.4015	rbernpois(p.pos=0.994, lambda=0.0255)	0.0170	0.1430	0.8400
rate*1.3	0.4745	rbernpois(p.pos=0.994, lambda=0.0255)	0.0170	0.1430	0.8400
mu*1.2	0.3650	rbernpois(p.pos=1, lambda=0.390761152718768)	0.0170	0.1430	0.8400
mu*1.5	0.3650	rbernpois(p.pos=1, lambda=0.889903825500934)	0.0170	0.1430	0.8400
p.fsl=0.025,0.17,0.805	0.3650	rbernpois(p.pos=0.994, lambda=0.0255)	0.0250	0.1700	0.8050
p.fsl=0.04,0.2,0.76	0.3650	rbernpois(p.pos=0.994, lambda=0.0255)	0.0400	0.2000	0.7600

Zusammenstoesse (collisions of trains):

	Scenario						
	reference	more events		more injuries per event		more severe injuries	
Summary measure / Test applied	**ref**	**rate*1.2**	**rate*1.5**	**mu*2**	**mu*3**	**p.fsl=0.05, 0.2,0.75**	**p.fsl=0.08, 0.25,0.67**
Number of events							
Overlap of bootstrap ABC CI	0.09212	0.11980	0.33328	0.09217	0.09121	0.09216	0.09211
Wilcoxon test, R default	0.04805	0.09205	0.28990	0.04744	0.04640	0.04728	0.04702
Wilcoxon test, exact	0.04791	0.09295	0.29300	0.04701	0.04609	0.04702	0.04677
CI for odds ratio	0.04406	0.10789	0.34588	0.04406	0.04406	0.04406	0.04406
Bootstrap ABC CI for ratio of means	0.07356	0.11604	0.33549	0.07358	0.07260	0.07367	0.07255
Wald test for Poisson GLM	0.04406	0.10789	0.34588	0.04406	0.04406	0.04406	0.04406
LR test for Poisson GLM	0.05155	0.09910	0.32293	0.05155	0.05155	0.05155	0.05155
Bayesian approach, two-sided (threshold: 0.032)	0.05006	0.09987	0.32731	0.05006	0.05006	0.05006	0.05006
Exact rate ratio test assuming Poisson counts	0.03323	0.07643	0.28143	0.03323	0.03323	0.03323	0.03323
Rate ratio test assuming Poisson counts, mid-p	0.04741	0.09811	0.32344	0.04741	0.04741	0.04741	0.04741
FWSI							
Overlap of bootstrap ABC CI	0.02583	0.03819	0.06047	0.07696	0.13357	0.08100	0.13790
Wilcoxon test, R default	0.03715	0.05790	0.09564	0.08846	0.13597	0.07054	0.09910
Wilcoxon test, exact	0.04258	0.06689	0.11110	0.10885	0.16933	0.09082	0.13041
Bootstrap ABC CI for ratio of means	0.50902	0.50130	0.47685	0.51811	0.52373	0.56204	0.58050
FWI							
Overlap of bootstrap ABC CI	0.03233	0.04566	0.06961	0.08779	0.15138	0.07666	0.12336
Wilcoxon test, R default	0.04508	0.06628	0.11890	0.07489	0.10420	0.05013	0.05442
Wilcoxon test, exact	0.04059	0.06620	0.12355	0.07578	0.10677	0.04571	0.05032
Bootstrap ABC CI for ratio of means	0.51933	0.45185	0.36706	0.46579	0.46501	0.56474	0.57458
Numbers of fatal and severe injuries							
Bayesian approach, two-sided (threshold: 0.057)	0.04995	0.06907	0.10210	0.19243	0.33023	0.14737	0.24454
Model fitted to the original data ("ref" scenario)							
Poisson point process intensity:		0.0369					
number of injuries per event:		Negative binomial with size=0.102, mu=1.30					
multinomial probabilities for fatal/severe/light injuries:		0.016, 0.144, 0.840					

Scenario	ppp.intensity	size.generator	p.f	p.s	p.l
ref	0.0369	rnbinom(size=0.102, mu=1.3)	0.0160	0.1440	0.8400
rate*1.2	0.0443	rnbinom(size=0.102, mu=1.3)	0.0160	0.1440	0.8400
rate*1.5	0.0554	rnbinom(size=0.102, mu=1.3)	0.0160	0.1440	0.8400
mu*2	0.0369	rnbinom(size=0.102, mu=2.6)	0.0160	0.1440	0.8400
mu*3	0.0369	rnbinom(size=0.102, mu=3.9)	0.0160	0.1440	0.8400
p.fsl=0.05,0.2,0.75	0.0369	rnbinom(size=0.102, mu=1.3)	0.0500	0.2000	0.7500
p.fsl=0.08,0.25,0.67	0.0369	rnbinom(size=0.102, mu=1.3)	0.0800	0.2500	0.6700

Schiffe (ship accidents):

	Scenario						
	reference	more events		more injuries per event		more severe injuries	
Summary measure / Test applied	**ref**	**rate*1.5**	**rate*2**	**mu*2**	**mu*3**	**p.fsl=0.15, 0.25,0.6**	**p.fsl=0.25, 0.35,0.4**
Number of events							
Overlap of bootstrap ABC CI	0.10353	0.18990	0.47030	0.10295	0.10321	0.10307	0.10309
Wilcoxon test, R default	0.04732	0.17157	0.43300	0.04724	0.04754	0.04676	0.04675
Wilcoxon test, exact	0.04289	0.17052	0.43310	0.04228	0.04290	0.04200	0.04211
CI for odds ratio	0.03444	0.20416	0.50990	0.03444	0.03444	0.03444	0.03444
Bootstrap ABC CI for ratio of means	0.06824	0.18637	0.47063	0.06865	0.06846	0.06892	0.06849
Wald test for Poisson GLM	0.03434	0.20416	0.50990	0.03434	0.03434	0.03434	0.03434
LR test for Poisson GLM	0.05390	0.18232	0.47333	0.05390	0.05390	0.05390	0.05390
Bayesian approach, two-sided (threshold: 0.039)	0.04996	0.18836	0.49129	0.04996	0.04996	0.04996	0.04996
Exact rate ratio test assuming Poisson counts	0.02862	0.14510	0.42265	0.02862	0.02862	0.02862	0.02862
Rate ratio test assuming Poisson counts, mid-p	0.04610	0.18149	0.47328	0.04610	0.04610	0.04610	0.04610
FWSI							
Overlap of bootstrap ABC CI	0.01420	0.03247	0.05745	0.04731	0.08196	0.04347	0.08395
Wilcoxon test, R default	0.04309	0.08824	0.14447	0.10352	0.15419	0.08022	0.12513
Wilcoxon test, exact	0.03053	0.06989	0.12474	0.08991	0.14663	0.07081	0.12459
Bootstrap ABC CI for ratio of means	0.43309	0.52516	0.57132	0.54557	0.59188	0.52529	0.58160
FWI							
Overlap of bootstrap ABC CI	0.02516	0.05402	0.09275	0.07176	0.11458	0.05511	0.09609
Wilcoxon test, R default	0.03537	0.09681	0.19260	0.08237	0.11787	0.03922	0.04483
Wilcoxon test, exact	0.04319	0.11576	0.22032	0.10378	0.14951	0.05233	0.06348
Bootstrap ABC CI for ratio of means	0.66835	0.66765	0.62418	0.68779	0.68926	0.69276	0.68657
Numbers of fatal and severe injuries							
Bayesian approach, two-sided (threshold: 0.197)	0.04989	0.09207	0.14592	0.15794	0.27098	0.14641	0.26278
Model fitted to the original data ("ref" scenario)							
Poisson point process intensity:		0.0173					
number of injuries per event:		Negative binomial with size=0.249, mu=0.617					
multinomial probabilities for fatal/severe/light injuries:		0.063, 0.187, 0.750					

Scenario	ppp.intensity	size.generator	p.f	p.s	p.l
ref	0.0173	rnbinom(size=0.249, mu=0.617)	0.0630	0.1870	0.7500
rate*1.5	0.0260	rnbinom(size=0.249, mu=0.617)	0.0630	0.1870	0.7500
rate*2	0.0346	rnbinom(size=0.249, mu=0.617)	0.0630	0.1870	0.7500
mu*2	0.0173	rnbinom(size=0.249, mu=1.234)	0.0630	0.1870	0.7500
mu*3	0.0173	rnbinom(size=0.249, mu=1.851)	0.0630	0.1870	0.7500
p.fsl=0.15,0.25,0.6	0.0173	rnbinom(size=0.249, mu=0.617)	0.1500	0.2500	0.6000
p.fsl=0.25,0.35,0.4	0.0173	rnbinom(size=0.249, mu=0.617)	0.2500	0.3500	0.4000

TechnDefekt_Nahv (technical defects in urban transport):

	Scenario						
	reference	more events		more injuries per event		more severe injuries	
Summary measure / Test applied	**ref**	**rate*1.5**	**rate*2**	**mu*2**	**mu*3**	**p.fsl=0.04, 0.08,0.88**	**p.fsl=0.06, 0.12,0.82**
Number of events							
Overlap of bootstrap ABC CI	0.10125	0.19576	0.49605	0.10178	0.10162	0.10189	0.10159
Wilcoxon test, R default	0.04644	0.17536	0.45393	0.04689	0.04671	0.04699	0.04648
Wilcoxon test, exact	0.04287	0.17420	0.45477	0.04310	0.04284	0.04314	0.04305
CI for odds ratio	0.03676	0.21023	0.53302	0.03676	0.03676	0.03676	0.03676
Bootstrap ABC CI for ratio of means	0.07004	0.19297	0.49574	0.06989	0.06931	0.06984	0.06932
Wald test for Poisson GLM	0.03661	0.21023	0.53302	0.03661	0.03661	0.03661	0.03661
LR test for Poisson GLM	0.05455	0.18638	0.49851	0.05455	0.05455	0.05455	0.05455
Bayesian approach, two-sided (threshold: 0.038)	0.04900	0.19782	0.52063	0.04900	0.04900	0.04900	0.04900
Exact rate ratio test assuming Poisson counts	0.02927	0.15208	0.45110	0.02927	0.02927	0.02927	0.02927
Rate ratio test assuming Poisson counts, mid-p	0.04731	0.18571	0.49844	0.04731	0.04731	0.04731	0.04731
FWSI							
Overlap of bootstrap ABC CI	0.00422	0.01086	0.01960	0.01873	0.04323	0.02263	0.05084
Wilcoxon test, R default	0.03175	0.06470	0.10554	0.09854	0.17771	0.11204	0.20761
Wilcoxon test, exact	0.01684	0.03854	0.06683	0.06245	0.12592	0.07209	0.14752
Bootstrap ABC CI for ratio of means	0.25287	0.33795	0.40467	0.39425	0.47593	0.41173	0.49628
FWI							
Overlap of bootstrap ABC CI	0.03843	0.09435	0.19420	0.14375	0.27179	0.07371	0.11571
Wilcoxon test, R default	0.04628	0.16533	0.42354	0.11063	0.15861	0.05173	0.05735
Wilcoxon test, exact	0.04430	0.16763	0.42803	0.11199	0.16095	0.05080	0.05685
Bootstrap ABC CI for ratio of means	0.29748	0.26139	0.30489	0.29786	0.38846	0.36040	0.40682
Numbers of fatal and severe injuries							
Bayesian approach, two-sided (threshold: 0.262)	0.04283	0.07525	0.11095	0.11370	0.19626	0.11986	0.20570
Model fitted to the original data ("ref" scenario)							
Poisson point process intensity:		0.0185					
number of injuries per event:		Bernoulli/positive Poisson with p.pos=0.980, lambda=0.0839					
multinomial probabilities for fatal/severe/light injuries:		0.019, 0.039, 0.942					

Scenario	ppp.intensity	size.generator	p.f	p.s	p.l
ref	0.0185	rbernpois(p.pos=0.98, lambda=0.0839)	0.0190	0.0390	0.9420
rate*1.5	0.0278	rbernpois(p.pos=0.98, lambda=0.0839)	0.0190	0.0390	0.9420
rate*2	0.0370	rbernpois(p.pos=0.98, lambda=0.0839)	0.0190	0.0390	0.9420
mu*2	0.0185	rbernpois(p.pos=1, lambda=1.65155011094964)	0.0190	0.0390	0.9420
mu*3	0.0185	rbernpois(p.pos=1, lambda=2.89567751396917)	0.0190	0.0390	0.9420
p.fsl=0.04,0.08,0.88	0.0185	rbernpois(p.pos=0.98, lambda=0.0839)	0.0400	0.0800	0.8800
p.fsl=0.06,0.12,0.82	0.0185	rbernpois(p.pos=0.98, lambda=0.0839)	0.0600	0.1200	0.8200

ZusammenstoesseZmZ (collisions of a train with a train):

	Scenario						
	reference	more events		more injuries per event		more severe injuries	
Summary measure / Test applied	**ref**	**rate*1.5**	**rate*2**	**mu*3**	**mu*5**	**p.fsl=0.05, 0.25,0.7**	**p.fsl=0.08, 0.35,0.57**
Number of events							
Overlap of bootstrap ABC CI	0.09939	0.14551	0.33525	0.09917	0.09948	0.09978	0.09964
Wilcoxon test, R default	0.04533	0.13685	0.32548	0.04530	0.04562	0.04548	0.04573
Wilcoxon test, exact	0.03641	0.13165	0.32154	0.03619	0.03692	0.03666	0.03626
CI for odds ratio	0.02982	0.15001	0.37190	0.02982	0.02982	0.02982	0.02982
Bootstrap ABC CI for ratio of means	0.10516	0.14445	0.33440	0.10521	0.10486	0.10543	0.10502
Wald test for Poisson GLM	0.02981	0.15001	0.37190	0.02981	0.02981	0.02981	0.02981
LR test for Poisson GLM	0.05608	0.13962	0.34650	0.05608	0.05608	0.05608	0.05608
Bayesian approach, two-sided (threshold: 0.049)	0.04954	0.14439	0.36184	0.04954	0.04954	0.04954	0.04954
Exact rate ratio test assuming Poisson counts	0.02448	0.09951	0.28235	0.02448	0.02448	0.02448	0.02448
Rate ratio test assuming Poisson counts, mid-p	0.04462	0.13755	0.34613	0.04462	0.04462	0.04462	0.04462
FWSI							
Overlap of bootstrap ABC CI	0.02176	0.05091	0.08688	0.09614	0.15959	0.06221	0.10044
Wilcoxon test, R default	0.04160	0.09462	0.16064	0.09681	0.12483	0.06975	0.08875
Wilcoxon test, exact	0.03972	0.09636	0.17116	0.11678	0.16109	0.08116	0.11196
Bootstrap ABC CI for ratio of means	0.50238	0.54546	0.53727	0.58691	0.61644	0.56516	0.59219
FWI							
Overlap of bootstrap ABC CI	0.02667	0.06183	0.10510	0.10875	0.17590	0.05983	0.09243
Wilcoxon test, R default	0.03667	0.09937	0.19220	0.05839	0.06842	0.03979	0.04290
Wilcoxon test, exact	0.04523	0.12092	0.22165	0.08369	0.10190	0.05486	0.06297
Bootstrap ABC CI for ratio of means	0.60187	0.56025	0.48509	0.61649	0.63795	0.61984	0.62298
Numbers of fatal and severe injuries							
Bayesian approach, two-sided (threshold: 0.06)	0.04995	0.09573	0.15736	0.29666	0.45214	0.16023	0.26416
Model fitted to the original data ("ref" scenario)							
Poisson point process intensity:		0.0114					
number of injuries per event:		Bernoulli/positive negative binomial with p.pos=0.412, size=0.699, mu=6.42					
multinomial probabilities for fatal/severe/light injuries:		0.019, 0.149, 0.832					

Scenario	ppp.intensity	size.generator	p.f	p.s	p.l
ref	0.0114	rbernnbinom(p.pos=0.412, size=0.699, mu=6.42)	0.0190	0.1490	0.8320
rate*1.5	0.0171	rbernnbinom(p.pos=0.412, size=0.699, mu=6.42)	0.0190	0.1490	0.8320
rate*2	0.0228	rbernnbinom(p.pos=0.412, size=0.699, mu=6.42)	0.0190	0.1490	0.8320
mu*3	0.0114	rbernnbinom(p.pos=0.464053396571847, size=0.699, mu=19.26)	0.0190	0.1490	0.8320
mu*5	0.0114	rbernnbinom(p.pos=0.47851735005759, size=0.699, mu=32.1)	0.0190	0.1490	0.8320
p.fsl=0.05,0.25,0.7	0.0114	rbernnbinom(p.pos=0.412, size=0.699, mu=6.42)	0.0500	0.2500	0.7000
p.fsl=0.08,0.35,0.57	0.0114	rbernnbinom(p.pos=0.412, size=0.699, mu=6.42)	0.0800	0.3500	0.5700

One-Sided Tests Considered

In order to focus on a deterioration of the safety level, it could be more adequate to use one-sided tests instead of the two-sided tests, which are currently applied in practice. One-sided tests are specifically designed to detect deteriorations of the safety level (and will typically be more powerful for such changes). Essentially, the same tests as in the two-sided case were considered; however, these were modified in order to provide such a one-sided version (again at the 5% significance level, unless otherwise indicated):

- For the number of events:
    - Overlap of bootstrap ABC CI. For the reference period, a 95% Bootstrap ABC upper confidence limit is calculated, and for the current period, a corresponding lower confidence limit is calculated; these one-sided confidence intervals are adjusted in order to more closely approximate the desired type I error rate of 5% in the following step, which indicates a change if the two intervals do not overlap.
    - Wilcoxon test, R default. This uses Wilcoxon's rank sum test (one-sided); for small samples and no ties, the exact distribution is used, while for larger samples, an asymptotic approximation is used.
    - Wilcoxon test, exact. Wilcoxon's rank sum test using the exact distribution in all cases (one-sided).
    - CI for odds ratio. This checks whether an approximate 95% upper confidence limit for the odds ratio is at least 1 (Andrášik, 2015).
    - Bootstrap ABC CI for ratio of means. This checks whether a 95% Bootstrap ABC upper confidence limit for the ratio of the expected values in the two periods is at least 1.
    - Wald test for Poisson GLM. One-sided Wald test for the effect of the binary predictor corresponding to the period in a generalized linear model (GLM) with a Poisson response and a log link used to model the number of events.
    - LR test for Poisson GLM. This uses a p-value of 1 whenever the estimated effect of the binary predictor in the same GLM corresponds to an improvement of the safety, and one half of the two-sided p-value of the likelihood ratio (LR) test for the same effect in this model otherwise.
    - Bayesian approach. Checks whether the posterior probability from a Bayesian model for the number of events (Andrášik, 2019) is below the specified threshold. The threshold is chosen in such a way that the resulting probability of indicating a change is approximately 0.05 when the same distribution is used for both periods.
    - Exact rate ratio test (one-sided) for the comparison of two Poisson counts, taking into account the time at risk associated with each of the counts.
    - A modification of the exact rate ratio test using the one-sided mid-p-value.
- For the FWSI (calculated from the numbers of fatal and severe injuries):
    - Overlap of bootstrap ABC CI. As above.
    - Wilcoxon test, R default. As above.
    - Wilcoxon test, exact. As above.
    - Bootstrap ABC CI for ratio of means. As above.

- For the FWI (calculated from the numbers of all three types of injuries):
  - Overlap of bootstrap ABC CI. As above.
  - Wilcoxon test, R default. As above.
  - Wilcoxon test, exact. As above.
  - Bootstrap ABC CI for ratio of means. As above.
- For the numbers of fatal and severe injuries used separately:
  - Bayesian approach. Checks whether the posterior probability from a Bayesian model for the FWSI (Andrášik, 2019) is below the specified threshold. The threshold is chosen in such a way that the resulting probability of indicating a change is approximately 0.05 when the same distribution is used for both periods.

Results for the One-Sided Tests

The estimated probabilities of indicating a change are given in the following tables, for each of the five indicators considered. The scenarios used for each of the indicators are the same as in the two-sided case.

Personenunfall (accidents to persons):

	Scenario						
	reference	more events		more injuries per event		more severe injuries	
Summary measure / Test applied	**ref**	**rate*1.1**	**rate*1.3**	**mu*1.2**	**mu*1.5**	**p.fsl=0.025, 0.17,0.805**	**p.fsl=0.04, 0.2,0.76**
Number of events							
Overlap of bootstrap ABC CI	0.06089	0.28390	0.89291	0.06134	0.06029	0.06116	0.06091
Wilcoxon test, R default	0.04801	0.24498	0.86317	0.04801	0.04814	0.04889	0.04836
Wilcoxon test, exact	0.04840	0.24619	0.86401	0.04837	0.04858	0.04933	0.04868
CI for odds ratio	0.05160	0.26902	0.89690	0.05160	0.05160	0.05160	0.05160
Bootstrap ABC CI for ratio of means	0.06144	0.28591	0.89472	0.06169	0.06067	0.06148	0.06129
Wald test for Poisson GLM	0.05160	0.26902	0.89690	0.05160	0.05160	0.05160	0.05160
LR test for Poisson GLM	0.04910	0.26203	0.89265	0.04910	0.04910	0.04910	0.04910
Bayesian approach (threshold: 0.052)	0.05000	0.26505	0.89471	0.05000	0.05000	0.05000	0.05000
Exact rate ratio test assuming Poisson counts	0.04515	0.24997	0.88557	0.04515	0.04515	0.04515	0.04515
Rate ratio test assuming Poisson counts, mid-p	0.04980	0.26408	0.89426	0.04980	0.04980	0.04980	0.04980
FWSI							
Overlap of bootstrap ABC CI	0.04820	0.07875	0.17301	0.12054	0.30037	0.19007	0.56218
Wilcoxon test, R default	0.04847	0.09260	0.23883	0.14891	0.39802	0.20695	0.57828
Wilcoxon test, exact	0.04831	0.09245	0.23860	0.14872	0.39772	0.20669	0.57798
Bootstrap ABC CI for ratio of means	0.04697	0.07668	0.16854	0.11771	0.29367	0.18625	0.55552
FWI							
Overlap of bootstrap ABC CI	0.04867	0.08954	0.23088	0.14883	0.41555	0.18264	0.53732
Wilcoxon test, R default	0.04851	0.11751	0.37630	0.21091	0.58836	0.18699	0.51802
Wilcoxon test, exact	0.04851	0.11751	0.37630	0.21091	0.58836	0.18699	0.51802
Bootstrap ABC CI for ratio of means	0.04828	0.08875	0.22861	0.14727	0.41114	0.18124	0.53485
Numbers of fatal and severe injuries							
Bayesian approach (threshold: 0.089)	0.04991	0.08743	0.21093	0.14390	0.38704	0.21346	0.61616

Zusammenstoesse (collisions of trains):

	Scenario						
	reference	more events		more injuries per event		more severe injuries	
Summary measure / Test applied	**ref**	**rate*1.2**	**rate*1.5**	**mu*2**	**mu*3**	**p.fsl=0.05, 0.2,0.75**	**p.fsl=0.08, 0.25,0.67**
Number of events							
Overlap of bootstrap ABC CI	0.05409	0.16696	0.44856	0.05424	0.05372	0.05433	0.05404
Wilcoxon test, R default	0.04900	0.14887	0.40611	0.04925	0.04775	0.04905	0.04853
Wilcoxon test, exact	0.04758	0.14585	0.40164	0.04788	0.04623	0.04748	0.04684
CI for odds ratio	0.05212	0.16591	0.45693	0.05212	0.05212	0.05212	0.05212
Bootstrap ABC CI for ratio of means	0.05457	0.16786	0.45054	0.05497	0.05407	0.05469	0.05488
Wald test for Poisson GLM	0.05212	0.16591	0.45693	0.05212	0.05212	0.05212	0.05212
LR test for Poisson GLM	0.04660	0.15345	0.43751	0.04660	0.04660	0.04660	0.04660
Bayesian approach (threshold: 0.063)	0.04968	0.16119	0.45099	0.04968	0.04968	0.04968	0.04968
Exact rate ratio test assuming Poisson counts	0.03731	0.13002	0.39603	0.03731	0.03731	0.03731	0.03731
Rate ratio test assuming Poisson counts, mid-p	0.04797	0.15654	0.44218	0.04797	0.04797	0.04797	0.04797
FWSI							
Overlap of bootstrap ABC CI	0.05606	0.07813	0.11249	0.14435	0.22764	0.15358	0.23580
Wilcoxon test, R default	0.05826	0.08839	0.14207	0.13153	0.19328	0.10706	0.14532
Wilcoxon test, exact	0.04349	0.06858	0.11464	0.11288	0.17708	0.09428	0.13638
Bootstrap ABC CI for ratio of means	0.35378	0.34003	0.32041	0.35756	0.38575	0.39474	0.43557
FWI							
Overlap of bootstrap ABC CI	0.05669	0.08161	0.12221	0.15565	0.24854	0.13787	0.20947
Wilcoxon test, R default	0.05604	0.09965	0.18099	0.11204	0.15877	0.06419	0.07094
Wilcoxon test, exact	0.04895	0.08836	0.16539	0.10262	0.14774	0.05766	0.06530
Bootstrap ABC CI for ratio of means	0.18759	0.16235	0.15196	0.22516	0.28400	0.26505	0.32842
Numbers of fatal and severe injuries							
Bayesian approach (threshold: 0.058)	0.04975	0.06955	0.10380	0.19387	0.33243	0.14829	0.24606

Schiffe (ship accidents):

	Scenario						
	reference	more events		more injuries per event		more severe injuries	
Summary measure / Test applied	**ref**	**rate*1.5**	**rate*2**	**mu*2**	**mu*3**	**p.fsl=0.15, 0.25,0.6**	**p.fsl=0.25, 0.35,0.4**
Number of events							
Overlap of bootstrap ABC CI	0.05154	0.27397	0.59347	0.05117	0.05097	0.05140	0.05115
Wilcoxon test, R default	0.05272	0.25674	0.55477	0.05263	0.05275	0.05221	0.05275
Wilcoxon test, exact	0.04379	0.23674	0.53466	0.04379	0.04434	0.04359	0.04408
CI for odds ratio	0.05669	0.29585	0.62584	0.05669	0.05669	0.05669	0.05669
Bootstrap ABC CI for ratio of means	0.05172	0.27319	0.59269	0.05151	0.05153	0.05142	0.05121
Wald test for Poisson GLM	0.05669	0.29585	0.62584	0.05669	0.05669	0.05669	0.05669
LR test for Poisson GLM	0.04759	0.26712	0.59274	0.04759	0.04759	0.04759	0.04759
Bayesian approach (threshold: 0.074)	0.05051	0.28403	0.61757	0.05051	0.05051	0.05051	0.05051
Exact rate ratio test assuming Poisson counts	0.03353	0.21745	0.52953	0.03353	0.03353	0.03353	0.03353
Rate ratio test assuming Poisson counts, mid-p	0.04759	0.26712	0.59274	0.04759	0.04759	0.04759	0.04759
FWSI							
Overlap of bootstrap ABC CI	0.04476	0.08220	0.12403	0.11057	0.16959	0.10789	0.18034
Wilcoxon test, R default	0.06912	0.12529	0.19135	0.14103	0.19953	0.11351	0.16587
Wilcoxon test, exact	0.03054	0.06990	0.12478	0.08995	0.14679	0.07087	0.12473
Bootstrap ABC CI for ratio of means	0.36621	0.41918	0.43495	0.43466	0.46498	0.43524	0.48018
FWI							
Overlap of bootstrap ABC CI	0.06376	0.11670	0.17476	0.14683	0.21643	0.12363	0.19368
Wilcoxon test, R default	0.05714	0.14665	0.26795	0.12563	0.17290	0.06221	0.06901
Wilcoxon test, exact	0.04550	0.12360	0.23627	0.11129	0.16125	0.05541	0.06776
Bootstrap ABC CI for ratio of means	0.43673	0.38618	0.34570	0.44248	0.45919	0.49540	0.52838
Numbers of fatal and severe injuries							
Bayesian approach (threshold: 0.198)	0.05069	0.09500	0.15063	0.16212	0.27530	0.14906	0.26658

TechnDefekt_Nahv (technical defects in urban transport):

	Scenario						
	reference	more events		more injuries per event		more severe injuries	
Summary measure / Test applied	**ref**	**rate*1.5**	**rate*2**	**mu*2**	**mu*3**	**p.fsl=0.04, 0.08,0.88**	**p.fsl=0.06, 0.12,0.82**
Number of events							
Overlap of bootstrap ABC CI	0.05161	0.28258	0.61814	0.05138	0.05106	0.05095	0.05031
Wilcoxon test, R default	0.05182	0.26436	0.57708	0.05175	0.05147	0.05176	0.05127
Wilcoxon test, exact	0.04438	0.24551	0.55928	0.04405	0.04408	0.04439	0.04372
CI for odds ratio	0.05678	0.30716	0.65024	0.05678	0.05678	0.05678	0.05678
Bootstrap ABC CI for ratio of means	0.05226	0.28238	0.61775	0.05180	0.05136	0.05111	0.05092
Wald test for Poisson GLM	0.05678	0.30716	0.65024	0.05678	0.05678	0.05678	0.05678
LR test for Poisson GLM	0.04726	0.27670	0.61773	0.04726	0.04726	0.04726	0.04726
Bayesian approach (threshold: 0.073)	0.04915	0.29088	0.63972	0.04915	0.04915	0.04915	0.04915
Exact rate ratio test assuming Poisson counts	0.03339	0.22502	0.55568	0.03339	0.03339	0.03339	0.03339
Rate ratio test assuming Poisson counts, mid-p	0.04726	0.27670	0.61773	0.04726	0.04726	0.04726	0.04726
FWSI							
Overlap of bootstrap ABC CI	0.03310	0.05365	0.07480	0.07412	0.12079	0.07977	0.13022
Wilcoxon test, R default	0.08785	0.13614	0.18474	0.17710	0.26198	0.19120	0.29525
Wilcoxon test, exact	0.01684	0.03854	0.06683	0.06245	0.12592	0.07209	0.14753
Bootstrap ABC CI for ratio of means	0.23175	0.29876	0.34542	0.33851	0.39082	0.35175	0.40127
FWI							
Overlap of bootstrap ABC CI	0.07592	0.18921	0.31361	0.25025	0.38904	0.15033	0.22066
Wilcoxon test, R default	0.05199	0.25127	0.54649	0.15738	0.22099	0.06346	0.07276
Wilcoxon test, exact	0.04684	0.23980	0.53450	0.15122	0.21325	0.05920	0.06887
Bootstrap ABC CI for ratio of means	0.09419	0.19028	0.30540	0.25254	0.38082	0.17319	0.24257
Numbers of fatal and severe injuries							
Bayesian approach (threshold: 0.262)	0.04263	0.07507	0.11076	0.11353	0.19610	0.11969	0.20561

ZusammenstoesseZmZ (collisions of a train with a train):

	Scenario						
	reference	more events		more injuries per event		more severe injuries	
Summary measure / Test applied	**ref**	**rate*1.5**	**rate*2**	**mu*3**	**mu*5**	**p.fsl=0.05, 0.25,0.7**	**p.fsl=0.08, 0.35,0.57**
Number of events							
Overlap of bootstrap ABC CI	0.04902	0.20973	0.45384	0.04936	0.04881	0.04938	0.04925
Wilcoxon test, R default	0.05470	0.21135	0.43649	0.05539	0.05527	0.05586	0.05516
Wilcoxon test, exact	0.03911	0.17463	0.39473	0.03957	0.03961	0.03917	0.03908
CI for odds ratio	0.05309	0.22428	0.47821	0.05309	0.05309	0.05309	0.05309
Bootstrap ABC CI for ratio of means	0.04882	0.20863	0.45126	0.04917	0.04866	0.04907	0.04896
Wald test for Poisson GLM	0.05309	0.22428	0.47821	0.05309	0.05309	0.05309	0.05309
LR test for Poisson GLM	0.04801	0.20993	0.45699	0.04801	0.04801	0.04801	0.04801
Bayesian approach (threshold: 0.088)	0.05010	0.22255	0.48029	0.05010	0.05010	0.05010	0.05010
Exact rate ratio test assuming Poisson counts	0.03140	0.16088	0.39255	0.03140	0.03140	0.03140	0.03140
Rate ratio test assuming Poisson counts, mid-p	0.04790	0.20980	0.45698	0.04790	0.04790	0.04790	0.04790
FWSI							
Overlap of bootstrap ABC CI	0.05129	0.10094	0.15485	0.17824	0.26915	0.12422	0.18755
Wilcoxon test, R default	0.06106	0.13053	0.21505	0.13115	0.16529	0.09763	0.12203
Wilcoxon test, exact	0.03987	0.09677	0.17244	0.11764	0.16261	0.08168	0.11281
Bootstrap ABC CI for ratio of means	0.39712	0.40494	0.38714	0.46998	0.52533	0.44817	0.48440
FWI							
Overlap of bootstrap ABC CI	0.05707	0.11366	0.17669	0.19475	0.29052	0.11687	0.17124
Wilcoxon test, R default	0.05798	0.15019	0.27035	0.09078	0.10537	0.06326	0.06748
Wilcoxon test, exact	0.04745	0.12917	0.23876	0.09046	0.11080	0.05867	0.06713
Bootstrap ABC CI for ratio of means	0.38177	0.32763	0.28978	0.46251	0.53032	0.43393	0.47030
Numbers of fatal and severe injuries							
Bayesian approach (threshold: 0.061)	0.04986	0.09670	0.15863	0.29767	0.45297	0.16123	0.26564

Interpretation

For the number of events, several tests with a reasonable performance are available. The approaches based on the CI for the odds ratio and the Wald test in a Poisson GLM seem to perform best – they roughly respect the significance level of 0.05 under the reference distribution and very often have the highest power under increased event rates among all tests compared. The LR test for a Poisson GLM and the Bayesian approach also seem to reasonably respect the significance level and provide high power. The two versions of Wilcoxon's test seem to be a bit less powerful, but could still be used. The exact rate ratio test is often highly conservative (i.e., its rejection rate under the reference scenario is much lower than 0.05) and in these cases also has low power. As expected, the mid-p version of the rate ratio test is much less conservative, while it still seems to roughly respect the significance level; it seems to almost always reach higher power than both versions of the Wilcoxon test and to sometimes even outperform the Bayesian approach and/or the LR test for a Poisson GLM. The methods based on the overlap of the bootstrap ABC CI and on the bootstrap ABC CI for the ratio of the means tend to be liberal, mainly in the two-sided case, i.e., they give too many false alerts – so these should not be used in the two-sided case and used only with caution in the one-sided case. (For the overlap of the bootstrap ABC CI, the reason for the liberal behaviour in the two-sided case could be that the adjustment of the confidence interval is actually based on a one-sided motivation anyway.) Note that the estimates for the two most powerful approaches (based on the CI for the odds ratio and the Wald test in a Poisson GLM) are very often (but not always) identical; it turns out that the two approaches lead to the same decision in most of the cases.

For the FWSI, the two versions of Wilcoxon's test work roughly as desired in the two-sided case, but have relatively little power under changes of the event rate (which can be explained by the fact that many events do not result in fatal or severe injuries). When considering indicators with few events (Schiffe, TechDefekt_Nahv, ZusammenstoesseZmZ) in the one-sided case, the two versions seem to differ more – while the exact version can be quite conservative, the R default version is sometimes liberal (especially for TechnDefekt_Nahv). The bootstrap ABC CI for the ratio of the means is much too liberal in many indicators, both for the one-sided and two-sided cases. For "Personenunfall", the overlap method using the bootstrap ABC CI is also liberal in the two-sided case, which illustrates that at least for certain indicators, this method will not work as desired when used in the two-sided case (presumably for the same reason as conjectured in the case of the number of events). In the one-sided case, this approach provides a higher power than Wilcoxon's test under certain scenarios (e. g., more injuries per event or more severe injuries for "Zusammenstoesse").

For the FWI, the picture is similar as for the FWSI, except that the differences between the two versions of Wilcoxon's test are not as large, and that the Wilcoxon tests often also have very low power for detecting changes in the severity of injuries. This might partly be a problem of ranking. In addition, even in the one-sided setting, the overlap method using the bootstrap ABC CI is liberal for "TechnDefekt_Nahv".

The numbers of fatal and severe injuries are only used by the corresponding Bayesian test. While this is somehow a natural competitor to the tests using the FWSI, it uses more information than just the FWSI. In the two-sided case, it usually outperforms the two versions of Wilcoxon's test for the FWSI in detecting more injuries per event or more severe injuries, often quite dramatically; in the one-sided case, the picture is less clear. This approach could be attractive to

investigate changes in the quantities underlying the FWSI. A practical problem with this test is that one has to find a reasonable threshold to obtain a specified error rate under the reference distribution. Therefore, this test would probably need some fine-tuning for each application to a new indicator, and the determination of the threshold would then rely on the parametric model for the data, which is just an approximation of reality. This problem of the determination of a threshold arises as soon as one wants to control the probability of a false alarm, i. e., the (frequentist) type I error probability, which was done here to ensure comparability with the other approaches. If one is ready to decide on the action to be taken directly based on the (approximate) posterior probability of a deterioration of the safety level, then this problem is not relevant.

As a general note, given that each estimate of type I error rate or power is based on 100,000 simulated data sets, its standard error can be at most $\sqrt{0.5 \cdot 0.5/100{,}000} \approx 0.0016$; for a true rate of 0.05 (i. e., for a typical type I error rate), it reduces to $\sqrt{0.05 \cdot 0.95/100{,}000} \approx 0.00069$. Therefore, in particular, differences in the simulated powers of less than about 0.002 to 0.003 only provide very weak evidence of a true difference.

Finally, for the Bayesian approaches, the choice of the prior parameters for the simulation followed Andrášik (2019), i. e., $\alpha = 2$ and $\beta = 7$. Meanwhile, a slightly adapted choice has been proposed; for a reference period of four years and a current period of one year, it would result in a choice of $\alpha = 1.97$ and $\beta = 6.88$. The small difference in the prior parameters is not expected to substantially affect the results of the simulation.

Conclusions

In summary, for the number of events, the CI for the odds ratio and the Wald test in a Poisson GLM seem to be the best (and most often equivalent) options, followed by the Bayesian approach, the LR test for a Poisson GLM and the mid-p version of the rate ratio test; both versions of Wilcoxon's test are less powerful, but valid options. In the one-sided setting, the tests based on the overlap of the bootstrap ABC CI and on the bootstrap ABC CI for the ratio of means could be considered as well, but they can be liberal, depending on the kind of data analyzed.

For directly analysing the FWSI, Wilcoxon's test (again in both versions) seems to be the best option in the two-sided case, while it has some problems in the one-sided case for rare events. In the one-sided case, the test based on the overlap of the bootstrap ABC CI could again be considered as well, as it may be more powerful. The Bayesian approach based on the numbers of fatal and severe events could provide an interesting, sometimes more powerful alternative if the problem of determining a reasonable threshold can be solved or if the decision can be based on the posterior probability resulting from the Bayesian method.

For the FWI, no broadly applicable options other than both versions of Wilcoxon's test are currently available, and these cannot be expected to be powerful in detecting changes in the severity of injuries. The test based on the overlap of the bootstrap ABC CI could again be considered in the one-sided case, but may be quite liberal.

In certain cases, the behavior of the same test under similar scenarios, but for different indicators, is surprisingly variable. Even if the indicators used in the simulations were intended to provide a representative selection, it can therefore not be excluded that further undesirable properties of certain tests occur when these are applied to different indicators.

References

Andrášik, R. (2015): *A new method for evaluation of safety targets based on the odds ratio*, technical report for the Swiss Federal Office of Transport.
Andrášik, R. (2019): *Evaluation of safety targets based on Bayesian inference*, technical report for the Swiss Federal Office of Transport.
Vock, M. (2015): *Simulationen zur Validität unterschiedlicher methodischer Ansätze*, technical report for the Swiss Federal Office of Transport.

