

<p align="right">  
  
</p>

# Appendix - COVID-19 test fraud detection: Findings from a pilot study comparing conventional and statistical approaches

---

**Michael Bosnjak**<sup>1</sup>, **Stefan Dahm**<sup>2</sup>, **Ronny Kuhnert**<sup>2</sup>, **Dennis Weihrauch**<sup>3</sup>, **Angelika Schaffrath Rosario**<sup>2</sup>, **Julia Hurraß**<sup>3</sup>, **Patrik Schmich**<sup>2,4</sup>, **Lothar Wieler**<sup>2,5</sup> und **Johannes Nießen**<sup>3</sup>

<sup>1</sup> Trier University, Department for Psychological Research Methods |  
([corresponding author](#))  
<sup>2</sup> Robert Koch Institute, Department for Epidemiology and Health Monitoring  
<sup>3</sup> City of Cologne, Health Authority, Infectious and Environmental Hygiene  
<sup>4</sup> Federal Ministry of Health, Projekt Group BIPAM  
<sup>5</sup> Hasso Plattner Institute, Department Digital Global Public Health

## Zitieren

Bosnjak M, Dahm S, Kuhnert R, Weihrauch D, Schaffrath Rosario A, Hurraß J, Schmich P, Wieler L und Nießen J (2024): Appendix - COVID-19 test fraud detection: Findings from a pilot study comparing conventional and statistical approaches. [Dataset] Zenodo.  
DOI:[10.5281/zenodo.10608926](https://doi.org/10.5281/zenodo.10608926).

---

The methods and results of the publication "COVID-19 test fraud detection: Findings from a pilot study comparing conventional and statistical approaches" are described in more detail in this appendix. The R-syntax for the calculation is provided, as well as a pseudo data set with which the syntax can also be tested.

## Organisational and administrative information

---

The publication “COVID-19 test fraud detection: Findings from a pilot study comparing conventional and statistical approaches”, is a joint project of the Department for Psychological Research Methods - Trier University, Department 2 | Epidemiology and Health Monitoring - Robert Koch Institute, the Department Infectious and Environmental Hygiene - Health Authority of the City of Cologne and the Department Digital Global Public Health - Hasso Plattner Institute. The appendix presented here provides additional results and data for the publication and was curated by Department 2 | Epidemiology and Health Monitoring of the Robert Koch Institute. Questions regarding the content of the data can be addressed directly to the corresponding author Michael Bosnjak ([bosnjak@uni-trier.de](mailto:bosnjak@uni-trier.de)).

The publication of the data as well as the quality management of the (meta-)data is done by the department MF 4 | Research Data and Information Management. Questions regarding data management and the publication infrastructure can be directed to the Open Data Team of the Department MF4 at [OpenData@rki.de](mailto:OpenData@rki.de).

## Data schema and simulated data

---

We used data on claims for COVID-19 antigen tests submitted for reimbursement by 907 test centers operating in a German city with approximately one million residents for the timespan April 8, 2021 through August 28, 2022.

The data were transmitted on a daily basis via an online portal provided for this purpose by the ministry of a federal German state. Transmission was mandatory by law for the test centers.

- Rechtliche Grundlage verlinken

For each claim, the following information was provided: test center category (pharmacy, doctor's or dentist's office, private test center), date of testing, number of tests performed per day, number of positive tests per day.

## Data schema of the analysed data

Variable	Type	Values	Description
typ	string	Pharmacy , Doctors or dentists office , Private test center	Facility type (pharmacy, doctor's or dentist's office or private test center)
tnr	interger	1 ... 999	Testcenter identification number
date	date	yyyy-mm-dd	Date of the tests in ISO 8601 format
week_yr	string	ww_yyyy	Identifer for the calendar week (ww) of the Year (yyyy)
getestet	interger	$\geq 0$	Number of tests per day
positiv	interger	$\geq 0$	Number of positiv tests per day
investigation	interger	0 , 1	Indicator for investigation by the health authorities ( 1 = yes, 0 = no)

## Simulated data

Based on the data schema of the analysed data, we have simulated data for overall 608 test centers for which we assumed that they operated within the time period 2021-03-01 till 2021-12-31, containing 136,192 observations. Furthermore, we assumed that 15% of the test centers invoiced for fraudulent test results. For these 'criminal' test centers are simulated high numbers of tests, low positive rates, deviations from Benford's Law and deviations from the assumption of equally distributed last digits.

The resulting characteristics of simulated data are shown in table 1. The simulated data does not contain any information from the original datas and can be used to test our provided R-scripts used in the analysis.

**Table 1: Characteristic of simulated data**

Facility type	
test centers [N]	
test centers suspected of fraud [N]	
Tests [N]	
Tests per day	
Positive tests [%]	
Mean	
Max	

```

<td colspan="1" valign="bottom">median</td></tr>
<tr><td colspan="1">Pharmacy</td>
<td colspan="1">290</td>
<td colspan="1">4</td>
<td colspan="1">6795707</td>
<td colspan="1">137.5</td>
<td colspan="1">1387</td>
<td colspan="1">98</td>
<td colspan="1">1.42</td></tr>
<tr><td colspan="1">Doctor's or dentist's office</td>
<td colspan="1">357</td>
<td colspan="1">6</td>
<td colspan="1">8312270</td>
<td colspan="1">141.4</td>
<td colspan="1">1663</td>
<td colspan="1">110</td>
<td colspan="1">1.75</td></tr>
<tr><td colspan="1">Private test center</td>
<td colspan="1">160</td><td colspan="1">3</td>
<td colspan="1">4457176</td>
<td colspan="1">159.3</td>
<td colspan="1">2019</td>
<td colspan="1">113</td>
<td colspan="1">1.16</td></tr>
<tr><td colspan="1">Total</td>
<td colspan="1">608</td>
<td colspan="1">13</td>
<td colspan="1">19565153</td>
<td colspan="1">143.7</td>
<td colspan="1">2019</td>
<td colspan="1">106</td>
<td colspan="1">1.50</td></tr>
</table>

```

## Methods and Results

---

We used four statistical methods to detect fraud, which are described in the following sections.

- Outlier identification from the mean number of tests per day invoiced (high number of tests)
- Deviations from the the mean positive rate as identified via Poisson regression (low positive rate)
- Deviations from Benford's Law
- Deviations from the assumption of equally distributed last digits

## Outlier identification from the mean number of tests per day invoiced (high number of tests)

In our first statistical approach aimed at identifying disproportionately high test volumes invoiced, the numbers of tests invoiced per day are classified as conspicuous if they fall outside the 90% percentile in terms of the mean number of tests performed per day within a test center category.

### Analysis script

The analysis on outlier identification from the mean number of tests per day invoiced was performed using an R script. The content of the script is provided in as R file:

[supporting\\_material/scripts/Fraud\\_Methods\\_Description.r](#)

Figure 1 shows the corresponding distributions resulting from the analysis.

**Figure 1: Histograms of the mean number of tests performed per day (x-axis) by test center type (pharmacies, doctor's or dentist's offices, private test centers). The dashed vertical line indicates the 90% percentile of each distribution. Test centers falling on the right sides of these lines are considered statistically conspicuous. The numbers above the bars indicate the number of test centers within each bar. (Dateiname.jpg)**



A total of 91 testing centers (6 pharmacies, 39 physician practices/dentists, and 46 private testing sites) were classified as suspicious using this approach. Table 1 shows the basic statistics of tests performed per day, divided into conspicuous by the statistical methods.

**Table 2: Basic statistics of the mean number of tests per day and test centers by facility type, non suspected and suspected of fraud by the conventional approach**

Facility type			Statistically conspicuous			Statistically not conspicuous			Total		
N			Median			N			Median		
Max			N			Median			Max		
N			Median			N			Median		
Max			N			Median			Max		
Pharmacy											

```

<td colspan="1">6</td>
<td colspan="1">252.2</td>
<td colspan="1">679.3</td>
<td colspan="1">54</td>
<td colspan="1">64.1</td>
<td colspan="1">161.5</td>
<td colspan="1">60</td>
<td colspan="1">71.0</td>
<td colspan="1">679.3</td></tr>
<tr><td colspan="1">Doctor's or dentist's office</td>
<td colspan="1">39</td>
<td colspan="1">51.7</td>
<td colspan="1">1032</td>
<td colspan="1">351</td>
<td colspan="1">3.5</td>
<td colspan="1">23</td>
<td colspan="1">390</td>
<td colspan="1">3.8</td>
<td colspan="1">1032</td></tr>
<tr><td colspan="1">Private test center</td>
<td colspan="1">46</td>
<td colspan="1">515.4</td>
<td colspan="1">5520</td>
<td colspan="1">411</td>
<td colspan="1">106.6</td>
<td colspan="1">342.2</td>
<td colspan="1">457</td>
<td colspan="1">116.3</td>
<td colspan="1">5520</td></tr>
</table>

```

## Deviations from the mean positive rate as identified via Poisson regression (low positive rate)

The positive rates were modeled by a Poisson regression model with random effects using the logarithms of the number of positive tests per day and per test center as dependent variable and the logarithms of the respective total number of tests as offset. The variability between the test centers were modeled by random intercepts for test centers. In addition, to account for changes in the positive rates over time for example induced by changing incidence, calendar week specific random intercepts were introduced in the model. Differences in positive rates between the facility types were controlled by fixed effects.

$$\log(pos_{ijkl}) = \log(test_{ijkl}) + A + \beta_j + \gamma_k + \alpha_l \cdot typ_l + \varepsilon_{ijkl}$$

$pos_{ijkl}$  : Number of positive tests at day  $i$  in test center  $j$  ( $j = 1, \dots, 907$ ) the week  $k$  ( $k=1, \dots, 73$ ) and in facility type  $l$  ( $l = 1, 2, 3$ )  
 $test_{ijkl}$  : Number of tests at day  $i$  in test center  $j$  in week  $k$  in facility type  $l$   
 $A$  : Overall mean

$\beta_j$ : Test center – specific deviation from the global intercept (random effect)  
 $\gamma_k$ : Calendar week specific deviation from the global intercept (random effect)  
 $\alpha_l$ : Facility type : Pharmacy, doctor's or dentist's office or private test center (fixed effect)  
 $x_l$ : Factor accounting for differences in positive rates by facility type (fixed)  
 $\epsilon_{ijkl}$ : Residual at day i in test center j in week k and facility type l not explained by the regression.

## Analysis script

The analysis on deviations from the mean positive rate as identified via Poisson regression was performed using an R script. The content of the script is provided in as R file:

[supporting\\_material/scripts/Fraud\\_Methods\\_Poisson\\_Regression.r](#)

Corona tests were performed in the time span April 8, 2021 through August 28, 2022 on 73 weeks respective 508 days in 907 test centers, but not all centers operated for the entire period. This resulted in a total of  $N = 118,908$  positive rates.

**Table 3: Statistics of estimated fixed effects**

Variable	Estimate	Standard Error	P-value
$A$ (global intercept)	-5.004	0.257	< 0.0001
$typ_2$ (Facility type: doctor's or dentist's office)	0.149	0.192	0.45
$typ_3$ (Facility type: private test center)	-0.477	0.182	0.0086

The differences between the mean positive rates by facility type (Table 2) were significant in the regression at  $P = 0.009$ . In particular, the rate of positive tests was lower for the private test centers than for the pharmacies or the physicians' practices. The estimated random intercepts for the 907 test centers ( $\beta_j$ ) were centered by 0 and had a variance of 1.65. The variance of the 73 week specific random intercepts ( $\gamma_k$ ) was estimated to be 2.67. Both variables ( $\beta$  and  $\gamma$ ) were significant with  $P < 0.0001$ .

A low center specific random intercept  $j$  indicates a low mean positive rate for the tests in resp. center. Therefore, the reporting of tests conducted by a test center was considered conspicuous if its estimated random intercept was significantly low. The estimated test center intercepts ( $\beta_j$ ) and their standard deviations  $sd(\beta_j)$  were used to generate test values comparable to the t-values of the t-test:

$$r_j = \frac{\beta_j}{sd(\beta_j)} \quad , j = 1, \dots, 907$$

The test values  $r_j$  ranged from -23.0 to 49.0 corresponding to positive rates of 0.5% resp. 10.6%. A value of  $r_j < -6$  was regarded as significant. According to this criterion, the 907 test centers could be classified to 88 conspicuous and 819 not conspicuous test centers (s. Table 3), where the mean positive rate in conspicuous test centers amounted to 0.6% and the not conspicuous test centers had a mean positive rate of 2.3%.

**Table 4: Summary of classifications into statistical suspicious versus non spicious test centers according to the Poisson regression model used**

Facility type	Statistically conspicuous [N]/[Positive rate %]	Statistically not conspicuous [N]/[Positive rate %]	Total [N]/[Positive rate %]
Pharmacy	11 / 0.88	49 / 2.82	60 / 2.44
Doctor's or dentist's office	16 / 0.59	374 / 3.95	390 / 2.75
Private test center	61 / 0.54	396 / 2.24	457 / 1.95
Total	88 / 0.58	819 / 2.33	907 / 2.02

## Deviations from Benford's Law

### Requirements for Benford Analysis:

For data to undergo Benford analysis, it must exhibit a specific range and be reported over a certain number of days. The Benford principle may be compromised if test centers frequently operate at maximum capacity, leading to the reporting of similar test numbers. This can lead to a false positive result for a test center according to Benford's Law. Test centers reporting tests on only a few days are ineligible for Benford analysis because the distribution of the first digit lacks the necessary variability for evaluation. To address this, we stipulate that the number of tests must be reported over a minimum of 30 days.

### Analysis script

The analysis for Deviations from Benford's Law was performed using an R script. The content of the script is provided in as R file:

[supporting\\_material/scripts/Fraud\\_Methods\\_Benford.r](#)



Figure 2 shows the distribution of the leading digit according to Benford's law and the distribution over all data available for the observation period. Overall, there is good agreement. The leading 1 as well as the 2 occur slightly disproportionately according to Benford's law.

**Figure 2. Distribution of leading digit of total reporting numbers (line) versus expected values of Benford's law (bars).**



A chi-square test is calculated for each of these 665 test centers. The chi-square test value determines the degree of deviation.

**Figure 3: Illustration of the distribution of the leading digit of the five test centers (lines) with the largest deviations from Benford's law.**



In Table 4, we have summarized the number of test centers classified by conventional methods and Benford's Law. The threshold for test centers considered to be conspicuous according to Benford's Law was set to those 10% with the largest chi-square test value.

**Table 4. Number of test centers by facility type, (non) suspected of fraud by the conventional approach, and (non) suspected of fraud by the statistical approach focusing on the deviation from Benford's law.**

		Facility type		Suspected of fraud by the health authorities (conventional approach)		Not suspected of fraud by the health authorities (conventional approach)	
	Statistically conspicuous		Statistically not conspicuous		Total		
	Statistically conspicuous		Statistically not conspicuous		Total		
Pharmacy	0	0	0	5	55	60	
Doctor's or dentist's office							

<td colspan="1">4</td>
<td colspan="1">4</td>
<td colspan="1">8</td>
<td colspan="1">197</td>
<td colspan="1">205</td>
</tr>
<tr><td colspan="1">Private test center</td>
<td colspan="1">10</td>
<td colspan="1">65</td>
<td colspan="1">75</td>
<td colspan="1">44</td>
<td colspan="1">227</td>
<td colspan="1">321</td>
</tr>
<tr><td colspan="1">Total</td>
<td colspan="1">10</td>
<td colspan="1">69</td>
<td colspan="1">79</td>
<td colspan="1">57</td>
<td colspan="1">529</td>
<td colspan="1">586</td>
</tr>
</table>

Based on table 4, the percentage of positive overlap between the traditional and Benford's law-based methods amounts to 12.7% (10/79), the percentage of negative overlap 90.3% (529/586), and the share of incrementally identified potentially fraudulent test centers identified by Benford's law which were undetected by traditional approaches amounts to 9.7% (57/586) related to all test centers.

In contrast to the publication, only the test sites that could also be evaluated by the method are considered here. This means that the positive, negative overlap and the proportion of undetected by traditional approaches are higher than in the results in the publication.

## Deviations from the assumption of equally distributed last digits

### Requirements for Last Digit Analysis:

Similar to the Benford approach, the last digit method requires data with a sufficient range for evaluating the distribution of the last digit. A minimum number of daily reports ( $\geq 30$ ) is essential for this evaluation. Unlike Benford's law, the distribution of the last digit includes zero. However, in the single-digit range 0-9, testing centers were not obligated to report "0" on days without testing, leading to a systematic underrepresentation of zero. To mitigate this bias, only test reports with more than 9 tests per day are considered for the distribution. In summary, for the last digit method evaluation, a test center must have reported more than 9 tests per day for at least 30 days. This method could only be applied to 512 of the 907 test sites.

## Analysis script

The analysis on deviations from the assumption of equally distributed last digits was performed using an R script. The content of the script is provided in as R file:

[supporting\\_material/scripts/Fraud\\_Methods\\_Last\\_Digit.r](#)

Figure 4 shows the distribution of the last digit comparing the expected distribution versus all reported data in the observation period. Overall, there is good agreement. The last digits 1 and 2 seem to be slightly overrepresented.

**Figure 4. Distribution of the last digit of the total number of reports (line) versus the expected distribution (bar).**



A chi-square test is calculated for each of the 512 test centers. The chi-square test value determines the degree of deviation.

The five test centers with the greatest deviation from the expected distribution can be seen in Figure 5. It seems that the zero and the number 5 were deliberately avoided as the last digit in the test center 48 and 68. Test center 486 reported only an average of 11 tests per day and all numbers below 10 are not considered. Therefore, the distribution has only limited informative value.

**Figure 5. Distribution of the last digit of the five test sites (line) with greatest deviation from the expected distribution (bar).**



The threshold for test centers considered to be conspicuous according to the Last Digit Law was set to those 10% with the largest chi-square test value, yielding 52 test centers. In Table 5, we have summarized the number of test centers classified by traditional methods and the Last Digit Law.

**Table 5. Number of test centers by facility type, (non) suspected of fraud by the conventional approach, and (non) suspected of fraud by the statistical approach focusing on the deviation from the law of equally distributed last digits.**

Facility type		Suspected of fraud by the health authorities (conventional approach)		Not suspected of fraud by the health authorities (conventional approach)	
Statistically conspicuous					
Statistically not conspicuous					
Total					
Statistically conspicuous					

	Statistically not conspicuous	
Total		
Pharmacy		
0		
0		
0		
1		
58		
59		
Doctor's or dentist's office		
1		
3		
4		
14		
48		
62		
Private test center		
7		
68		
75		
29		
283		
312		
Total		
8		
71		
79		
44		
389		
433		

Based on table 5, the percentage of positive overlap between the traditional and Last-Digit-Law-based methods amounts to 10.1% (8/79), the percentage of negative overlap 89.8% (389/433), and the share of incrementally identified potentially fraudulent test centers identified by the last digit law which were undetected by traditional approaches amounts to 10.2% (44/433).

As with Benford, only the test sites that could also be assessed by the method are considered here. This means that the positive, negative overlap and the proportion of undetected by traditional approaches are higher than in the results in the publication.

## Comparison and combination of the four statistical methods used

If all test sites that are conspicuous by at least one statistical method are considered, the positive overlap also increases by 50.5% and the negative overlap decreases to 78.1%. If test sites are considered to be conspicuous in the combination of at least 2 methods, the positive and negative overlap is comparable to that of the individual methods.

**Table 6 Positive and negative overlap and share of incrementally identified potentially fraudulent test centers by statistical approaches**

Statistical approach	Positive overlap	Negative overlap	Share of incrementally identified potentially fraudulent test centers
At least identified by one method	50.5% (47/93)	78.1% (636/814)	21.9% (178/814)
At least identified by two methods	9.7% (9/93)	93.6% (762/814)	6.4% ((61-9)/814)

## Content and structure of the appendix data

The tables, figures and their data, R-scripts and simulated data described in the section [Methods and Results](#) are made available in the appendix as open data. The following section describes the structure of the dataset in more detail.

### Data Tables

The results of the analyses conducted are provided as data tables. The files are named as the corresponding tables in the [Methods and Results](#) section as `table_1.tsv`, etc.

[table\\_1.tsv](#)

[table\\_2.tsv](#)

[table\\_3.tsv](#)

### Data Formatting

The data is provided in tab-separated .tsv files. The character set of the .tsv file is UTF-8. The delimiter between individual values is a tab [tab].

- Character set: UTF-8

- .tsv delimiter: Tab [tab]

## Supporting Materials

The Supporting Materials folder contains

- All figures used in the appendix
- The underlying data for the figures
- The R scripts used for the analysis
- Sample data for testing the analysis scripts

### Figures and Figure Data

The figures listed in the appendix and their underlying data are provided in the "Figures" subfolder as .png and .tsv files.

[supporting\\_material/figures](#)

The figures are named according to the numbering in the appendix as `figure_1.png`, etc. The underlying data for the figures are provided following the naming of the figures as `figure_1_data.tsv`, etc. The data schemas of the figure data are explained below.

#### Data Schema Figure 1

Variable	Type	Variations	Description
testcenter_id	integer	1 ... 907	Test center identification number
type	string	Pharmacy , Doctors or dentists office , Private test center	Facility type (pharmacy, doctor's or dentist's office, or private test center)
mean	float		Text
p90	float		Text
suspicious_des	integer	0 or 1	Text

#### Data Schema Figure 2 and Figure 4

Variable	Type	Variations	Description
leading_figure	integer		Text
theoretical_percent	float		Text
observed_percent	float		Text

#### Data Schema Figure 3 and Figure 5

Variable	Type	Variations	Description
testcenter_id	integer	1 ... 907	Test center id
leading_figure	integer		Text
theoretical_percent	float		Text
observed_percent	float		Text

## R Scripts

The analysis scripts for the individual methods are provided as .r files in the following subfolder:

[supporting\\_material/scripts](#)

The naming of the scripts is based on the methods used and is linked in the "Analysis script" sections.

## Simulated data

Simulated data for testing the evaluation R-scripts are provided. These correspond to the data schema described in the section [Data schema of the analysed data](#).

The simulated data are available in the folder supporting\_material in the file

[simulated\\_data.csv](#)

Further information on the simulated data can be found in the corresponding [section](#).

## Metadata

To increase findability, the provided data are described with metadata. The Metadata are distributed to the relevant platforms via GitHub Actions. There is a specific metadata file for each platform; these are stored in the metadata folder:

#### Metadata/

Versioning and DOI assignment are performed via [Zenodo.org](https://zenodo.org). The metadata prepared for import into Zenodo are stored in the [zenodo.json](#). Documentation of the individual metadata variables can be found at <https://developers.zenodo.org/representation>.

#### Metadata/zenodo.json

The zenodo.json includes the publication date ( "publication\_date" ) and the date of the data status:

```
"dates": [  
  {  
    "start": "2023-09-11T15:00:21+02:00",  
    "end": "2023-09-11T15:00:21+02:00",  
    "type": "Collected",  
    "description": "Date when the Dataset was created"  
  }  
],
```

## Guidelines for Reuse of the Data

---

Open data from the RKI are available on [Zenodo.org](https://zenodo.org), [GitHub.com](https://github.com), [OpenCoDE](https://gitlab.opencode.de), and [Edoc.rki.de](https://edoc.rki.de):

- <https://zenodo.org/communities/robertkochinstitut>
- <https://github.com/robert-koch-institut>
- <https://gitlab.opencode.de/robert-koch-institut>
- <https://edoc.rki.de/>

## License

The "Appendix - COVID-19 test fraud detection: Findings from a pilot study comparing conventional and statistical approaches" dataset is licensed under the [Creative Commons Attribution 4.0 International Public License | CC-BY](#).

The data provided in the dataset are freely available, with the condition of attributing the Robert Koch Institute as the source, for anyone to process and modify, create derivatives of the dataset and use them for commercial and non-commercial purposes.

Further information about the license can be found in the [LICENSE](#) or [LIZENZ](#) file of the dataset.