VILNIUS TECH
Vilnius Gediminas Technical University

12th International Scientific Conference

BUSINESS AND MANAGEMENT 2022

May 12–13, 2022, Vilnius, Lithuania

ISSN 2029-4441 / eISSN 2029-929X
ISBN 978-609-476-288-8 / eISBN 978-609-476-289-5
Article Number: bm.2022.725
https://doi.org/10.3846/bm.2022.725

ADVANCED ECONOMIC DEVELOPMENT *http://vilniustech.lt/bm*

EVALUATION OF THE MATHEMATICAL MODELLING METHODS AVAILABLE IN THE MARKET

Vaidas GAIDELYS*, Emilija NAUDŽIŪNAITĖ

School of Economics and Business, Kaunas University of Technology, Donelaičio g. 73, Kaunas, Lithuania

Received 25 January 2022; accepted 13 April 2022

Abstract. The major purpose of this research is to analyse and select the relevant mathematical modelling methods that will be employed for developing an algorithm. To fulfil the major purpose, three following objectives were raised. First, to select and substantiate the most common mathematical modelling methods. Second, to test the pre-selected methods under laboratory conditions so that the most relevant method for implementing the target project could be identified. Third, to prepare at least 3 models for application. The research results indicate that when evaluating the respiratory virus (SARS-CoV-2 causing COVID-19) concentration and survival rate dependence on a number of traits, the methods of descriptive statistics, confidence intervals, hypothesis testing, dispersion analysis, trait dependence analysis, and regression analysis are employed. All the above-listed methods were tested under laboratory conditions and thus can be applied to evaluate the effectiveness of the project product – a device designed to prevent transmission of respiratory viruses through air droplets. Selection of a particular method depends on a set of traits to be analysed, a trait type (quantitative, qualitative), a trait distribution type, and parameters. In the context of COVID-19, there is an urgent need to bring new products to market. Since most of the new products developed are directly related to research, it is very important to calculate the algorithms required to provide the service. Therefore, in order to calculate the optimal algorithm, it is necessary to analyze the algorithms already on the market. In this way, the products developed can gain a competitive advantage over competitors' products. Given that the equipment placed on the market will be equipped with HINS radiation sources, such a product will become original and new on the market. Therefore, it is necessary to evaluate several methods of mathematical modelling. It is also necessary to take into account that the placing on the market of a product takes place in the context of global competition.

Keywords: respiratory viruses, COVID-19, dentistry, diagnostics, virus spread prevention, virus survival, ultraviolet rays, clinical trials, modelling methods.

JEL Classification: 014, 03, 04.

Introduction

The spread of COVID-19 is associated with transmission of aerosols and droplets containing severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) to healthy subjects through breathing, talking, sneezing and coughing. When providing a variety of dental services, dentists work close to a patient's face and often use high-speed dental devices that aerosolize body fluids such as saliva and blood, which puts these professionals at a significantly increased risk of getting infected with or transmitting SARS-CoV-2. When using water-cooled, rotating or vibrating devices in the oral cavity, together with a patient's normal breathing, coughing and sneezing, aerosols and droplets ranging in size from small inhalable aerosols to large droplets that can contaminate nearby surfaces and personal protective equipment (PPE) are formed. Due to the low settling rate, smaller aerosols remain in the air for a long time and can be inhaled into the upper and lower respiratory tract. Droplets larger than 10 micrometres in diameter do not remain long in the air and settle quickly on nearby surfaces. Although they can be inhaled into the upper respiratory tract, they are less likely to reach the lower respiratory tract, but can cause surface contamination or direct contact with open mucous membranes. Larger droplets can carry larger amounts of the virus than smaller droplets or aerosols. Thus, dentists need good protection from inhaled aerosols and larger droplets. To ensure the safety of patients

* Corresponding author. E-mail: *Vaidas.Gaidelys@ktu.lt*

and dentists, it is necessary to use measures to limit the generation and inhalation of aerosols and droplets. Masks and respirators are the main measures used to restrain the release and inhalation of aerosols, but they are insufficient. Scientific literature review (Derruau et al., 2021) provides answers to 72 questions related to dentistry and Covid-19 disease. The answers were developed by 11 authors after analysing the information provided in 377 scientific publications. More than 90% of the publications were issued in 2020–2021. The authors answered the questions on a variety of relevant topics, including epidemiology, virology, immunology, diagnostics, oral SARS-CoV-2 transmission, clinical trials for COVID-19 treatment, current treatment options, vaccines, infection prevention and control options in the dental practice. It was emphasized that new protective measures, devices and physical barriers needed to be developed to more reliably prevent transmission of the virus when providing dental services.

A variety of mathematical models can be employed to evaluate the effectiveness of new anti-viral measures. The report focuses on the use of ultraviolet (UV) radiation to prevent virus transmission through air droplets when providing dental services; the results of the report will be employed for developing a new UV-based respiratory virus eradication device in the context of the project "Preventing transmission of respiratory viruses (COVID-19) through air droplets when providing dental services" (No. 01.2.1-LVPA-T-858-01-0025). Therefore, the aim of this study is to select the most appropriate mathematical modelling methods for evaluating the effectiveness of a device developed to prevent transmission of respiratory viruses (SARS-CoV-2 causing COVID-19 disease) through air droplets when providing dental services.

Because the newly developed product will be placed on the market in the face of global competition. It is important to evaluate the mathematical modelling methods that will be used to develop the new algorithm. It should be noted that the market segmentation of R&D products entering the market under COVID-19 disease differs.

1. Literature background

Length SARS-CoV-2 is transmitted to humans through air droplets and aerosols, as well as contact routes. In both cases, the virus spreads from an infected patient's nose and / or mouth by breathing, talking, and sneezing. During respiration, warm (36 °C) and moist (6.2% of water) gases in the alveoli rise to the mouth and nose, where they cool and condense before being ejected (0.6–1.4 m/s) through the airways. The droplets emitted (0.8–1 μm in diameter) contain water and mucous particles from the alveoli and upper respiratory tract, which may carry viruses. The droplets form a bio-aerosol and can infect people nearby, and also linger in the air (Derruau et al., 2021). Aerosols disperse in the air at a distance of 0.5–3 m. Thus far, the question of the virus viability and concentration in the air is open (Abkarian et al., 2020). When sneezing, air expulsion is stronger (up to 13 m/s), so the air carries a larger amount of nasal/oral mucosal infectious substances, forming large droplets up to 100 μm in diameter (Derruau et al., 2021). Within a few milliseconds, the droplets diffuse at a distance of 0.7 m. The heaviest particles fall down, contaminate surfaces and become phomits (dead objects containing disease-causing bacteria). Within 10–20 s, the largest droplets lose water through evaporation, usually at low relative humidity and high air temperature. The small particles with a small amount of water can linger in the air for many hours or even days. As a result, viral air contamination can get worse over time, usually in enclosed spaces without adequate ventilation. Airborne viruses settle directly into the human airways. Being highly contagious, airborne viruses are an important route for disease transmission (Zhang et al., 2020).

To evaluate the stability of the viruses, a solution containing SARS-CoV-2 viruses was sprayed on various surfaces (Van Doremalen et al., 2020). The stability of the viruses on plastic and stainless steel surfaces was greater, 72 h and 48 h respectively, than on copper and cardboard, 4 h and 24 h respectively. Another study revealed that protective masks could be contaminated with SARS-CoV-2 for several days (Chin et al., 2020). These results confirm the likelihood of the infection through contact with equipment and dental tools because the virus can remain viable for several days on plastic and steel surfaces which are often used in medicine (Chin et al., 2020).

Literature offers many recommendations on how to reduce the risk of the infection when providing dental services. Patients with suspected and confirmed COVID-19 disease are recommended to refuse aerosol-generating procedures (AGPs). When AGPs are required for dental treatment and cannot be delayed, the risk can be reduced by rinsing the mouth, insulating the spaces with rubber curtains, and employing intensive saliva suction. In case an AGP is performed, the dental office must be naturally or mechanically ventilated before admitting a new patient (Centers for Disease Control and Prevention, 2020). Proper ventilation of the dental office by bringing in fresh and clean outdoor air can play an important role in preventing airborne infections by reducing the concentration of infectious respiratory aerosols in the indoor air. There are three types of ventilation: natural (window), mechanical and mixed ventilation. Dental offices are recommended at least 6 (ideally 12) air changes per hour (Centers for Disease Control and Prevention, 2020). The World Health Organization (WHO) recommends an average natural ventilation rate of ≥60 L/s/per patient or ≥12 air changes per hour with mechanical room ventilation (World Health Organization [WHO], 2020).

Based on Italian experience, dental practice-related risks and recommendations for doctors working under conditions of the COVID-19 pandemic are provided in article (Izzetti et al., 2020). It emphasizes the need to develop new treatments and safe service provision

protocols that would significantly reduce the amount of droplets, aerosols and patient contact. Standard protective measures are insufficient to protect against the effects of aerosols and droplets. The immediate risk of inhalation is usually associated with the use of hand tools and ultrasonic scalers which generate aerosols and droplets that are often mixed with saliva and blood. It is recommended to: 1) avoid and reduce the use of nozzles so as to diminish aerosol and/or droplet generation, and use hand-pieces with antiretractive or antireflux valves instead; 2) place a rubber cover to significantly reduce the dispersion of aerosol and droplets in the treatment room; 3) employ surgical aspiration to control the dispersion of airborne particles; 4) use measures to reduce the risk of saliva stimulation and cough. The analysis of 22 studies revealed that human coronaviruses such as SARS and MERS can linger on surfaces for up to 9 days, but can be effectively eradicated within one minute by disinfection with 62–71% ethanol, 0.5% hydrogen peroxide, and 0.1% (1 g/l) sodium hypochlorite (Ge et al., 2020). The article also presents the strategies for reducing droplet formation while providing various dental services. For instance, in the case of endodontics, a rubber curtain should be affixed; in the cases of dental filling and pediatric dentistry, such measures as limited use of rotating tools, chemical treatment of decay, insulating rubber curtains, and active disinfection of surfaces are recommended.

What dental instruments and tools generate the largest amounts of aerosols and droplets? This question was partly answered in the review published at the end of 2020 (Innes et al., 2020). Its purpose is to identify which dental instruments and tools generate droplets and aerosols, and to describe droplet spread and deposition. When preparing the review, the authors were looking for relevant articles published before 2020-08-20 in the following databases: "Medline" (OVID), "Embase" (OVID), "Cochrane", "Scopus", "Web of Science" and LILACS. A total of 83 studies met the inclusion criteria and covered: ultrasonic scalers (USS, n = 44), high-speed air rotors (HSAR, n = 31), oral surgical instruments (n = 1), low speed hand tools (n = 4), air-water syringes (n = 4), polishing tools (n = 4), prophylactic tools (n = 2), and hand scalers (n = 2). When conducting experiments with the above-mentioned tools, they were found to substantially contaminate a dentist's waist, hand, and a patient's body. The heterogeneity of the studies did not allow for accurate comparisons, but a hierarchy of contamination risks was developed: higher risks (USS, HSAR, air and water syringes, polishing, extraction by employing motorized hand tools); medium risk (low speed hand tools, prophylactic tools, extraction devices); lower risks (water syringes and hand scalers). The authors of the review state that due to heterogeneity of the studies, only qualitative conclusions can be drawn and suggest to continue with researching and developing standardized methodologies that would facilitate the synthesis of this type of research.

Articles (Lindsley et al., 2020; Wilson et al., 2020; Morawska et al., 2020) discuss the relevant scientific issues that need to be addressed in the nearest time. The purpose of the research introduced in article (Bizzoca et al., 2020) is to rank the risk of routine dental procedures, taking into account the threats of the SARS-CoV-2 virus. After ranking the risk of infection, the teams of dentists in each dental practice are provided the recommendations on procedure safety and personal protective equipment (PPE). Taking into account the risk of the virus transmission, the authors of this research analysed 42 routine dental procedures. Risk ranks were estimated for each procedure based on the following ranking system: direct contact with saliva (1 point), direct contact with blood (2 points), low spray/aerosol quantities transmitted through air and water syringes (3 points), high spray/aerosol quantities produced by rotating, ultrasonic and piezoelectric tools (4 points); and duration of a procedure. By employing this new risk ranking system, the authors categorised various dental procedures by their risk ranks: low (1–3), medium (4–5), and high (≥6). The safety protocol for each procedure was aligned with the estimated risk rank. Risk ranks also contributed to selection of the relevant PPE. Considering the major purpose of the report, the following sections address only eradication of viruses by employing ultraviolet radiation.

The effects of ultraviolet (UV) rays on microorganisms have been analysed for many years. They eradicate a wide variety of bacteria, viruses, fungi, and other microorganisms that lose their ability to regenerate or mutate. A 2021 review of 377 dentistry and COVID-19 related scientific papers, published over the last year and a half, provides: "... non-contact disinfection technology, such as ultraviolet radiation or evaporated hydrogen peroxide, can complement, but not replace, manual surface cleaning for virus removal. The efficacy of alternative disinfection methods (e.g. ultrasonic waves, UV radiation and blue LED light) against SARS-CoV-2 has not been sufficiently studied". In this section of the report, we review the properties of UV light and scientific articles that apply UV rays to eradicate respiratory viruses (SARS-CoV-2 causing COVID-19 or related viruses) and analyse the effectiveness of the measures or devices used.

The total dose, to which the irradiated aerosol virus is exposed, is equal to the product of the UVG irradiation intensity and the irradiation duration (Tseng & Li, 2005). The dependence of the part of the viruses remaining after UVG irradiation on irradiation duration is called the viral survival function.

The data for estimating the viral survival function is collected by employing biosamplers; then the presence of SARS-CoV-2 virus is determined from the samples, and viability of the virus is measured in TCID50 units or the virus concentration PFU/ml = 0.7·TCID50 (Kenarkoohi et al., 2020; Van Doremalen et al., 2020).

Linear regression methods are used to construct the viral survival function, or if the data do not meet the model presumptions, various alternatives to the linear regression analysis model are employed. If the viral survival function is nonlinear, a logarithmic transformation

is performed before the linear regression analysis. The examples of the model application are provided in article (Lin & Li, 2002). It presents the survival functions for several viruses at different doses of UVGA.

There are many irradiation-based technologies that can be used to deal with the COVID-19 pandemic (Sabino et al., 2020). UV-C germicidal irradiation can be applied to disinfect surfaces, air and water. Its effects depend on the wavelength of the rays and the relative humidity of the air. The effectiveness of UV-C decreases with rising relative humidity. When removing viruses from surfaces, the effect of UV-C depends on the type of a surface. UV-C 254 nm rays are harmful to the human eyes and skin, so they should only be used when they do not cause any direct effects on humans. The major drawbacks of UV-C application are as follows: the potential of damaging open human tissues, a higher risk of cancer, and potential destructive effects on materials and surfaces in the long run. An overview of the advantages and disadvantages of different irradiation technologies is provided in article (Sabino et al., 2020).

In their previous works, the authors of article (Buonanno et al., 2020) proved that 222 nm UVC radiation effectively eradicates airborne viruses and extended their studies to investigate the effectiveness of ultraviolet light against the respiratory human coronaviruses alpha HCoV-229E and beta HCoV-OC43. Low doses of 1.7 and 1.2 mJ/cm^2 inactivated 99.9% of aerosolized coronavirus 229E and OC43. Because the genomes of all human coronaviruses are similar, it is likely that UVC radiation inactivation effectiveness will be similar when eradicating other human coronaviruses, including SARS-CoV-2. Based on the results of beta-HCoV-OC43 research, continuous exposure of infected areas to UV light at the current standard dose (~3 mJ/ m^2/8 h) would inactivate the virus by ~90% within ~8 minutes, by 95% – within ~11 minutes, by 99% – within ~16 minutes, and by 99.9% – within ~25 minutes. Thus, exposure to ultraviolet light at the current low normative doses can practically inactivate virus-infected areas. The following mathematical methods were used to research the dependence of coronavirus survival on the 222 nm UVC radiation dose: descriptive statistics, hypothesis testing, parameter confidence intervals, linear regression, robust regression, and bootstrap for small samples (Rotomskis & Streckytė, 2007).

The dependence of the virus survival on the dose of UVC 222-nm radiation was estimated by dividing the PFU/ml fraction in each UV dose (PFUUV) by zero-dose fraction (PFUcontrols): S = PFUUV/PFUcontrols (Buonanno et al., 2020). For each replicate experiment, the logarithm of the survival value was estimated to bring the distribution of residual errors closer to the normal distribution and thus satisfy the model presumptions. The robust linear regression analysis was employed to construct a linear regression equation in which ln[S] is a dependent variable, and a UV dose D (mJ/cm^2) is an independent variable, ln [S] = –k × D, where k represents a 222 nm UVCI radiation inactivation rate constant or a sensitivity coefficient (Panov & Borisova-Papancheva, 2015).

The SARS-CoV2-induced COVID-19 pandemic has led to widespread interest in effective and reliable disinfection methods to combat the virus, including ultraviolet germicidal inactivation (UVGI). Because the coronavirus is new, the recent literature is still lacking the research in its susceptibility to ultraviolet light. The paper presents the estimates of the effects of UVGI on SARS-CoV-2 that were obtained based on the studies focused on susceptibility of a close genomic relative, SARS-CoV-1, to UV light. Article (Arguelles, 2020) compares the genomic sequences of the two coronavirus species and reveals that the theoretical susceptibility of both species to UV is almost identical and differs by only 1.48% (Desboulets, 2018). The nonlinear regression analysis method was applied to SARS-CoV-1 survival data, obtained from the literature reviewed. The approximate UV-C dose required to inactivate the virus was found to be below the detection limit of the assay at 36,144 J / m^2 (≥5-log). By using this dose as a UVGI benchmark to affect SARS-CoV-2, a minimum exposure time $t \approx (1.5 \times 106)\pi \cdot (r^2/P)$ can be estimated; here r represents the distance from a UV-C source to a tested surface, P stands for the wattage of the germicidal bulb, and time t is expressed in seconds. For instance, irradiation with a 15 W UV-C bulb, 6 inches from the surface to be disinfected must last at least 2 hours. In this paper, a nonlinear regression formula for evaluating UVGI irradiation efficiency was obtained from the experimental results (Arguelles, 2020). Other researchers found that at the appropriate irradiation dose, a set of UV-C lamps was effective in reducing the spread of respiratory viruses in dental offices (Botta et al., 2020). In article (Khaiboullina et al., 2021), the method of dispersion analysis (ANOVA) was employed to research the effectiveness of the virus eradication with UV-C rays (Dutton, 2021).

To evaluate the effectiveness of UV-C radiation in inactivating surfaces (Dos Santos & de Castro, 2021), various hospital surfaces infected with 6 different types of viruses were researched. The surfaces were disinfected with a portable UV device, 254 nm UV-C light and 45.6 mW/cm^2 radiation intensity at a distance of 1 cm from the surfaces. The light dose was 0.912 J/cm^2. After the disinfection, virus concentrations on the surfaces were quantified and compared to the control (non-irradiated) group. The comparison was performed by applying a statistical t-criterion for paired samples. The differences were considered statistically significant when $p \leq 0.05$ (Dos Santos & de Castro, 2021).

A recent study (Gilbert et al., 2020) revealed that UV rays could decontaminate the protective N95 mask tissue contaminated with SARS-CoV-2. The effectiveness of UVGI radiation depends on a mask manufacturer, material and the medium containing the virus (liquid, air or surface) (Čekanavičius & Murauskas, 2014). Although there is a risk that UV rays can damage the protective mask materials and their effectiveness in filtering

particles, many different studies on the effects of UV rays did not find any significant deterioration in the quality of protective masks, even at UVGA doses several times higher than the dose required for virus disinfection (Yang et al., 2021).

1.1. Economic significance in developing a product usage algorithm

During the COVID-19 pandemic, the speed with which a product was placed on the market became particularly important. Therefore, the sooner a new product is placed on the market, the higher its demand may be. And the payback of such a product is faster. Another very important factor is the algorithm of using the developed product, in this case mathematical modeling becomes important, which should determine the optimal algorithm and its application sequence. In order to determine the optimal algorithm, it is necessary to take into account the most important conditions determining the time and conditions of application of the algorithm. In this case, the time period in which the COVID-19 virus is eliminated becomes paramount. Another issue is the ease of use of the equipment for the user. The less you have to pay attention to the equipment and how to turn it on or off, the less time and money is required. Therefore, it can be said that it is rational to turn on the equipment at the beginning of the work and to turn it off at the end, because the service life of HINS radiation sources is long enough and studies show that the virus becomes infectious after 9 seconds of use, but PCR viruses are not detected until 30 minutes after use. Based on these results, we have found that it is possible to automate the algorithm of turning the devices on and off, thereby reducing the user's time spent using the equipment. By estimating the staff time costs in this way, it is possible to determine the impact of the algorithm developed on the basis of mathematical modeling on the additional costs of using the equipment (Matys & Grzech-Leśniak, 2020).

1.2. Results

The literature analysis revealed that when evaluating the effectiveness of the measures undertaken to prevent the spread of respiratory viruses (in particular, SARS CoV-2 causing COVID-19) through air droplets, virus concentrations are measured at different points: inside and outside a patient's mouth; points at various distances to the virus source (a patient's mouth, a phantom mouth or an infected surface); input and output of the virus transmission prevention device; on the surfaces of the dental office; a patient's skin, a doctor's protective mask or shield, clothing, dental tools and devices.

The two main dependent traits of these studies are as follows:

- The respiratory virus (SARS CoV-2 causing COVID-19) concentration (in the air or on surfaces).
- The respiratory virus (SARS CoV-2 causing COVID-19) survival time (in the air or on surfaces).

Most of the experiments were aimed at investigating the dependencies of the above-mentioned traits on various quantitative and/or qualitative characteristics:

Type of aerosols. Aerosols with SARS-CoV-2 virus and its strains (Great Britain, South Africa, Brazil, etc.), artificially generated aerosols with similar diffusion properties as real ones (used in simulation experiments, e.g. when researching the effectiveness of aerosol extraction from a certain environment or the effects of suction devices on the spread of the virus.

The UVGA source and irradiation characteristics. A source type and wavelength (e.g., $\lambda = \in(200; 280)$nm); UV radiation intensity I (W/m^2) (when getting further from the source, UV radiation intensity is decreasing exponentially); irradiation duration (s or min.); irradiation dose D (J/m^2) which is equal to the product of UV radiation intensity and irradiation duration $D = I\ t$.

Type of an irradiated object. Air (aerosols in a patient's mouth, aerosols at a device's input, aerosols in the air of the dental office), surfaces (skin, dental instruments, a dentist's and a patient's personal protective equipment (protective clothing, protective masks and shields)).

Characteristics of an irradiated object. Temperature, length, width, height, volume, area, age, condition and other qualitative and quantitative characteristics.

Type of a dental service. Provided by a dentist, endodontist, oral hygienist, periodontist, orthopaedist.

Risk rate of a dental procedure. Low, medium, high.

Characteristics of a dental office. Area, volume, air ventilation system, air ventilation intensity, air temperature, air humidity.

In virtually all studies, the respiratory virus (SARS-CoV-2 causing COVID-19) concentration or survival dependence was examined for only 1–3 characteristics. Other traits were considered fixed or not mentioned at all.

Literature addresses the issues of the transfer of SARS-CoV-2 aerosol particles that still need to be answered in the near future: How does the ability of an airborne virus to infect diminish? How do speaking, coughing and breathing affect viral emissions? What is the mechanism of aerosol generation and how does the amount of contaminated aerosols change when providing different medical, including dental, services? How does a worker's risk of getting infected with the virus depend on the concentration of aerosols, the type of personal protective equipment, and the length of time in the infected environment? What engineering measures are most effective in preventing the infection caused by breathing and touching contaminated surfaces? For instance, ultraviolet germicidal irradiation (UVGA) is known to be effective in eradicating viruses in the air and on surfaces, but little is known about the effectiveness of UVGA in eradicating SARS-CoV-2, including the dose of UVGA required to eradicate the virus, depending on temperature, humidity, and other traits.

Literature analysis also revealed that when evaluating the respiratory virus (SARS-CoV-2 causing COVID-19) concentration and survival rate dependence on a number of traits, the methods of descriptive statistics, confidence intervals, hypothesis testing, dispersion analysis, trait dependence analysis, and regression analysis are employed. All the above-listed methods have been tested under laboratory conditions and thus can be applied to evaluate the effectiveness of the project product – a device designed to prevent transmission of respiratory viruses through air droplets. Selection of a particular method depends on a set of traits to be analyzed, a trait type (quantitative, qualitative), a trait distribution type, and data applicability for one or another method. The following sections provide a brief description of the methods selected as well as recommendations and presumptions for their relevant application. All of the methods selected have been implemented programmatically by employing data analytics software SAS, R, SPSS, etc. (Elliott & Woodward, 2015; Lalanne & Mesbah, 2016, 2017).

The main purpose of descriptive statistics is to provide a concise description of the data collected. Qualitative and quantitative data are systematised by employing numerical characteristics: frequencies, relative frequencies, mean, median, standard deviation, quantiles, and so forth. Data can be graphically represented by bar charts, pie charts, fishbone diagrams, histograms, etc. There are many methods of graphical data representation. Which method is best for a particular situation, depends on the method applied and the traits being analysed.

Numerical data characteristics are used to briefly describe and compare the sets of the data collected. The characteristics are divided into two main groups. Position characteristics describe the magnitude of data values. The major position characteristics are mean, median, mode, and quantiles. Dispersion characteristics describe the variability of data values. The major dispersion characteristics are general dispersion, standard deviation, quartile width, and coefficient of variation.

Position characteristics. The major position characteristics are mean, median, mode, and quartiles.

The sample mean is estimated as follows:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i ,$$

where n marks the sample size. The sample mean describes the mean value of a sample.

A sample's q-quantile divides the variation line into $q\times100$ and $(1-q)\times100$ percentages ($0<q<1$), i.e. $q\times100$ percent sample values are not higher than quantile, and $(1-q)\times100$ percent – not lower than quantile x_q. The first quartile (lower) is the median of the lower side; the third quartile (upper) is the median of the upper side, i.e. when $p = 0.25$ and $p = 0.75$, we get the lower $x_{0.25}$ and the upper $x_{0.75}$. quartiles. The median divides the data into two halves, while the purpose of the quartiles is to divide the data into quarters. The median itself is the second quartile, so about 50% of the numbers are no larger than the median. The quantiles $x_{0.01}, x_{0.02}, \ldots, x_{0.99}$ are referred to as empirical percentiles.

Dispersion characteristics. Quartile width (QW) is commonly used to eliminate the distortion caused by variances. The quartile width is equal to the difference between the upper (third) and lower (first) quartiles $QW = x_{0.75} - x_{0.25}$; it describes the dispersion of 50% middle data values. Quartile width is considered a reliable characteristic of dispersion.

Sample dispersion: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$, describes the degree of trait X value dispersion around the sample mean. The definition of a sample dispersion proposes that it is measured in square units. Thus, to describe the degree of dispersion, another characteristic $s = \sqrt{s^2}$ if often used; this characteristic is referred to as a sample's standard deviation.

The degree of random value dispersion is also represented by the coefficient of variation:

$$V_K = \frac{s}{\overline{x}} .$$

Quantity $s_{\overline{x}} = \frac{s}{\sqrt{n}}$ is referred to as a standart error of mean. The larger is sample size n, the smaller is estimation error $\overline{x}$.

Asymmetry and excess coefficients. A sample's (empirical) asymmetry coefficient (skewness) $\overline{A}_s$ represents asymmetry of the empirical distribution. If the distribution is symmetric with respect to the mean, then $\overline{A}_s = 0$. In the case of right asymmetry, $\overline{A}_s > 0$, while in the case of left asymetry, $\overline{A}_s < 0$. In the case of a normal distribution, $\overline{A}_s = 0$.

A sample's (empirical) excess coefficient (kurtosis) represents the peak ($\overline{E}_k > 0$) or flatness ($\overline{E}_k < 0$) of the empirical distribution density (histogram), compared to the normal distribution. In the case of a normal distribution, $\overline{E}_k = 0$. If the maximum of the empirical distribution density function is higher (lower) than the normal law, then the empirical distribution is said to have a positive (negative) excess.

Estimates of the population parameters (mean, standard deviation, median, quantile, correlation coefficient, regression coefficients, etc.) are random variables. Their realizations obtained from random samples are scattered around the true value of the population parameter. In practice, it is important to know the range to which an unknown population parameter may belong.

Taking into account parameter $\hat{\Theta}$ (sample mean, median, quantile, standard deviation, a regression equation, etc.), we want to know the range in which the values of the population parameter Θ are going to vary under a given probability.

Therefore, once a point estimate $\hat{\Theta}$ has been found, it is necessary to answer the question about the accuracy and reliability of this estimate. When solving practical problems, it is important to know what possible errors can be expected, if we replace an unknown parameter

Θ with its point estimate $\hat{\Theta}$. What is the probability that the errors will not exceed particular ranges? These questions are especially relevant when the sample size is small because then the point estimate is highly random.

Because $\hat{\Theta}$ is a random variable, $P(\hat{\Theta}-\varepsilon<\Theta<\hat{\Theta}+\varepsilon)=1-\alpha$ is required to be valid with a high probability $(1-\alpha)\in\{0.9;0.95;0.95\}$. The interval $(\hat{\Theta}-\varepsilon;\hat{\Theta}+\varepsilon)$ is referred to as a confidence interval, probability $1-\alpha$ – as a confidence level, and quantity ε – as an error. We will present an algorithm for estimating the mean confidence interval when distribution of an observed quantity is normal.

Finding the confidence interval of the normal distribution mean. A random sample $x=(x_1;x_2;\ldots x_n)$ is given. The distribution of observed trait $\mathbf{X}$ is known to be normal $X\sim N(\mu,\sigma)$. Both parameters, mean μ and standard deviation σ are unknown.

The problem is to find the confidence interval of the unknown population mean μ.

Solution algorithm:

The best point estimate of μ is sample mean $\bar{x}$

$$\hat{\mu}=\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i\,.$$

For constructing the confidence interval, the statistics are used:

$$t=\frac{\bar{x}-\mu}{\bar{s}}\sqrt{n},\quad t\sim St(n-1).$$

The confidence level $1-\alpha$ is selected.

Quantiles of the Student's distribution are estimated: $t_{\alpha/2;n-1}$ ir $t_{1-\alpha/2;n-1}$.

The mean confidence interval is constructed:

$$PI_{1-\alpha}(\mu)=(\bar{X}-\frac{\bar{S}}{\sqrt{n}}\cdot t_{1-\frac{\alpha}{2};n-1}\;;\;\bar{X}-\frac{\bar{S}}{\sqrt{n}}\cdot t_{\frac{\alpha}{2};n-1}).$$

The formulas for finding the confidence intervals for various distribution parameters are provided in sources (Čekanavičius & Murauskas, 2002, 2014; McHugh, 2011; Judd et al., 2017). Three interrelated variables are used to construct the confidence interval: the confidence level, sample size n, and accuracy. When seeking the highest possible level of confidence, the width of the confidence interval increases greatly, which diminishes the accuracy of the information.

The confidence interval of the normal distribution $X\sim N(\mu,\sigma^2)$ dispersion is as follows:

$$PI_{1-\alpha}(\sigma^2)=(\frac{\bar{S}^2\cdot(n-1)}{\chi^2_{1-\frac{\alpha}{2};n-1}}\;;\;\frac{\bar{S}^2\cdot(n-1)}{\chi^2_{\frac{\alpha}{2};n-1}}).$$

$\bar{S}^2=\frac{1}{n-1}\cdot\sum_{i=1}^{n}(X_i-\bar{X})^2$ represents an unshifted point estimate of the population dispersion σ^2.

Event probability confidence interval:

$$PI_{1-\alpha}(p)=(\hat{p}-z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}}\;;$$
$$\hat{p}-z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}}),$$

here $\hat{p}=\frac{m}{n}$.

When comparing virus concentrations on two different surfaces, the hypotheses concerning population mean differences for two independent or two paired samples are tested by employing Student's statistics. In this section, the hypotheses concerning two mean differences are tested for only two independent samples. When it is necessary to test the hypothesis of the equality of three and more means, the method of dispersion analysis is applied.

For comparing unknown means of two traits, when samples $(X_1,X_2,\ldots,X_n)$ ir $(Y_1,Y_2,\ldots,Y_m)$ are obtained by observing two independent normal random quantities $X\sim N(\mu_X,\sigma_X^2)$ and $Y\sim N(\mu_Y,\sigma_Y^2)$ with unknown means μ_X ir μ_Y and dispersions, Student's criteria are applied. Suppose we need to test the following hypotheses:

$$H_0:\mu_X=\mu_Y,$$
$$H_a:\mu_X\neq\mu_Y,\ (\text{or } H_a:\mu_X>\mu_Y \text{ or } H_a:\mu_X<\mu_Y).$$

For testing hypothesis H_0, when dispersions are equal $\sigma_X^2=\sigma_Y^2=\sigma^2$, Student's statistics is applied:

$$T=\frac{\bar{X}-\bar{Y}}{\sqrt{S_p^2(1/n+1/m)}},$$

it possesses Student's distribution with (n+m–2) degrees of freedom; here $S_p^2=\frac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}$ is a joint population dispersion σ^2 estimate. By inserting expression S_p^2 into the statistics T formula, we obtain the following equation:

$$T=\frac{\bar{X}-\bar{Y}}{\sqrt{(n-1)S_Y^2+(m-1)S_Y^2}}\sqrt{\frac{nm(n+m-2)}{n+m}}\,.$$

Statistics T is also used for constructing the confidence interval of mean difference 1-α:

$$(\bar{X}-\bar{Y})\pm t_{1-\alpha/2;n+m-2}\sqrt{S_p^2(1/n+1/m)}.$$

When dispersions of the variables observed are unequal $\sigma_X^2\neq\sigma_Y^2$, the so-called Behrens-Fisher problem is faced. Several approximate solutions to this problem are known. Student's statistics is used to test hypothesis H_0:

$$T=\frac{\bar{X}-\bar{Y}}{\sqrt{S_x^2/n+S_y^2/m}};$$

it possesses Student's distribution with $\nu = \frac{1}{\nu_x + \nu_y}$ degrees of freedom; here

$$\nu_x = \left(\frac{\frac{S_X^2}{n}}{\frac{S_X^2}{n} + \frac{S_Y^2}{m}} \right)^2 \cdot \frac{1}{n-1}, \quad \nu_y = \left(\frac{\frac{S_Y^2}{m}}{\frac{S_X^2}{n} + \frac{S_Y^2}{m}} \right)^2 \cdot \frac{1}{m-1}.$$

When trait distributions are unknown, the Mann-Whitney-Wilcoxon rank sum criterion is applied to two independent samples. Suppose we have two samples. Let's denote:

n1 – the number of members in the first sample,
n2 – the number of members in the second sample,
R1 – sum of the ranks of the first sample members,
R2 – sum of the ranks of the second sample members.
Here n1 and n2 are not necessarily equal.
We test the hypothesis:
H_0 : Distributions of traits X and Y are equal,
H_a : Distributions of traits X and Y are not equal.

When samples are large (n1>20, n2>20), the hypothesis concerning the overlap of the distributions in two populations is tested by applying the following statistics:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1.$$

If the null hypothesis that all n1+n2 observations have the same distribution is valid, then statistics U has a normal distribution with mean $\mu = \frac{n_1 n_2}{2}$ and standard deviation $\sigma = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$. Then statistics $Z = \frac{U_1 - \mu}{\sigma}$ has the standard normal distribution with mean equal to 0 and standard deviation equal to 1.

Compatibility hypotheses are recommended to be tested in a variety of ways. Based on different criteria, the gap between the theoretical and empirical distribution is evaluated by employing various measures. Thus, by applying several criteria, more information can be obtained.

The Kolmogorov-Smirnov criterion is used to test the compatibility hypothesis:

H_0 : The function of a random quantity X distribution is $F(x)$.

Let us define the Kolmogorov-Smirnov criterion. Suppose that the function of a random quantity X distribution is $F(x)$. Based on the sample data, the empirical distribution function $F_n(x)$ is estimated. To test the compatibility hypothesis, Kolmogorov and Smirnov proposed the statistics that measures the difference between the empirical and theoretical distributions and evaluates the difference between the empirical distribution function $F_n(x)$ and the theoretical distribution function $F(x)$:

$$Z = \sqrt{n} \max_x |F_n(x) - F(x)|,$$

here n – the sample size.

One of the most popular compatibility criteria is chi-square criterion χ^2. It is used to test hypotheses about the distribution of a trait (normal, binomial, etc.) in a population. Criterion χ^2 indicates whether the difference between the empirical and theoretical distributions is significant, i.e. it is verified whether the empirical distribution is compatible with the theoretical model. Criterion χ^2 applies to the aggregate data, so the sample has to be quite large.

The general scheme of estimating the criterion is as follows:

– In case a discrete variable is observed, the sample values (category frequencies) are calculated initially.
– In case a continuous variable is observed, the range of values is divided into non-intersecting intervals, and then interval frequencies are calculated.

Suppose category (interval) frequencies are represented by $O_1, O_2, \ldots, O_k$, here k is a number of categories (intervals). By using the proporties of the teoretical distribution (indicated in the formulation of hypothesis H_0), we estimate how many values of the variable should be assigned to each category (would fall into each interval), if the hypothesis about distribution of the variable was valid, i.e. the probable frequencies $E_1, E_2, \ldots, E_k$ are found.

The differences in expected and observed frequencies are estimated. The larger are these differences, the less likely is the hypothesis about the distribution to be valid. The decision-making rules are based on the magnitude of the differences between the expected and observed frequencies.

Further we will describe how compatibility criterion χ^2 applies, when a continuous variable is observed. Suppose that based on the variable observed, the population can be divided into k categories. The share of the population assigned to the jth category is marked p_j, $j = 1, \ldots, k$ (the equivalent formulation: a random variable that acquires the jth value with probability p_j is observed). In case the hypothesis about distribution H_0 is valid, the distribution of the variable observed is known, and the probability of the variable being assigned to the jth category is p_j^0. The hypothesis concerning compatibility of the empirical and theoretical distributions is formulated:

$$\begin{cases} H_0 : p_1 = p_1^0, \quad p_2 = p_2^0, \ldots, p_k = p_k^0, \\ H_a : p_j \neq p_j^0, \quad \text{at least one of } j = 1, \ldots, k. \end{cases}$$

When H_0 is valid, $E_j = np_j^0$ observations out of the sample with n observations should be assigned to the jth category. If they actually are assigned to O_j, then the difference $O_j - E_j$ indicates whether hypothesis H_0 is probable. In the case of a discrete distribution, compatibility of the empirical and theoretical distributions is verified by employing the following statistics:

$$\chi^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j}.$$

If hypothesis H_0 is valid, then the criterion statistics possesses distribution χ^2 with k–1 degrees of freedom. When verifying hypothesis H_0, the criterion with the right critical area is used.

$$P(\chi^2 \geq \chi^2_{1-\alpha,k-1}) = \alpha .$$

H_0 non-rejection area is $\chi^2_{H_0} = [0; \chi^2_{1-\alpha;\, k-1})$, and the critical area is $\chi^2_K = [\chi^2_{1-\alpha;\, k-1}, \infty)$. If the statistical value under observation is assigned to H_0 non-rejection area $\chi^2_{imt} \in \chi^2_{H_0}$, then H_0 is not rejected, otherwise, it is rejected.

When comparing virus concentrations at different levels of qualitative factors, the dispersion analysis is employed. This section examines the differences between three or more population mean differences. The actual population means are often unknown, so their differences are measured based on sample mean differences.

When sample mean differences are significant, it is very unlikely to be a coincidence. We then say that sample means differ statistically significantly and there is a high probability that the means of the populations themselves also differ in this sense. Suppose we are interested in:

The first studies based on dispersion analysis were published by Fisher. Dispersion analysis is the research of the dependence of random distributions on particular factors that are qualitative. One of the major purposes of dispersion analysis is to research whether the means of dependent variable Y, measured in different populations, substantially vary.

A completely randomised one-factor dispersion analysis model. Suppose that distribution of a random variable Y can depend on a factor A which acquires I different values a_1, a_2, ..., a_I. Based on factor A, I independent populations are distinguished. The same dependent variable Y (measured on an interval or a ratio scale) is estimated in each population. In the population where factor A acquires value $A = a_i$, the variable is denoted by Y_i, and its sample mean – by $\overline{Y_i}$. Dispersion analysis can be based on a number of models (Judd et al., 2017). Further we will discuss a completely randomised dispersion analysis model with constant factors that is most commonly applied.

Suppose that distribution of observed random variable Y depends on factor A which is at I different levels. Thus, we have I samples, each sized n_i, $i = \overline{1, I}$, $n = \sum_{j=1}^{I} n_i$.

Each observation y_{ij} is divided into two components:

$$y_{ij} = \beta_i + e_{ij}, \quad j = 1,...,n_i, \; i = 1,...,I ,$$

here β_i denotes unknown population means $M(Y_i) = \beta_i$, and e_{ij} represents independent random variables with a standard normal distribution $N\left(0, \sigma^2\right)$.

The null hypothesis of one-factor dispersion analysis proposes that the means of all population variables are equal.

$$H_0 : \beta_1 = ... = \beta_I .$$

To verify the null hypothesis, Fisher statistics is used:

$$F = \frac{\overline{SS_A}}{\overline{SS_e}},$$

here

$SS_A = \sum_{i=1}^{I} \left(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot}\right)^2 n_i$ – sum of the deviation squares that describes the effect of factor A on the mean of observed random variable Y;

$SS_e = \sum_{i=1}^{I} \sum_{j=1}^{n_i} \left(y_{ij} - \overline{y}_{i\cdot}\right)^2$ – sum of the deviation squares that describes the effect of random error factor E on the mean of observed random variable **Y**, in the model defined by random e_{ij};

$\overline{y}_{\cdot\cdot}$ – empirical mean of sample Y; $\overline{y}_{i\cdot}$ – empirical mean of sample Y_i;

$SS_p = \sum_{i=1}^{I} \sum_{j=1}^{n_i} \left(y_{ij} - \overline{y}_{\cdot\cdot}\right)^2$ – total sum of deviation squares;

$SS_p = SS_A + SS_e$;

$\overline{SS_A} = \frac{1}{I-1} SS_A$, $\overline{SS_e} = \frac{1}{n-I} SS_e$, – means of deviation squares (a sample's factor and residual dispersions).

If H_0 is not rejected, then the observation data do not contradict the null hypothesis, i.e. it can be considered that factor A does not affect the mean of observed (researched) random variable Y. If the observations do not contradict the null hypothesis H_0 concerning the equality of the means, then the analysis can be concluded. In this case, all observants can be aggregated into a single sample of size n, obtained by observing the normal random variable with mean β_0 and dispersion σ_e^2. The opposite conclusion hardly satisfies a researcher because the question naturally arises as to how Y depends on the levels of factor A. What factor levels are the cause of non-homogeneity? Can the levels of factor A be grouped so that the difference in means within the groups is insignificant? Multiple comparisons can be used to solve this problem (Ehtezazi et al., 2021).

When applying the Fisher criterion, dispersions of the populations must be equal. If the hypothesis concerning equality of the dispersions is rejected, then, to test the hypothesis of the means equality, the Welch or Brown-Forsythe criteria should be applied instead of the Fisher criterion (Tseng & Li, 2005).

The criteria for multiple comparisons are divided into a priori and a posteriori (post hoc). A priori comparisons are planned before the dispersion analysis or instead of it. A posteriori (post hoc) comparisons are made after the results of the analysis (the results of testing the hypothesis about the equality of several means) are known. There are many different criteria for multiple comparisons. Most criteria (only with different levels of significance) can be used as both a priori and a posteriori multiple comparisons.

Post Hoc criteria. Commonly used a posteriori (post hoc) multiple comparison criteria, such as LSD (Least Significant Difference), Bonferroni, Sidak, Scheffe, R-E-G-W F (Ryan-Einot-Gabriel-Welsch F criterion), R-E-G-W Q (Ryan-Einot-Gabriel-Welsch Q criterion), S-N-K (Student-Newman-Keuls criterion), Tukey, Tukey b, Duncan, Hochberg GT2, Gabriel, Waller-Duncan, and Dunnett criteria are applied when dispersions of the populations under consideration are equal. Tamhane T2, Dunnett T3, Games-Howell, and Dunnett C criteria are applied when dispersions of the populations under consideration are not equal (McHugh, 2011).

Each of the criteria has both advantages and disadvantages. Some criteria apply when all sample sizes are the same, while others can be used when the sample sizes are not the same. Some criteria apply when population dispersions are equal, while others can be used when the dispersions are unequal. Some criteria are more likely to reveal statistically significant differences, while others (more conservative criteria) are not likely to do that, and so forth. Selection of an appropriate post-hoc multiple comparison criterion is not an easy task; it depends on the data under consideration and requires deeper research (Čekanavičius & Murauskas, 2002; Aggarwal & Ranganathan, 2017). Further we will provide two criteria most relevant for implementation of this project.

The Bonferroni criterion. Based on the Bonferroni criterion, the level of the experiment significance α_E (i.e. the probability of at least once incorrectly derermining a statistically significant difference between two means when comparing all potential pairs) is selected, and all mean pairs are compared by applying Student's t-criterion at the significance level $\alpha = \alpha_E / C$, when $C = I(I-1)/2$. The Bonferroni criterion is not applied when there are many population means because α decreases dramatically, and a statistically significant mean difference is rarely obtained, although the actual population means differ (i.e. the probability of the second type error increases significantly) (McHugh, 2011).

The Tukey criterion. The Tukey's HSD (*Honestly Significant Difference)* criterion is one of the most commonly used criteria. It is a good alternative to the Bonferroni criterion with many samples. The Tukey's criterion is very conservative, i.e. it is less likely to reject the null hypothesis, i.e. to recognize mean differences as statistically significant. By employing the Tukey criterion, homogeneity groups are formed. If the samples, selected from the populations, vary in size, then the results of the homogeneity groups and the HSD criterion may also differ because formation of the homogeneity groups is based on harmonic means. The power of the Tukey's HSD criterion when researching a larger number of samples is higher compared to the power of the Bonferroni criterion, and vice versa – the power of the Bonferroni criterion is higher when researching a smaller number of samples.

When the data do not meet the presumptions of the dispersion analysis, the nonparametric Kruskal-Wallis criterion is applied to k-independent samples. Then the following hypotheses are tested:

H_0 : Variable distributions are equal, when alternative

H_a : Variable distributions are not equal.

The hypothesis proposes that all samples are extracted from the populations in which the variable under consideration has the same distribution and the same mean. The hypothesis is tested by employing the statistics $H = \frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{R^2_j}{n_j} - 3 \cdot (n+1)$, where n_j represents the number of the jth's sample members, R_j – sum of the jth's sample member ranks, k – the number of samples, and $n = n_1 + n_2 + \ldots + n_k$ – total number of all sample observations. When the hypothesis is valid, statistics H has approximately χ^2 distributions with k–1 degrees of freedom. The Kruskal-Wallis criterion requires that the size of all samples is at least 5.

It is often necessary to answer the question of whether the traits observed are dependent or independent. If they are found to be dependent, then strength of the correlation is evaluated. Correlation coefficient is a measure of trait interdependence. By testing the hypothesis of a population correlation coefficient or calculating its confidence interval, we answer the question about the interdependence of traits in the population in terms of linearity, monotony, compatibility, etc.

Nevertheless, the correlation coefficient does not reveal the cause of the correlation. Two variables X and Y can be strongly correlated for three reasons: variable X affects variable Y; variable Y affects variable X; both variables X and Y are affected by the third variable; thus, the relationship revealed by the correlation analysis cannot be interpreted as causality, but only as a measure of association or relationship.

To evaluate the strength of the linear relationship between two quantitative traits, Pearson's correlation coefficient is employed. It applies when distributions of traits X and Y under consideration are normal. Pearson's correlation coefficient for a sample is estimated by the formula:

$$\widehat{\rho} = r = \frac{k_{xy}}{s_x s_y} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sqrt{\overline{x^2} - (\overline{x})^2}\sqrt{\overline{y^2} - (\overline{y})^2}}.$$

The linear relationship is stronger when value $|r|$ is closer to 1. If r>0, then, as the values of one trait are increasing, the values of the other are increasing linearly. If r<0, then as the values of one trait are increasing, the values of the other are decreasing linearly; r does not reveal any non-linear dependency. The larger is the sample, the closer is r to the unknown population correlation coefficient ρ.

Testing the Pearson correlation coefficient hypothesis. Which value of the population correlation coefficient's ρ point estimate r can be considered statistically significant? At which value of r it can be stated that there exists a statistically significant linear relationship between observed traits X and Y?

Suppose two traits X and Y with unknown correlation coefficients in population ρ are observed. To answer the question of whether these values are linearly correlated, the following hypotheses are verified:

$H_0 : \rho = 0$;

$H_a : \rho \neq 0$.

If H_0 is rejected, then X and Y are statistically significantly correlated in the population under consideration; the strength of the relationship may vary from very weak (ρ close to zero) to a functional relationship ($|\rho|=1$).

When solving practical problems, we rarely know a type of the trait distribution observed or know that the distributions are not normal. In this case, the coefficients not related to distributions should be employed. Spearman's rank correlation coefficient ρ_S describes strength of the relationship between X and Y in terms of monotony, i.e. as X is increasing, Y is increasing (not necessarily linearly) or decreasing monotonically. This coefficient is used to evaluate the strength of the relationship between trait ratios, intervals and order scales. To test the hypothesis concerning significance of Spearman's rank correlation coefficient, Student's statistics is applied.

1.3. Evaluation of the qualitative trait correlation

Suppose we observe a pair of qualitative traits (*X, Y)*. Trait *X* acquires *I* different values, while trait *Y* acquires *J* different values. Let us denote the observed frequency of pairs (x_i; y_j) by o_{ij}. The results of the observations are provided in the trait dependence Table 1.

Table 1. Trait dependence table

X\Y	y_1	y_2	…	y_J	Σ
x_1	o_{11}	o_{12}	…	o_{1J}	$o_{1\bullet}$
x_2	o_{21}	o_{22}	…	o_{2J}	$o_{2\bullet}$
⋮	⋮	⋮		⋮	⋮
x_I	o_{I1}	o_{I2}	…	o_{IJ}	$o_{I\bullet}$
Σ	$o_{\bullet 1}$	$o_{\bullet 2}$	…	$o_{\bullet J}$	n

It is often necessary to answer the question of whether the observed traits are dependent or independent. Then the following hypotheses are tested:

H_0 : "Traits are independent".

H_a : "Traits X and Y are dependent".

To verify the hypotheses, criterion χ^2 with the right critical area is applied:

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}} \sim \chi^2\left(\left(I-1\right)\left(J-1\right)\right),$$

here O_{ij} – observed frequency; E_{ij} – expected frequency.

When H_0 is rejected, strength of the relationship is evaluated. There are over 100 correlation coefficients that describe strength of the relationship between particular qualitative traits. We will provide several relationship measures when the observed variables are measured based on the name scale. The same rule applies to all the above-mentioned relationship measures: the higher is their absolute value, the greater is trait dependence; the closer to zero is their value, the weaker is trait dependence.

Coefficient Phi is a relationship measure for tables 2×2; it is also referred to as the overlap coefficient and expressed as follows:

$$\varphi = \sqrt{\chi^2 / n}.$$

Coefficient's φ variation interval for table 2×2 is [0; 1].

Cramer's V coefficient:

$$V = \frac{\varphi}{\sqrt{\min\left(I-1, J-1\right)}}, \quad 0 \le V \le 1.$$

For table 2×2, Cramer's V coefficient overlaps with coefficient φ.

We present several relationship measures when the observed variables are measured based on the order scale. The same rule applies to all the above-mentioned relationship measures: the higher is their absolute value, the greater is trait dependence; the closer to zero is their value, the weaker is trait dependence. The values of the measures may vary between –1 and 1.

The Kendall rank correlation coefficients measure the strength of the relationship in terms of compatibility between the variables measured on the ratio, interval, and order scales. All possible values of the rank pairs representing the observed variables (X,Y) are compared with each other and it is verified whether they are compatible, incompatible or bounded. If comparing the respective values of the rank pairs (rx_i, ry_i) and (rx_j, ry_j), representing variables X and Y, we estimate that:

($rx_i > rx_j$) and ($ry_i > ry_j$) or ($rx_i < rx_j$) and ($ry_i < ry_j$), then the pair is considered compatible. A number of the compatible pairs in a sampe is denoted P;

($rx_i > rx_j$) and ($ry_i < ry_j$) or ($rx_i < rx_j$) and ($ry_i > ry_j$), then the pair is considered incompatible. A number of the incompatible pairs in a sampe is denoted by Q;

($rx_i = rx_j$) and ($ry_i \neq ry_j$), then the pair is considered bounded by x; a number of such pairs in a sample is denoted by T_x;

($rx_i \neq rx_j$) and ($ry_i = ry_j$), then the pair is considered bounded by y; a number of such pairs in a sample is denoted by T_y;

($rx_i = rx_j$) and ($ry_i = ry_j$), then the pair is considered bounded by x and y; a number of such pairs in a sample is denoted by T_{xy}.

The total number of bounded pairs in a sample is denoted by $T = T_x + T_y + T_{xy}$.

Kendall rank correlation coefficient τ_a evaluates strength of the relationship in terms of compatibility and applies when a sample does not have any bounded pairs.

$$\tau_a = \frac{P-Q}{\frac{1}{2}n\cdot(n-1)}, \quad -1 \le \tau_a \le 1.$$

When a sample has bounded pairs, Kendall rank correlation coefficient τ_b is employed:

$$\tau_b = \frac{P-Q}{\sqrt{P+Q+T_x}\sqrt{P+Q+T_y}}, \quad -1 \le \tau_b \le 1.$$

If in the ***I*x*J*** trait dependence table I=J, strength of the relationship is evaluated by employing Kendall coefficient τ_b; if I≠J, Stewart's τ_c is recommended:

$$\tau_c = \frac{m \cdot (P-Q)}{n^2 \cdot (m-1)},$$

here $m = \min(I, J)$, $-1 \le \tau_c \le 1$.

The Goodman-Kruskal coefficient γ indicates if a sample has more compatible or incompatible pairs ($|\gamma|$ is equal to the value of the probability difference that shows to which extent the probability that a random pair is compatible is higher than the probability that it is incompatible when γ>0, and vice versa when γ<0):

$$\gamma = \frac{P-Q}{P+Q}.$$

When γ = 0, a sample has the equal number of compatible and incompatible pairs; when γ = 1, all pairs are compatible; when γ = –1, all pairs are incompatible. Note: coefficient γ does not consider bounded pairs. Somers proposed a coefficient's γ modification – coefficient d that considers bounded pairs:

$$d = \frac{P-Q}{P+Q+(T_x+T_y)/2}.$$

He also proposed two more coefficients for the case if there is no symmetry between variables X and Y (asymmetric Somers' coefficients):

$$d_{XY} = \frac{P-Q}{P+Q+T_x}, \quad d_{YX} = \frac{P-Q}{P+Q+T_y}.$$

Variables X and Y are considered symmetrical if the success of the variable value forecast does not depend on whether X values are forecasted when knowing y values, or vice versa. Otherwise, the variables are considered asymmetric.

Regression analysis is aimed at investigating causal relationships between dependent and independent variables. Within the framework of the target research area, the dependent variables are as follows: virus concentration in the air or on surfaces, and virus survival in the air or on surfaces. Regression analysis is a powerful mathematical model, but, unfortunately, this analysis is often misapplied or misinterpreted, which leads to inaccurate forecasts or unreasonable decisions. For instance, linear regression analysis is still applied in the cases when its presumptions are not met (researchers often fail to verify the normality of the regression residues, homoscedasticity, the effect of exclusions on the regression coefficients, multicollinearity, and other presumptions). Thus, reliable conclusions in regression analysis are only drawn if the presumptions of a model's relevance are met. This issue is discussed further in the section.

The regression model relates a dependent variable Y with other – independent – variables X_1, X_2 ,..., X_K. By employing the regression equation, values of the dependent variable Y can be forecasted based on the values of independent variables with a certain degree of reliability. The most common model of linear regression is as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_K X_K + \varepsilon.$$

The model of multiple linear regression analysis is most conveniently written by employing matrices (Darlington & Hayes, 2017). Suppose we have n observations – Y_1 , Y_2 ,..., Y_n – of a dependent variable, and n observations – X_{1j}, X_{2j},..., X_{nj} – of each independent variable X_j , $j = 1, K$. Then the model is written as follows:

$Y = X\beta + \varepsilon$, here

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{bmatrix},$$

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Y – the vector of (n×1) dimension dependent variable values; X – the matrix of n×(K×1) dimension independent variable values; β – the vector of (n×1) dimension regression equation coefficients; ε – the vector of (n×1) dimension random errors (Darlington & Hayes, 2017).

In the model of multiple linear regression analysis, variables Y, X_1, X_2 ,..., X_K are quantitative, measured on interval and ratio scales, or binary (dichotomous) variables. The model of linear regression analysis can be applied if the data meet particular conditions. Most of the presumptions of the regression analysis are the requirements to be met by random errors ε_i , which indicate to which extent the observed value Y varies from the value that would be obtained when making forecasts based on the regression equation. The basic presumptions of regression analysis are as follows:

random errors ε_i are normally distributed random quantities; means of all ε_i are equal to zero; $E\varepsilon_i = 0$; dispersions of all ε_i are equal (presumption of homoskedasticity); $D\varepsilon_i = \sigma^2$; the data do not possess any variances (Astivia & Zumbo, 2019).

In the classical model of multiple linear regression analysis, variables Y, X_1, X_2 ,..., X_K are quantitative, measured on interval or ratio scales.

When constructing a regression model for systems that depend on qualitative parameters (e.g., there are

two different surfaces on which the virus concentration is measured, or two different experimental conditions, etc.), the regression model is supplemented by binary qualitative variables, also called pseudovariables. All binary variables can acquire only two values. For instance:

$$Y = \beta_0 + \beta_1 D + \beta_2 X + \varepsilon, \text{ here } D \in \{0,1\}.$$

When a categorical variable has $m>2$ categories, it is replaced by a $(m-1)$ binary variable. Regression equations can be constructed with several binary variables, interactions of binary variables or interactions of quantitative and pseudovariables, for instance, $Y = \beta_0 + \beta_1 D + \beta_2 X + \beta_3 (D'X) + \varepsilon$. A detailed description of regression models with pseudovariables and interactions is presented in literature (Darlington & Hayes, 2017).

If the variables are not suitable for linear regression, they are undergoing a transformation. This problem is often solved by logarithmizing both the dependent variable and some of the independent variables. Sometimes the variables to be included in the regression model are raised to an appropriate degree, or the interactions (multiplications) of two or more variables are included. If the variables cannot be transformed to suit linear regression analysis, then nonlinear regression analysis is employed. Although it is implemented in a number of statistical packages (SPSS, SAS, STATISTICA, STATA, R, etc.), the problem of optimising selection of an "initial (starting) point" for an algorithm is often encountered, i.e. it is difficult to find the global minimum of function L (Lalanne & Mesbah, 2016, 2017). Thus, in most cases, the first attempt is not to apply nonlinear regression analysis, but to transform the data to suit a multiple linear regression analysis.

Based on a sample's multiple linear regression function $\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + ... + b_K X_K$, mean Y values for fixed values of independent variables are forecasted.

After constructing the regression equation, it is necessary to determine whether the regression equation obtained corresponds the data well. The smaller is the difference between the observed values Y_i and a sample's regression function-based forecasts, the better the regression function corresponds the research data. The difference is referred to as a residual error or simply a residual.

$$e_i = Y_i - \hat{Y}_i,$$

here e_i – residual; Y_i– the value observed; $\hat{Y}_i$ – regression equation-based value. The major measures of the relevance a regression model are the standard error of estimate and Adjusted R Square (r^2_{adj}). Suppose that based on the observation results $(x_{i1}, x_{i2}, ..., x_{iK}, y_i)$, $i = \overline{1,n}$, a sample's regression equation was constructed. Residual Error sum of square $SS_e = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2$, and the standard error of estimate $s_e = \sqrt{\frac{SS_e}{n-K-1}}$ describe dispersion of Y values around the regression function which is not explained by linear regression. The closer to zero is the standard error of estimate, the better is the model.

Adjusted R Square r^2_{adj} considers a sample size n and the number of independent variables in regression equation K.

$$r^2_{adj} = 1 - \frac{\frac{SS_e}{n-K-1}}{\frac{SS_p}{n-1}}.$$

It indicates which part of the variable's Y dispersion around the mean can be explained by Y linear regression with respect to independent variables $X_1, X_2, \ldots, X_K$. The closer to the unit is r^2_{adj}, the larger part of the dispersion is explained by linear regression, i.e. the better the regression function describes variable Y.

When solving regression analysis problems, the question of whether independent variable X_j affects Y variation often arises. Typically, the effects of X_j on Y variation are verfied by testing the null hypothesis:

$H_0 : \beta_j = 0$, i.e. the coefficient at population X_j in the regression equation is equal to zero. The alternative hypothesis $H_a : \beta_j \neq 0$ implies the existence of a linear relationship between X_j and Y, when j = 1,2, …, K. The hypotheses concerning the regression equation coefficients are tested by employing Student's statistics.

$$T_j = \frac{b_j}{s_{b_j}} \sim St\left(n-K-1\right), \; j = 0,1,...,K.$$

If the null hypothesis is rejected, then coefficient β_j statistically significantly differs from zero, i.e. Y values depend on X_j. Population coefficient's β_j confidence intervals (at the confidence level $1-\alpha$) are estimated by the formula:

$$b_j - t_{1-\alpha/2;n-K-1} \cdot s_{b_j} \leq \beta_j \leq b_j - t_{\alpha/2;n-K-1} \cdot s_{b_j},$$

here $t_{\alpha/2;n-K-1}$ denotes $\alpha/2$ quantile of a Student's distribution with $n–K–1$ degrees of freedom, when j = 0,1, …, K.

When comparing β_j coefficients, we cannot evaluate the relative significance of variables X_j by forecasting because the magnitude of β_j depends on X_j measurements and data distribution. Therefore, a standardized linear regression function is often sought. Dependent variable Y and independent variables $X_1, X_2, ..., X_K$ are replaced by z-values, and the least squares method is employed to estimate the standardized coefficients $BETA_j$, $j = \overline{1,K}$. The standardized coefficients $BETA_j$ indicate which variable X_j has a greater impact on Y forecast. $BETA_j$ higher by the absolute value indicates a greater Y's dependence on X_j (Čekanavičius & Murauskas, 2002).

The standard regression error is used to define the intervals within which individual Y values or the mean of Y values fall with a particular degree of reliability. We will explain how the mean of dependent variable Y is forecasted in the case of one-variable regression analysis. With a fixed X, the interval of Y mean values can be estimated. If the value of independent variable $X = X_p$ is fixed, then the confidence interval for Y's conditional

mean $E(Y / X = X_p)$ is estimated by the formula: here

$$\hat{Y} - t_{1-\alpha/2,n-2} \cdot S_{\hat{Y}} \le E(Y / X = X_p) \le \hat{Y} - t_{\alpha/2,n-2} \cdot S_{\hat{Y}}$$

$$S_{\hat{Y}} = S_e \sqrt{\frac{1}{n} + \frac{\left(X_p - \overline{X}\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}}.$$

The width of the confidence interval is affected by: confidence level $1-\alpha$, data distribution, sample size n, distance between point *Xp* and X mean. Similarly, the confidence interval for the forecasted individual Y values is estimated, only it is always wider than the confidence interval estimated for the conditional Y mean.

The model of multiple regression analysis is best suited to forecast when all independent variables X_1, X_2,..., X_K strongly correlate with Y, but do not correlate or weakly correlate with each other. When strong correlations between independent variables X_1, X_2,...,X_K are observed, the problem of mulicollinearity is encountered. Due to multicollinearity of the variables, the effects of the correlating variables on Y forecast cannot be well distinguished, the "wrong" sign of the regression function coefficient is obtained, and the coefficients of the regression equation become extremely unstable – several additional or removed observations can change them significantly (Judd et al., 2017). Multicollinearity of the variables is estimated by employing various statistics, Variance Inflation Factor VIF_j being the most common. If $4<VIF_j<10$, it proposes that variable X_j is multicollinear (medium to strong); if $VIF_j>=10$, then variable X_j is "excesivelly multicollinear" with Tolerance $TOL_j=1/VIF_j$. If $0.1<TOL_j<0.25$, it proposes that variable X_j is multicollinear (medium to strong); if $TOL_j<=0.1$, then variable X_j is "excesivelly multicollinear" (Darlington, 2017).

There are no universal methods for reducing multicollinearity, and there is no consensus on this issue in previously published studies. Authors often suggest increasing the sample, abandoning a part of multicollinear variables, replacing variables with major components, and other methods (Desboulets, 2018).

When a sample is small, even a single, very different observation can statistically significantly alter the values of the regression equation coefficients. Thus, when constructing regression models, it is important to identify any variances in the data. Currently, a number of variance identification methods are applied (Darlington & Hayes, 2017). Variances are identified by comparing the impact measure values, estimated for each observation, with the marginal values.

Here we provide some of the measures most commonly used in regression models. Variances are identified by employing the standardized residual, which is obtained from residual e_i by extracting the arithmetic mean of the sample of residues and dividing by standard deviation. The mean of the standardized residual is equal to 0, while standard deviation is equal to 1. An observation is considered a variance if the absolute value of the standardized residual exceeds 3 standard deviations.

The Cook's effect measure ($CookD_i$) indicates variance in the forecast when the ith observation is eliminated (Lalanne & Mesbah, 2017). If Cooks $D_i>4/n$, then the ith observation is considered a variance**,** i = 1,...,n; here n denotes a sample size.

Leverage h_i and **centered leverage ch_i.** $Ch_i = h_i - 1/n$ estimates the distance of the ith observation $(x_{1i}, x_{2i}, ..., x_{Ki})$ to the "centre" $(\overline{x}_1, \overline{x}_2, ..., \overline{x}_K)$. An observation is considered a variance if $ch_i>2(K+1)/n$ (for large samples) or $ch_i>3(K+1)/n$ (for small samples) (Centers for Disease Control and Prevention, 2020; World Health Organization, 2020); here K denotes a number of independent variables in the regression equation.

A studentized residual $e_i^s = \frac{e_i}{\sqrt{s_{(i)}^2(1-h_i)}}$, here $s_{(i)}$ denotes standard deviation when the ith observation is eliminated. Observation $(x_{1i}, x_{2i}, ..., x_{Ki}, y_i)$ can be considered a variance if the studentized residual $\left|e_i^s\right| > 2$. A more liberal approach can also be followed: an observation is considered a variance when $\left|e_i^s\right| > 3$.

The effect measure $DfFit_i$ indicates the effect of eliminating the ith observation on the forecasted value $\hat{Y}_i$, $DfFit_i = \hat{Y}_i - \hat{Y}_{i(i)}$, here $\hat{Y}_{i(i)}$ denotes a forecast that is obtained by the regression equation when the ith observation is eliminated. To identify the variances, the standardized $DfFit_i$ value $Std.DfFit_i$ is employed. If $\left|Std.DfFit_i\right| > 2 \cdot \sqrt{(K+1)/n}$, then the ith observation is considered a variance, and its elimination affects forecast $\hat{Y}_i$.

The effect measure $CovRatio_i$ – a covariance ratio – indicates the effect of eliminating the ith observation on the covariance matrix determinant. Elimination of the observations with $CovRatio_i$ value around 1 has an insignificant impact on the regression equation coefficients, while the observations with the values that do not fall into the interval (1–3K/n; 1+3K/n) are influential.

The effect measure $DfBetas_{ji}$ indicates the effect of the ith observation on regression coefficient β_j. $DfBetas_{ji} = b_j - b_{j(i)}$, here $b_{j(i)}$ denotes point estimate of coefficient β_j when the ith observation is eliminated, j=0,1,…,K. To identify the variances, standardized $DfBetas_{ji}$ is employed. If $\left|Std.DfBetas_{ji}\right| > 2/\sqrt{n}$, then the ith observation is considered a variance, and its elimination affects point estimate b_j of regression coefficient β_j, and at the same time – the conclusions based on this coefficient.

The alternatives to the linear regression analysis are applied when the data do not meet the model presumptions. There is no need to abandon linear regression for the slightest violation of the presumptions because the results of this analysis usually best reveal potential

dependencies and effect trends. Linear regression is sufficiently resistant to violations of the presumptions, and its result analysis is much more comprehensive. In addition, data transformation can always be employed. After logarithmizing the traits, all variances often disappear, and the data become similar to normal. Linear regression alternatives should only be employed when data transformations do no work.

In the case of the violated homoscedasticity presumption (Astivia & Zumbo, 2019), a regression of stabilized residual errors is applied; in the case when the data possess variances, robust regression is applied; in the case of the violated normality presumption, quantile regression is applied; in the case when trait dependencies are non-linear, the non-linear regression is applied.

Robust regression is only robust to variances. Observations are assigned particular weights which are considered when estimating coefficient values. If one observation is very different from the others, its weight is close to zero. In other words, then variances do not have any significant impact on the values of the regression equation coefficients. When applying a robust regression analysis, the weights assigned should be verified. This allows to subjectively assess whether estimation results are acceptable. The weight scale ranges from 0 to 1. If the weight is small, then the effect of an observation on value estimation is relatively insignificant, i.e. an observation is treated as uncharacteristic of the population. When the weight is equal to 0, an observation is not used for value estimation, i.e. it is eliminated from the regression analysis. However, it is necessary to check the reasons why this was done.

The idea of quantile regression is to model the quantile value instead of the mean value of a dependent variable with respect to the regressors selected (Furno & Vistocco, 2018). It is recommended to use when the data have variances and / or the presumption of normality is violated. Heteroscedasticity has little effect on quantile regression. Quantiles are insensitive to variances, but when the data possess significant variances, quantile regression is not applied. Variances affect the parameter estimates because when estimating the values of the regression coefficients, the difference between the observations and quantile (in the classical regression – the mean) is considered.

Non-linear regression is applied when the model cannot be made linear by transforming variables (Miguez et al., 2018).

Table 2. Examples of non-linear regression models

Name	Expression
Asymptomatic regression	b1 + b2 × exp(b3 × x)
Asymptomatic regression	b1 – (b2 × (b3 ** x))
Density	(b1 + b2 × x)** (–1/b3)
Gauss	b1 × (1 – b3 × exp(–b2 × x**2))
Gompertz	b1 × exp(–b2 × exp(–b3 × x))

In non-linear regression, the expression of the model is selected by a researcher. All values of unknown parameters are found iteratively. The initial values of unknown parameters are selected by a researcher. All estimates are sensitive to variances. Several examples of non-linear regression models are presented in Table 2.

Conclusions

Literature analysis revealed that many questions concerning the transfer of SARS-CoV-2 with aerosol particles still remain unanswered. Unanswered questions directly related to the project are as follows: How does the ability of an airborne virus to infect diminish? How do speaking, coughing and breathing affect viral emissions? What is the mechanism of aerosol generation and how does the amount of contaminated aerosols change when providing different dental services? How does a doctor's risk of getting infected with the virus depend on the concentration of aerosols, the type of personal protective equipment, and the length of time in the infected environment? What engineering measures are most effective in preventing the infection caused by breathing and touching contaminated surfaces? Ultraviolet germicidal irradiation (UVGA) is known to be effective in eradicating viruses in the air and on surfaces, but little is known about the effectiveness of UVGA in eradicating SARS-CoV-2, including the dose of UVGA required to eradicate the virus, depending on temperature, humidity, and other traits.

When evaluating the effectiveness of the measures undertaken to prevent transmission of respiratory viruses (SARS-CoV-2 causing COVID-19) through air droplets, two major characteristics – respiratory virus concentration and survival (in the air or on surfaces) – are considered. The studies are focused on the dependencies of the above-mentioned characteristics on a number of quantitative and/or qualitative traits: type of aerosols; UV irradiation characteristics (a UV source type, wavelength, beam intensity, irradiation duration, irradiation dose); characteristics of an irradiated object (temperature, length, width, volume, area, age, condition); type and degree of risk of a dental service, characteristics of a dental office (area, volume, air ventilation system, air ventilation intensity, air temperature, air humidity). In virtually all studies, the respiratory virus (SARS-CoV-2 causing COVID-19) concentration or survival rate dependence was examined for only 1-3 characteristics. Other traits were considered fixed or not mentioned at all.

When evaluating the respiratory virus (SARS-CoV-2 causing COVID-19) concentration and survival rate dependence on a number of traits, the methods of descriptive statistics, confidence intervals, hypothesis testing, dispersion analysis, trait dependence analysis, and regression analysis are employed. All the above-listed methods were tested under laboratory conditions and thus can be applied to evaluate the effectiveness of the

project product – a device designed to prevent transmission of respiratory viruses through air droplets. Selection of a particular method depends on a set of traits to be analysed, a trait type (quantitative, qualitative), a trait distribution type, and parameters. The report provides a brief description of the methods selected as well as recommendations and presumptions for their relevant application. All of the methods selected are implemented in software packages for data analysis (SAS, R, SPSS).

The analysis of mathematical methods has shown that mathematical methods are essential for the development of economic algorithms in the development of new products in the fight against the COVID-19 pandemic.

Taking into account the characteristics of the application of the algorithm, it is found that the less time the user spends using additional COVID-19 destruction equipment, the lower the cost of using such equipment in the final version.

References

Abkarian, M., Mendez, S., Xue, N., Yang, F., & Stone, H. A. (2020). Speech can produce jet-like transport relevant to asymptomatic spreading of virus. *Proceedings of the National Academy of Sciences*, *117*(41), 25237–25245. https://doi.org/10.1073/pnas.2012156117

Aggarwal, R., & Ranganathan, P. (2017). Common pitfalls in statistical analysis: Linear regression analysis. *Perspectives in Clinical Research*, *8*(2), 100–102. https://doi.org/10.4103/2229-3485.203040

Arguelles, P. (2020). *Estimating UV-C sterilization dosage for COVID-19 pandemic mitigation efforts*. Preprint.

Astivia, O. L. O., & Zumbo, B. D. (2019). Heteroscedasticity in multiple regression analysis: What it is, how to detect it and how to solve it with applications in R and SPSS. *Practical Assessment, Research, and Evaluation*, *24*, 1.

Bizzoca, M. E., Campisi, G., & Muzio, L. L. (2020). An innovative risk-scoring system of dental procedures and safety protocols in the COVID-19 era. *BMC Oral Health*, *20*(1), 1–8. https://doi.org/10.1186/s12903-020-01301-5

Botta, S. B., de Sá Teixeira, F., Hanashiro, F. S., de Araújo, W. W. R., Cassoni, A., & da Silveira Salvadori, M. C. B. (2020). Ultraviolet-C decontamination of a dental clinic setting: Required amount of UV light. *Brazilian Dental Science*, *23*(2), 1–10. https://doi.org/10.14295/bds.2020.v23i2.2275

Buonanno, M., Welch, D., Shuryak, I., & Brenner, D. J. (2020). Far-UVC light (222 nm) efficiently and safely inactivates airborne human coronaviruses. *Scientific Reports*, *10*(1), 1–8. https://doi.org/10.1038/s41598-020-67211-2

Čekanavičius, V., & Murauskas, G. (2002). *Statistika ir jos taikymai* [Statistics and its Applications]. TEV.

Čekanavičius, V., & Murauskas, G. (2014). *Taikomoji regresinė analizė socialiniuose tyrimuose* [Applied regression analysis in social research]. Vilnius University Publishing.

Centers for Disease Control and Prevention. (2020). *Guidance for dental settings: Interim infection prevention and control guidance for dental settings during the COVID-19 response*. https://www.cdc.gov/coronavirus/2019-ncov/hcp/dental-settings.html

Chin, A., Chu, J., Perera, M., Hui, K., Yen, H. L., Chan, M., Peiris, M., & Poon, L. (2020). *Stability of SARS-CoV-2 in different environmental conditions*. MedRxiv. https://doi.org/10.1016/S2666-5247(20)30003-3

Darlington, R. B., & Hayes, A. F. (2017). *Regression analysis and linear models*. Guilford.

Derruau, S., Bouchet, J., Nassif, A., Baudet, A., Yasukawa, K., Lorimier, S., Prêcheur, I., Bloch-Zupan, A., Pellat, B., Chardin, H., & Jung, S. (2021). COVID-19 and dentistry in 72 questions: An overview of the literature. *Journal of Clinical Medicine*, *10*(4), 779. https://doi.org/10.3390/jcm10040779

Desboulets, L. D. D. (2018). A review on variable selection in regression analysis. *Econometrics*, *6*(4), 45. https://doi.org/10.3390/econometrics6040045

Dos Santos, T., & de Castro, L. F. (2021). Evaluation of a portable Ultraviolet C (UV-C) device for hospital surface decontamination. *Photodiagnosis and Photodynamic Therapy*, *33*, 102161. https://doi.org/10.1016/j.pdpdt.2020.102161

Dutton, G. (2020). *UV-C Light Kills SARS-CoV-2, Triggering Novel Lighting Options for Public Spaces*. Retrieved March 01, 2021, from https://www.biospace.com/article/uv-c-light-kills-sars-cov-2-triggering-novel-lighting-options-for-public-spaces/

Ehtezazi, T., Evans, D. G., Jenkinson, I. D., Evans, P. A., Vadgama, V. J., Vadgama, J., Jarad, F., Grey, N., & Chilcott, R. P. (2021). SARS-CoV-2: Characterisation and mitigation of risks associated with aerosol generating procedures in dental practices. *British Dental Journal*, 1–7. https://doi.org/10.1038/s41415-020-2504-8

Elliott, A. C., & Woodward, W. A. (2015). *Mastering SAS for data analytics*. John Wiley & Sons.

Furno, M., & Vistocco, D. (2018). *Quantile regression: Estimation and simulation* (Vol. 2). John Wiley & Sons.

Ge, Z. Y., Yang, L. M., Xia, J. J., Fu, X. H., & Zhang, Y. Z. (2020). Possible aerosol transmission of COVID-19 and special precautions in dentistry. *Journal of Zhejiang University-SCIENCE B*, *21*, 361–368. https://doi.org/10.1631/jzus.B2010010

Gilbert, R. M., Donzanti, M. J., Minahan, D. J., Shirazi, J., Hatem, C. L., Hayward-Piatkovskyi, B., Dang, A. M., Nelson, K. M., Bothi, K. L., & Gleghorn, J. P. (2020). Mask reuse in the COVID-19 pandemic: Creating an inexpensive and scalable ultraviolet system for filtering facepiece respirator decontamination. *Global Health: Science and Practice*, *8*(3), 582–595. https://doi.org/10.9745/GHSP-D-20-00218

Innes, N., Johnson, I. G., Al-Yaseen, W., Harris, R., Jones, R., McGregor, S. KC. S., Robertson, M., Wade, W. G., & Gallagher, J. E. (2020). A systematic review of droplet and aerosol generation in dentistry. *Journal of Dentistry*, *105*, 103556. https://doi.org/10.1016/j.jdent.2020.103556

Izzetti, R., Nisi, M., Gabriele, M., & Graziani, F. (2020). COVID-19 transmission in dental practice: Brief review of preventive measures in Italy. *Journal of Dental Research*, *99*(9), 1030–1038. https://doi.org/10.1177/0022034520920580

Judd, C. M., McClelland, G. H., & Ryan, C. S. (2017). *Data analysis: A model comparison approach to regression, ANOVA, and beyond* (3rd ed.). Routledge. https://doi.org/10.4324/9781315744131

Kenarkoohi, A., Noorimotlagh, Z., Falahi, S., Amarloei, A., Mirzaee, S. A., Pakzad, I., & Bastani, E. (2020). Hospital indoor air quality monitoring for the detection of SARS-CoV-2 (COVID-19) virus. *Science of the Total Environment*, *748*, 141324. https://doi.org/10.1016/j.scitotenv.2020.141324

Khaiboullina, S., Uppal, T., Dhabarde, N., Subramanian, V. R., & Verma, S. C. (2021). Inactivation of human coronavirus by titania nanoparticle coatings and UVC radiation: Throwing light on SARS-CoV-2. *Viruses*, *13*(1), 19. https://doi.org/10.3390/v13010019

Lalanne, C., & Mesbah, M. (2016). *Biostatistics and computer-based analysis of health data using R*. Elsevier.

Lalanne, C., & Mesbah, M. (2017). *Biostatistics and computer-based analysis of health data using SAS*. Elsevier.

Lin, C. Y., & Li, C. S. (2002). Control effectiveness of ultraviolet germicidal irradiation on bioaerosols. *Aerosol Science and Technology*, *36*(4), 474–478. https://doi.org/10.1080/027868202753571296

Lindsley, W. G., Blachère, F. M., Burton, N. C., Christensen, B., Estill, C. F., Fisher, E. M., Martin, S. B., Mead, K. R., Noti, J. D., & Seaton, M. (2020). COVID-19 and the workplace: Research questions for the aerosol science community. *Aerosol Science and Technology*, *54*(10), 1117–1123. https://doi.org/10.1080/02786826.2020.1796921

Matys, J., & Grzech-Leśniak, K. (2020). Dental aerosol as a hazard risk for dental workers. *Materials*, *13*(22), 5109. https://doi.org/10.3390/ma13225109

McHugh, M. L. (2011). Multiple comparison analysis testing in ANOVA. *Biochemia Medica*, *21*(3), 203–209. https://doi.org/10.11613/BM.2011.029

Miguez, F., Archontoulis, S., & Dokoohaki, H. (2018). Nonlinear regression models and applications. In B. Glaz & K. M. Yeater (Eds.), *Applied statistics in agricultural, biological, and environmental sciences* (pp. 401–447). John Wiley & Sons, Inc.

Morawska, L., Tang, J. W., Bahnfleth, W., Bluyssen, P. M., Boerstra, A., Buonanno, G., Cao, J., Dancer, S., Floto, A., Franchimon, F., Haworth, Ch., Hogeling, J., Isaxon, Ch., Jimenez, J. L., Kurnitski, J., Li, Y., Loomans, M., Marks, G., Marr, L C., Mazzarella, L., Krikor Melikov, A., Miller, S., Milton, D. K., Nazaroff, W., Nielsen, P. V., Noakes, C., Peccia, J., Querol, X., Sekhar, Ch., Seppänen, O., Tanabe, S.-i., Tellier, R., Tham, K. W., Wargocki, P., Wierzbicka, A., & Yao, M. (2020). How can airborne transmission of COVID-19 indoors be minimized? *Environment International*, *142*, 105832. https://doi.org/10.1016/j.envint.2020.105832

Panov, V., & Borisova-Papancheva, T. (2015). Application of ultraviolet light (UV) in dental medicine. *Journal of Medical and Dental Practice*, *2*(2), 194–200. https://doi.org/10.18044/MedInform.201522.194

Rotomskis, R., & Streckytė, G. (2007). Savitoji fluorescencija. Iš *Fluorescencinė diagnostika biomedicinoje* [Fluorescence diagnostics in biomedicine] (pp. 78–123). Vilnius University Publishing.

Sabino, C. P., Ball, A. R., Baptista, M. S., Dai, T., Hamblin, M. R., Ribeiro, M. S., Santos, A. L., Sellera, F. P., Tegos, G. P., & Wainwright, M. (2020). Light-based technologies for management of COVID-19 pandemic crisis. *Journal of Photochemistry and Photobiology B: Biology*, *212*, 111999. https://doi.org/10.1016/j.jphotobiol.2020.111999

Tseng, C. C., & Li, C. S. (2005). Inactivation of virus-containing aerosols by Ultraviolet Germicidal Irradiation. *Aerosol Science and Technology*, *39*(12), 1136–1142. https://doi.org/10.1080/02786820500428575

Wilson, N. M., Norton, A., Young, F. P., & Collins, D. W. (2020). Airborne transmission of severe acute respiratory syndrome coronavirus-2 to healthcare workers: A narrative review. *Anaesthesia*, *75*(8), 1086–1095. https://doi.org/10.1111/anae.15093

Van Doremalen, N., Bushmaker, T., Morris, D. H., Holbrook, M. G., Gamble, A., Williamson, B. N., Tamin, A., Harcourt, J. L., Thornburg, N. J., Gerber, S. I., Lloyd-Smith, J. O., de Wit, E., & Munster, V. J. (2020). Aerosol and surface stability of SARS-CoV-2 as compared with SARS-CoV-1. *New England Journal of Medicine*, *382*(16), 1564–1567. https://doi.org/10.1056/NEJMc2004973

World Health Organization. (2020). *Infection prevention and control during health care when coronavirus disease (COVID-19) is suspected or confirmed: Interim guidance, 29 June 2020* (WHO/2019-nCoV/IPC/2020.4).

Yang, M., Chaghtai, A., Melendez, M., Hasson, H., Whitaker, E., Badi, M., Sperrazza, L., Godel, J., Yesilsoy, C., Tellez, M., Orrego, S., Montoya, C., & Ismail, A. (2021). Mitigating saliva aerosol contamination in a dental school clinic. *BMC Oral Health*, *21*(1), 1–8. https://doi.org/10.1186/s12903-021-01417-2

Zhang, R., Li, Y., Zhang, A. L., Wang, Y., & Molina, M. J. (2020). Identifying airborne transmission as the dominant route for the spread of COVID-19. *Proceedings of the National Academy of Sciences*, *117*(26), 14857–14863. https://doi.org/10.1073/pnas.2009637117