

1 **Whole-Genome Sequences of the Severe Acute Respiratory Syndrome Coronavirus-2**
2 **obtained from Romanian patients between March and June of 2020**

3 **Authors:**

- 4 1. Mihaela Lazar, Cantacuzino Military-Medical Research and Development National Institute
5 Bucharest, Romania
6 2. Odette Popovici, [Romanian National Institute Of Public Health](#), Bucharest, Romania
7 3. Barbara Mühlemann, Institute of Virology, Charité - Universitätsmedizin Berlin
8 4. Tim Durfee – DNASTAR Inc, USA
9 5. Razvan Stan - Cantacuzino Military-Medical Research and Development National Institute
10 Bucharest, Romania

11
12 Corresponding author's name and mailing address, telephone number, and e-mail address:
13 Mihaela Lazar, Cantacuzino Military-Medical Research and Development National Institute,
14 Splaiul Independentei 103, Bucharest, Romania, lazar.mihaela@cantacuzino.ro, +40 213069116.

15
16 **Abstract**

17 Impact of mutations on the evolution of Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) are
18 needed for ongoing global efforts to track and trace the current pandemic, in order to enact effective prevention and
19 treatment options. SARS-Co-V-2 viral genomes were detected and sequenced from 18 Romanian patients suffering
20 from coronavirus disease-2019. Viral Spike S glycoprotein sequences were used to generate model structures and
21 assess the role of mutations on protein stability. We integrated the phylogenetic tree within the available European
22 SARS-Co-V-2 genomic sequences. We further provide an epidemiological overview of the pre-existing conditions
23 that are lethal in relevant Romanian patients. Non-synonymous mutations in the viral Spike glycoprotein relating to
24 infectivity are constructed in models of protein structures. Continuing search to limit and treat SARS-CoV-2 benefit
25 from our contribution in delineating the viral Spike glycoprotein mutations, as well as from assessment of their role
26 on protein stability or complex formation with human receptor angiotensin-converting enzyme 2. Our results help
27 implement and extend worldwide genomic surveillance of coronavirus disease-2019.

28 Keywords: SARS-CoV-2; COVID-19; genome sequencing; Spike glycoprotein; surveillance.

29
30
31
32

33 **Introduction**

34 In December 2019, in China, an outbreak started due to a novel coronavirus strain that causes a
35 severe illness, the coronavirus disease 2019 (COVID-19). The virus, SARS-CoV-2, has since
36 spread globally [1]. Data reported to European Center for Disease Control show that clinical
37 presentation of COVID-19 ranges from no symptoms to severe pneumonia, and that many cases
38 have led to death [2]. According to available data, 32 % of the diagnosed COVID-19 cases in the
39 EU/EEA have been hospitalized, and 4% thereof had severe form of illness [3]. Additionally,
40 hospitalization rates have been significantly higher for adults 65 years and older [4]. The first
41 positive case of infection with SARS-CoV-2 in Romania was confirmed on 26th of February
42 2020. With the ongoing spread of the virus, it is becoming critical to detect and classify the virus
43 in patient samples. As such, full genome characterization of this virus is instrumental for
44 updating diagnostics criteria and assessing viral evolution. To assess its genetic variation in a
45 South-Eastern European population, we herein generated genome sequences using metagenomic
46 sequencing, from 18 Romanian patients. We delineated the non-synonymous mutations and
47 constructed models of the protein structures from these sequences. We explicitly focused on the
48 spike (S) glycoprotein because it mediates infection of human cells, and is the main target of
49 most vaccine strategies and antibody-based therapeutics [5].

50 **Materials and Methods**

51 Collection and processing of samples

52 Nasopharyngeal and oropharyngeal swabs were collected from individuals from different
53 Romanian geographical areas. Total nucleic acid extraction was performed using the Maxwell®
54 RSC Viral Total Nucleic Acid Purification Kit (Promega, USA) as described by the
55 manufacturer. SARS-CoV-2 nucleic acid was detected using the E gene assay as the first-line

56 screening tool, followed by confirmatory testing with the RdRp gene assay [6]. The RiboZero
57 Gold rRNA depletion protocol was used to remove human cytoplasmic and mitochondrial rRNA.
58 The total RNA quantity and integrity were measured with the Qubit RNA Assay Kit (Invitrogen,
59 Carlsbad, CA, USA). The TrueSeq Stranded Total RNA Library Prep Gold kit along with the
60 IDT for Illumina TruSeq RNA UD Indexes were used for sequencing-ready library preparation.
61 RNA fragmentation, first and second-strand cDNA synthesis, adenylation, adapter ligation and
62 amplification were done according to the TruSeq Stranded Total RNA protocol. After
63 amplification, the prepared libraries were quantified, pooled and loaded onto Illumina MiSeq
64 DNA Sequencer. Sequence data from each sample was aligned to the SAR-CoV-2 reference
65 genome (GenBank accession number: NC_045512.2) and variants called using SeqMan NGen
66 (DNASTAR, Madison, WI, USA). Alignments were manually inspected to confirm variant calls
67 and the viral sequence from each sample exported with SeqMan Pro (DNASTAR).

68 Data mining for structures

69 Atomic coordinates of the template, the novel SARS-CoV-2 spike protein in complex with
70 Receptor Binding Domain (RBD) of the human Angiotensin-converting enzyme 2 (hACE2),
71 PDB: 6LZG, were retrieved from RCSB protein data bank. Spike protein sequences from our
72 genomes were modeled using I-Tasser [7] and the quality of models of Spike S proteins was
73 assessed with MolProbity [8]. Full length models were superimposed over template and root
74 mean square deviations (RMSD) of carbon alpha backbones were determined using PyMol
75 (PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC).

76 Mutational analyses and molecular docking

77 Site Directed Mutator (<http://marid.bioc.cam.ac.uk/sdm2/>) was used to assess the impact of non-
78 synonymous single, double or triple mutations on the protein stability and to determine the $\Delta\Delta G$

79 values for each protein. For mutations in the Spike protein binding site to RBD from ACE2, the
80 difference in protein stability between Spike protein:RBD complex and Spike protein alone is
81 indicated. For in silico docking, the HawkDock server (<http://cadd.zju.edu.cn/hawkdock/>) was
82 used to conduct docking simulations between our modeled Spike S protein structures and ACE2.
83 All molecular structures were visualized with PyMol.

84

85 Results

86 We performed real time RT-PCR tests on 32233 individuals, starting on 16.02.2020 until
87 22.06.2020, and summarized the data in Table 1. We compared our data to public information on
88 national testing relevant for this pandemic, valid at the time of writing.

89 **Table 1. Aggregate of the testing data performed at Cantacuzino Military-Medical National Research-**
90 **Development Institute (INCDMMC) and at national level [9]**

As of 22.06.2020	INCDMMC data	National data
Total number of tests	32233	626330
Confirmed SARS-CoV-2 (% from total tested)	1630 (5.05 %)	24045 (3.83%)
Negative SARS-CoV-2 (% from total tested)	30603 (94.95 %)	602285 (96.17 %)

91

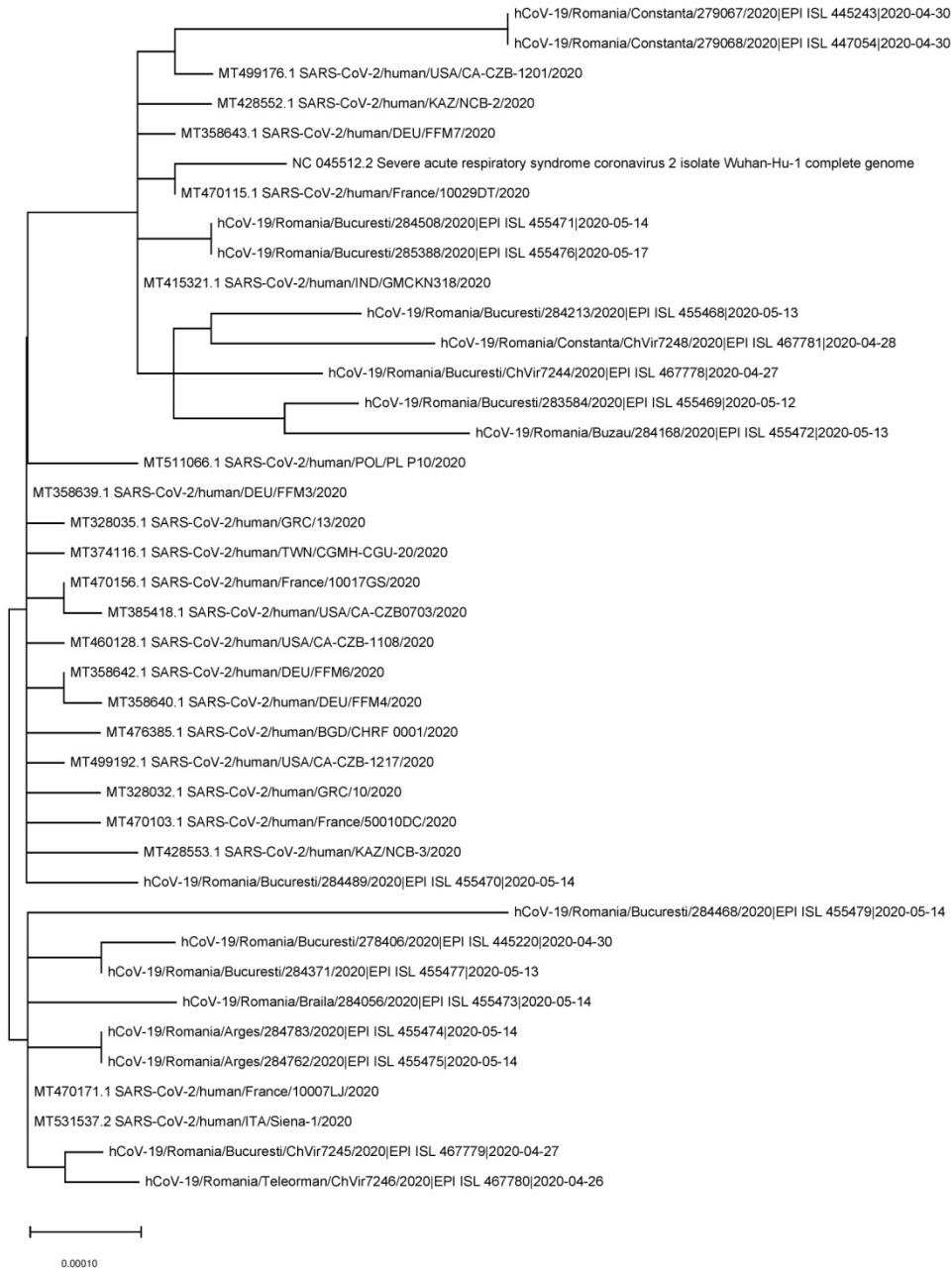
92 A recent analysis of risk factors in Romanian patients was performed for a subset of 842
93 deceased individuals or 7.1 % out of a subset of 11790 confirmed cases (data accessed on May
94 23rd, 2020) [10]. For both sexes, patients with pre-existing cardiovascular conditions were 2.89
95 times more likely to succumb than patients who were not affected by this type of illness.
96 Importantly, the next categories of patients who died were suffering from chronic renal diseases
97 (death more probable by 2.69) and diabetes (by 2.3 times). Given the high expression and
98 circulating levels of hACE2 in the heart and kidneys, these findings support a mechanism
99 whereby Spike glycoprotein may attach to hACE2 in a dose-response manner. We also note that
100 these and other high risk factors (e.g. chronic lung disease or cancer/other immune deficiencies)

101 are not equally distributed between sexes, with men having the highest risks when suffering from
102 cancer or other immune deficiencies (Odds Ratio of 3.25), while women succumbing mostly
103 when suffering with chronic renal diseases (Odds Ratio of 4.3). This finding may reflect possible
104 gender differences in the levels and activity of ACE2, as has been observed in murine models
105 [11].

106

107 Genetic phylogeny

108 The sequences of Romanian SARS-CoV-2 sequences showed high (~ 99.95 identity with Wuhan
109 seafood market pneumonia virus (Genbank accession number: NC_045512.2) and >99.6% with
110 sequences from France, Germany, Greece, Italy, Poland, India, USA. Phylogenetic analysis
111 showed that the Romanian sequences belonged to different clusters (Figure 1). The following
112 sequences were introduced in GISAID (Global Initiative on Sharing All Influenza Data) under
113 accession numbers: EPI_ISL_445220, EPI_ISL_445243, EPI_ISL_447054, EPI_ISL_455468,
114 EPI_ISL_455469, EPI_ISL_455470, EPI_ISL_455471, EPI_ISL_455472, EPI_ISL_455473,
115 EPI_ISL_455474, EPI_ISL_455475, EPI_ISL_455476, EPI_ISL_455477, EPI_ISL_455479,
116 EPI_ISL_467779, EPI_ISL_467780, EPI_ISL_467781, EPI_ISL_467778.



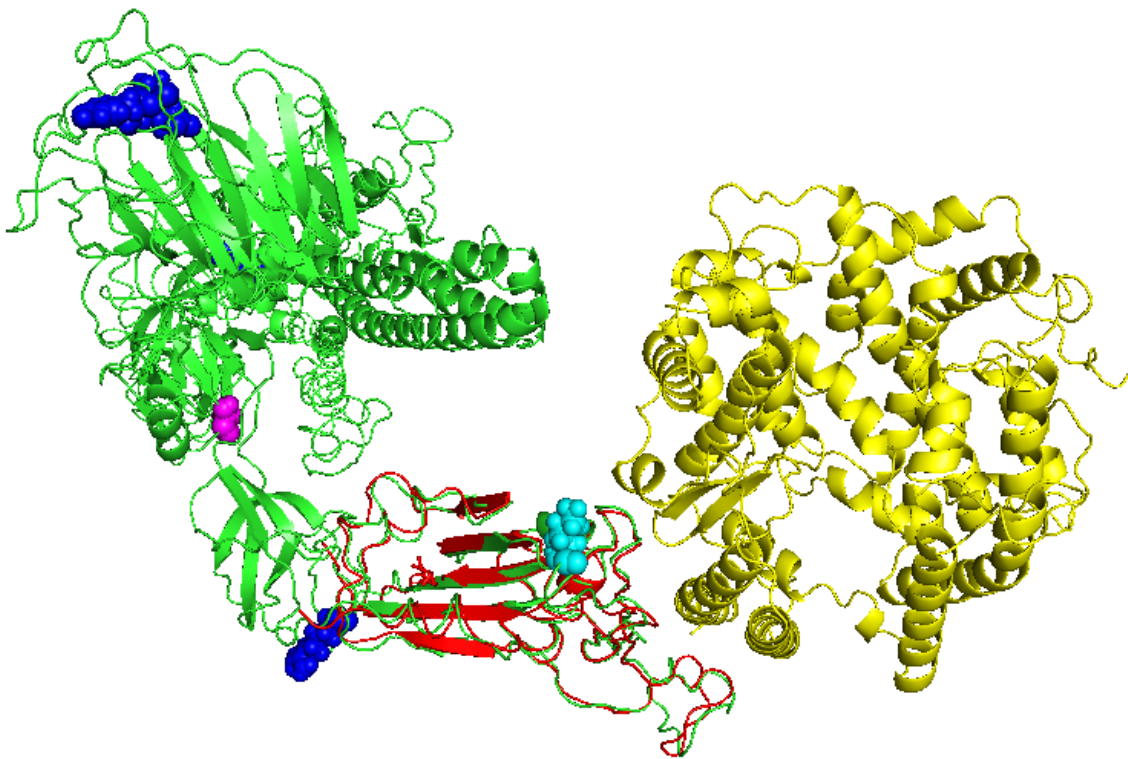
117

118 Fig. 1 Phylogenetic analysis of 18 SARS-CoV-2 complete genome sequences. The Wuhan reference genome from
119 GenBank accession number: NC_045512.2 and other European/American complete sequences are also shown from
120 different countries (n=21). The tree was built by using the best fitting substitution model (HKY) through MEGA X
121 software.

122

123 Molecular modeling

124 An overview of the 7 mutations of the Spike proteins detected from Romanian sequence is
125 presented in the structure of Figure 2 and listed in Table 2. We note that only the N439K variant
126 is involved in the binding interface, where the lysine now H-bonds with Gln 325 from the
127 hACE2. Mutants included single, double and triple Spike protein variants. The impact of the
128 mutations on the stability of the complexes made with hACE2 is also indicated in Table 2.



129
130 Figure 2: Docking of hACE2 (from PDB 6LZG, in yellow) with models of Spike glycoprotein variants (green) that
131 contains the location of mutations described in this study (blue) with respect to binding partner hACE2. D614G
132 mutation is shown in magenta; N439K mutation, that sits at the binding interface with hACE2 is shown in cyan. The
133 structural overlap between RBD from Spike protein in PDB 6LZG and a constructed model of the mutant Spike
134 protein is shown in red.

135
136 We used PDBsum to identify the critical amino acids at the binding interface, and further
137 compiled with PyMol their identity (Table 2), where residues interfacing with multiple H-bonds
138 are shown in bold. TM-scores [0-1] indicating the structural similarity to the overlapped model
139 have been derived with I-TASSER (TM-Scores of 0.5 and above represent a high probability to
140 match similar folds). The impact of the mutations on the stability of the complexes made with

141 hACE2 is also indicated in Table 2. Although all mutants have negative (favorable) binding
 142 energy, compared to the native complex, only the triple mutant has a significantly higher value
 143 (60% increase). We note that the double mutant *N439K, D614G* that directly affects the binding
 144 interface, has only negligible positive effect on binding, but ranks above the rest of the variants.

145

146

Table 2. Key residues involved in the formation of interfaces between ACE2: Spike S variants

Spike Protein	Viral Interface Residues	RMSD (Å)	TM score	Binding free energy - complex hACE2:variant Spike proteins (kcal/mol)
PDB 6LZG		-	-	-42.96
Model SARS-CoV-2 RBD [PDB: 6LZG]	K417, G446, Y449, Y453, L455, F456, A475, F486, N487, Y489, F490, Q493 , G496, Q498, T500 , A501, G502, Y505	-	-	-
<i>D614G</i> variant Spike S overlapped to SARS-CoV-2 RBD	Similar to SARS-CoV-2 RBD	0.97	0.63	-9.94
<i>P521R, D614G</i> variant Spike S overlapped to SARS-CoV-2 RBD	Similar to SARS-CoV-2 RBD	0.97	0.6	-27.25
<i>S98F, D614G</i> variant Spike S overlapped to SARS-CoV-2 RBD	Similar to SARS-CoV-2 RBD	0.99	0.58	-26.24
<i>N439K, D614G</i> variant Spike S overlapped to SARS-CoV-2 RBD	K417, K439, G446, Y449, Y453, L455, F456, A475, F486, N487, Y489, F490, Q493 , G496, Q498, T500 , A501, G502, Y505	0.96	0.61	-39.22
<i>V308L, D614G</i> variant Spike S overlapped to SARS-CoV-2 RBD	Similar to SARS-CoV-2 RBD	0.97	0.51	-28.74
<i>E96D, H245Y, D614G</i> variant Spike S overlapped to SARS-CoV-2 RBD	Similar to SARS-CoV-2 RBD	0.98	0.6	-67.21

147

148

149 **Conclusions**

150 Mutations in the spike surface glycoprotein may be conducive to conformational changes, which
151 can translate into changing antigenicity. As such, identifying the amino acids involved in
152 conformational changes of the SARS-CoV-2 spike surface glycoprotein structure is essential to
153 inform on patterns of mutations subject to positive selection. The role of detected mutations on
154 overall thermodynamic stability of the protein variants is thus imperative, as shown in Table 3.

155 **Table 3. Role of point mutations observed in the genomes on overall protein stability. $\Delta\Delta G$ parameter is a**
156 **measure of the change in energy between the folded and unfolded states caused by point mutations [12].**

Variant Spike proteins	$\Delta\Delta G$ (kcal/mol)	Effect on stability	Solvent accessibility compared to wild type
S98F	-0.77	destabilizing	0.6% increase
E96D	-1.34	destabilizing	9.2% increase
H245Y	0.78	stabilizing	25.8% increase
V308L	-1.27	destabilizing	3.2% increase
N439K	-0.2	destabilizing	20.7% increase
P521R	0.67	stabilizing	40.8% increase
D614G	-0.58	destabilizing	23% increase

157
158 Spike *D614G* variant has been the most widespread mutant encountered thus far in the 2020
159 genome datasets outside East Europe (3577 counts in GISAID database) [13]. *D614G* mutation
160 is embedded into an immunodominant antibody epitope and is recognized by monoclonal
161 antibodies isolated from recovered individuals, who had been infected with the original SARS-
162 CoV [14]. A very recent experimental study comparing Spike variants *D614G* against *D614S*
163 indicates that the former is less stable thermodynamically, which translates into markedly
164 increased infectivity of ACE-2 expressing cells, consistent with epidemiological data.¹⁵
165 Furthermore, pseudoviruses containing both of these variants were neutralized with comparable
166 efficiency by convalescent plasma [14]. However, another recent study demonstrated that Spike
167 variant *D614G* significantly infects ACE-2 expressing cells, when compared to the native Spike
168 protein, and that convalescent sera showed decreased neutralizing activity against *D614G*

169 pseudovirus [16]. Given these factors, this mutant could mediate immune avoidance, although
170 there is as-yet no definite clinical data to correlate the impact of this or any other mutation on
171 infectivity. We note that none of the other mutations we describe have been found thus far in
172 other cohorts, and that other mutations are themselves rare [13]. Tracking mutations in viral
173 Spike glycoprotein continues to be paramount for vaccine and antibody therapy strategies that
174 are currently being developed, and on evaluating new SARS-CoV-2 strains as they emerge and
175 evolve.

176

177 **Acknowledgements:**

178 We are grateful to Luiza Ustea, Nicoleta Paraschiv and Adrian Crețu for technical support. We
179 thank the staff working from the Viral Respiratory Infections Laboratory (National Influenza
180 Centre), colleagues from Cantacuzino Military-Medical Research and Development National
181 Institute, the public health staff from the National Centre for Communicable Diseases
182 Surveillance and Control, officials from the Ministry of Health and the Ministry of Defense.
183 Funding support was provided by Rompetrol Group NV (Fundatia pentru SMURD) and the
184 Romanian Association for Shoulder and Elbow Surgery. The funding sources had no role in the
185 study design, analysis or writing of report.

186

187

188 **REFERENCES**

- 189 [1] Zhou P, Yang XL, Wang XG, et al. A pneumonia outbreak associated with a new coronavirus
190 of probable bat origin. *Nature*. 2020; 579(7798):270–273.
- 191 [2] Zheng F, Tang W, Li H, Huang YX, Xie YL, Zhou ZG. Clinical characteristics of 161 cases
192 of corona virus disease 2019 (COVID-19) in Changsha. *Eur Rev Med Pharmacol Sci*. 2020;
193 24(6):3404–3410.

- 194 [3] European Center for Disease Control: Rapid Risk Assessment: Coronavirus disease 2019
195 (COVID-19) in the EU/EEA and the UK– ninth update. Retrieved on June 8th, 2020.
- 196 [4] Garg S, Kim L, Whitaker M, et al. Hospitalization Rates and Characteristics of Patients
197 Hospitalized with Laboratory-Confirmed Coronavirus Disease 2019 — COVID-NET, 14 States,
198 March 1–30, 2020. *MMWR Morb Mortal Wkly Rep.* 2020; 69:458–464.
- 199 [5] Robson B. COVID-19 Coronavirus spike protein analysis for synthetic vaccines, a
200 peptidomimetic antagonist, and therapeutic drugs, and analysis of a proposed achilles’ heel
201 conserved region to minimize probability of escape mutations and drug resistance. *Comput Biol*
202 *Med.* 2020; 121:103749.
- 203 [6] Corman VM, Landt O, Kaiser M, et al. Detection of 2019 novel coronavirus (2019-nCoV) by
204 real-time RT-PCR. *Eurosurveillance.* 2020; 25(3):2000045.
- 205 [7] Yang J, Zhang Y. I-TASSER server: new development for protein structure and function
206 predictions. *Nucleic Acids Research.* 2015; 43: W174-W181.
- 207 [8] Williams CJ, et al. MolProbity: More and better reference data for improved all-atom
208 structure validation. *Protein Science.* 2018; 27: 293-315.
- 209 [9] Informare COVID-19, Grupul de Comunicare Strategică, 13 Iunie 2020 (In Romanian).
210 [https://www.mai.gov.ro/informare-covid-19-grupul-de-comunicare-strategica-13-iunie-2020-ora-](https://www.mai.gov.ro/informare-covid-19-grupul-de-comunicare-strategica-13-iunie-2020-ora-13-00/)
211 [13-00/](https://www.mai.gov.ro/informare-covid-19-grupul-de-comunicare-strategica-13-iunie-2020-ora-13-00/). Retrieved 14.06.2020.
- 212 [10] Popovici O. Risk factors for death in confirmed cases of patients with COVID-19 (In
213 Romanian). [https://www.cnscbt.ro/index.php/analiza-cazuri-confirmate-covid19/1791-analiza-](https://www.cnscbt.ro/index.php/analiza-cazuri-confirmate-covid19/1791-analiza-epidemiologica-factori-de-risc-deces-cu-covid-19)
214 [epidemiologica-factori-de-risc-deces-cu-covid-19](https://www.cnscbt.ro/index.php/analiza-cazuri-confirmate-covid19/1791-analiza-epidemiologica-factori-de-risc-deces-cu-covid-19). Retrieved 09.06.2020.
- 215 [11] White MC, Fleeman R, Arnold AC. Sex differences in the metabolic effects of the renin-
216 angiotensin system. *Biol Sex Differ.* 2019; 10, 31.
- 217 [12] Kellogg EH, Leaver-Fay A, Baker D. Role of conformational sampling in computing
218 mutation-induced changes in protein structure and stability. *Proteins.* 2011;79(3):830–838.
- 219 [13] Korber B, et al. Spike mutation pipeline reveals the emergence of a more transmissible form
220 of SARS-CoV-2. *BioRxiv.* 2020.04.29.069054; doi: <https://doi.org/10.1101/2020.04.29.069054>.
- 221 [14] Wang Q, et al. Immunodominant SARS Coronavirus Epitopes in Humans Elicited both
222 Enhancing and Neutralizing Effects on Infection in Non-human Primates. *ACS Infectious*
223 *Diseases.* 2016; 2, 361-376.
- 224 [15] Zhang L, Jackson CB, Mou H, Ojha A, Rangarajan ES, Izard T, Farzan M, Choe H. The
225 D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity.
226 *BioRxiv.* 2020.06.12.148726; doi: <https://doi.org/10.1101/2020.06.12.148726>.
- 227 [16] Hu J, He CL, Gao Q, Zhang GJ, Cao XX, Long QX, Deng HJ, Huang LY, Chen J, Wang K,
228 Tang N, Huang AL. The D614G mutation of SARS-CoV-2 spike protein enhances viral
229 infectivity and decreases neutralization sensitivity to individual convalescent sera. *BioRxiv*
230 2020.06.20.161323; doi: <https://doi.org/10.1101/2020.06.20.161323>.