Optimizing Non-Pharmaceutical Intervention Strategies Against COVID-19 Using Artificial Intelligence

Vito Janko (vito.janko@ijs.si)
Jožef Stefan Institute https://orcid.org/0000-0002-9549-9742
Nina Reščič
Jožef Stefan Institute https://orcid.org/0000-0001-5428-3577
Aljoša Vodopija
Jožef Stefan Institute
David Susič
Jožef Stefan Institute
Carlo Maria De Masi
Jožef Stefan Institute
Tea Tušar
Jožef Stefan Institute
Anton Gradisek
Jožef Stefan Institute https://orcid.org/0000-0001-6480-9587
Sophie Vandepitte
Ghent University
Delphine De Smedt
Ghent University
Jana Javornik
Leeds University Business School
Matjaz Gams
Mitja Lustrek

Article

Keywords: COVID-19, non-pharmaceutical interventions, proposed intervention, artificial intelligence

Posted Date: November 5th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1047502/v1

Optimizing Non-Pharmaceutical Intervention Strategies Against COVID-19 Using Artificial Intelligence

Vito Janko[1,*], Nina Reščič[1,2], Aljoša Vodopija[1,2], David Susič[1,2], Carlo De Masi[1], Tea Tušar[1], Anton Gradišek[1], Sophie Vandepitte[3], Delphine De Smedt[3], Jana Javornik[4], Matjaž Gams[1,2], and Mitja Luštrek[1,2]

[1]Jožef Stefan Institute, Department of Intelligent Systems, Ljubljana, 1000, Slovenia
[2]Jožef Stefan Postgraduate School, Ljubljana, 1000, Slovenia
[3]Ghent University, Department of Public Health and Primary Care, Ghent, 9000, Belgium
[4]Leeds University Business School, Leeds, LS2 9JT, UK
[*]corresponding author(s): Vito Janko (vito.janko@ijs.si)

ABSTRACT

One key task in the early fight against the COVID-19 pandemic was to plan non-pharmaceutical interventions to reduce the spread of the infection while limiting the burden on the society and economy. With more data on the pandemic being generated, it became possible to model both the infection trends and intervention costs, transforming the creation of an intervention plan into a computational optimization problem. This paper proposes a framework developed to help policy-makers plan the best combination of non-pharmaceutical interventions and to change them over time. We developed a hybrid machine-learning epidemiological model to forecast the infection trends, aggregated the socio-economic costs from literature and expert knowledge, and used a multi-objective optimization algorithm to find and evaluate various intervention plans. The framework is modular and easily adjustable to a real-world situation, it is trained and tested with data collected from almost all countries of the world, and its proposed intervention plans generally outperform those used in real life in terms of both the number of infections and intervention costs.

Introduction

The first line of defence against the spread of the SARS-CoV-2 virus was the introduction of Non-Pharmaceutical Interventions (NPIs) by national governments. With the virus being aerosol-borne, some of the key measures included the use of face masks and restrictions on gatherings, which have often resulted in partial or full lockdowns. While efficient in reducing the numbers of infections[1,2], restrictive NPIs also presented immense Socio-Economic Costs (SECs) to the population[3]. Policy-makers were faced with an almost impossible task of carefully balancing NPI costs against the predicted NPI benefits, largely without having appropriate tools and data for evidence-based decisions.

To further add complexity to the problem, in a typical intervention plan adopted by policy-makers, a combination of NPIs would be used, each of them taking place for different periods of time. These plans were usually prepared by expert panels who had the challenge of selecting intervention plans without assurance that they would really flatten the infection curve enough to be lifted within the expected period[4,5]. Moreover, the full extent of SECs was unpredictable until a few months into the pandemic; until then, the intervention plans were primarily focused on containing the spread of the virus. While the SECs have now started affecting the decisions of policy-makers[6], they are still not sufficiently explored.

The prediction of daily infections and the impact of NPIs on the spread of the pandemic has been researched quite well[1,2], but little work has been done regarding the prescription of intervention plans. Few of the published approaches have proposed frameworks to find good intervention plans that also consider NPI costs and how to best combine NPIs. Yousefpour et al.[7], for example, proposed a framework based on SEIRD models and multi-objective optimization to prescribe NPIs. However, the optimization did not operate on real-life NPIs, and as such, this approach cannot be directly used by policy-makers. Chen et al.[8] created a linear programming tool to explore the trade-off between the expected mortality rate of COVID-19 and return to normal activities, while Yaesoubi et al.[9] developed a decision tool to determine when to trigger, continue, or stop physical distancing intervention in order to minimize both the deaths from COVID-19 and intervention duration. Both studies combined the objectives into a single function and the final result was a single intervention plan. Such approaches require a strong predefined preference on how to balance the objectives, which is often difficult to define in practice. In addition, none of the

three approaches was extensively tested on various epidemiological scenarios. For this reason, their generalization to real-world situations is unpredictable.

A more structured attempt to research the possibility of using artificial intelligence (AI) to automatically prescribe intervention plans was made by the $500K Pandemic Response Challenge[10], organized by XPRIZE and sponsored by Cognizant. The participants were tasked to find good trade-offs between the costs of NPIs and their benefits – and assemble three-month intervention plans for each territory (all countries and some sub-country regions). An approach proposed by the sponsor (Miikkulainen et al.[11]) involved the use of evolutionary algorithms to evolve neural networks that prescribe intervention plans. This approach was intended to point the way for the competitors, who would go on to develop better-performing approaches. The competition ended with two "Grand Prize Winners." One of them[12] combined two prescriptors: the first selected the most cost-effective intervention plans from a subset of possible plans with precomputed effectiveness, and the second greedily composed intervention plans from most cost-effective individual NPIs. The other winning submission – submitted by some of this paper's authors – was the starting point for the approach described here.

In this study, we developed a framework to help policy-makers design reasonable intervention strategies by dynamically adjusting NPIs. The framework is comprised of three components: a predictor based on the SEIRD epidemiological model that predicts infection trends, a compilation of SECs of NPIs, as found in the literature, and a prescriptor that finds diverse optimized intervention plans. The main methodological novelty of the predictor is that the key parameters of the SEIRD model can be dynamically adapted to any set of given NPIs using a machine learning model. Intuitively, the machine learning model decreases the disease transmission rate in the SEIRD model when strict NPIs are in place, and vice versa. In contrast to most related work, our prescriptor uses multi-objective optimization and does not combine the objectives into a single function. As such, it can find near-optimal trade-offs between the costs (SEC) and benefits (reduced number of infections) of NPIs, and presents the results in the form of a Pareto front approximation (i.e., a set of near-optimal intervention plans where no objective can be improved without making the other worse). Ideally, the obtained Pareto front approximation ranges from costly intervention plans, which significantly decrease infections, to cheap but not as effective ones – presenting a set of plans for the policy-maker to choose from. Our methodology was extensively tested: the predictor was tested on data from 194 territories and the prescriptor on data from 50 territories. It yields semantically sensible results, achieves similar or better prediction accuracy than previously proposed models, and furthermore, proposes better plans – at least based on our simulations – than those actually implemented by policy-makers in the studied period (March 2020 to April 2021).

Results

The framework is presented as a sum of its components. First, we discuss the algorithm that predicts the trends in the number of infections, taking into account the historic data together with the NPIs that were in place – for this study we consider 12 NPIs listed in Table 1, and we denote this set as *OxNPIs* as it is derived from Oxford's OxCGRT dataset[13]. Then, based on the literature data, the SECs of the OxNPIs are estimated. Bringing these two aspects together, multi-objective optimization is used to find the best trade-offs between the number of infections and the SEC of the intervention plan. Finally, we discuss a typical structure of these plans, the NPIs most/least often used in them, and both strengths and limitations of the proposed approach.

Predicting infections

The basis of the prediction model for infections is the SEIRD epidemiological model, which considers the dynamics between the pools of Susceptible, Exposed, Infectious, Recovered, and Deceased individuals[14]. This model consists of a set of differential equations where the key parameters are β, related to the probability of disease transmission per contact; the incubation period σ; and the mortality rate μ. These values can be obtained for a specific time-period/territory by fitting the model to the historical infection and mortality data (Figure 1a).

While the SEIRD model on its own is accurate in predicting the future in a "status-quo" situation (see Supplementary Information – Estimating the prediction error), it does not correctly predict the infection trends following a change of the NPIs – which is essential if the framework is to propose which NPIs to use in the future. Ideally, as the NPIs change, the parameters of the SEIRD system would be adjusted accordingly, taking into account their changed impact on the disease transmission rate. An example of such behavior can be seen in Figure 1b, as generated by our Hybrid Machine-Learning Epidemiological (HMLE) method. To achieve this dynamic forecast of infections, different machine-learning models were built to find the relations between the (β, σ, μ) parameters and OxNPIs. The details of combining the machine-learning predictions and the SEIRD model into the HMLE method are explained in the Methods section.

To assess the performance of the HMLE method, we show in Figure 2 that our predictor significantly outperforms the "standard predictor" provided by Cognizant in the second phase of the XPRIZE competition[10] (for details of this test, see Supplementary Information). The mean average error (MAE) is 5.9 times lower on day 70. To explore what contributes to the increased performance, we compared the full implementation to two additional versions of our method: 1) one that relies only on machine learning to set the parameter values of the SEIRD model without normalizing them using the last known

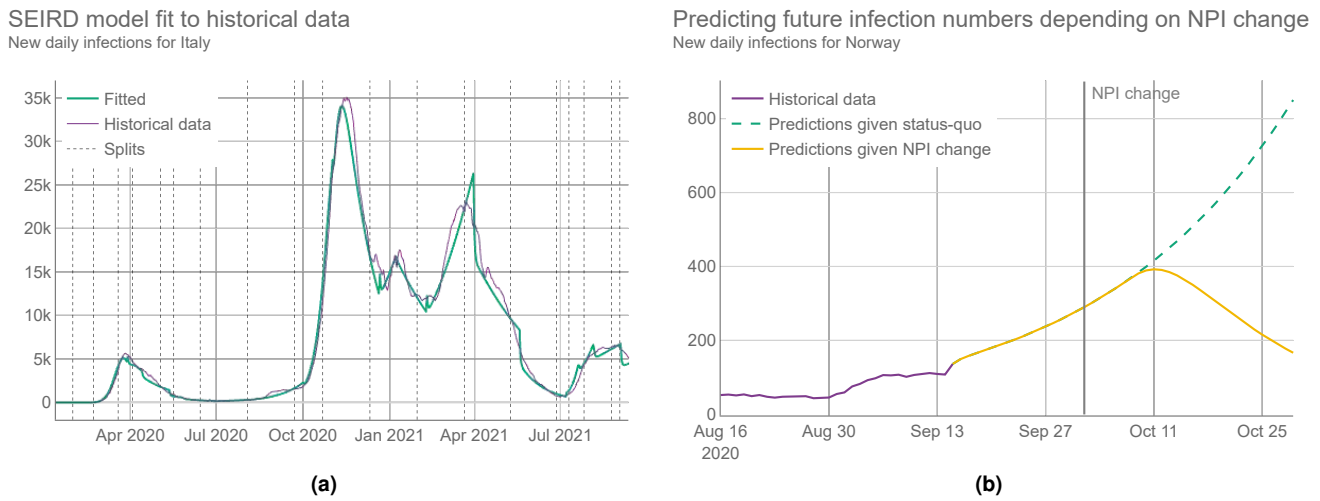

Figure 1. a) Daily infections for Italy over 20 months (purple) together with the predictions using best fitted SEIRD model (green). Fitting was conducted by first splitting the data into segments, represented by dashed vertical lines, where at least two NPIs were changed with respect to the previous segment. b) Daily infection predictions for Norway, made both by using only fitted parameters (green) and by using adapted parameters, which reflect the change to more strict NPIs (yellow).

fitted parameter values, and 2) one that retains the last known fitted parameter values throughout the forecast period, without using machine learning to account for NPI changes. The experimental results showed that the parameters predicted by the machine-learning model are less appropriate on average, than the last known fitted parameters; when normalized, however, they outperform the last known fitted parameters. The benefit of machine learning does not appear to be huge, but it is significant in case of important NPI changes, as demonstrated in Figure 1b.

Of all machine learning algorithms tested (see Supplementary Information), the Ridge classifier (a type of linear model) had the highest accuracy. Aside from prediction accuracy, the model has an additional advantage – it is easily interpretable. Figure 3 lists the coefficients corresponding to the normalized OxNPI strictness values. Given this normalization, the model's coefficient size can indicate relative NPI importance. Our model's most important intervention is the cancellation of public events, which is consistent with the related work that typically ranks it among the top NPIs[15]. Next is school closure, which additionally results in some parents staying at home, so its importance is not surprising. These two are followed by contact tracing – which is difficult to execute well, and other sources do not rate this NPI as high. In the fourth place are international travel controls, which played a big role in some countries, particularly in the early stages of the pandemic. The importance of this NPI was corroborated by Haug et al.[15] Other NPIs have notably lower coefficient values. This may come as a surprise for "C2: Workplace closing," "C4: Restrictions on gatherings," and "C6: Stay at home requirements," but it should be noted that 1) these three NPIs have a large overlap with each other and with other NPIs, and 2) they were usually instituted when the epidemiological situation was grave, with many NPIs in force simultaneously, thus making it very difficult to properly isolate the importance of each of them. This is why in these cases the assigned regression coefficient do not necessarily correctly reflect their relative importance. Nonetheless, their sum is close to the largest single coefficient. Of note, the NPI features were not the only ones included in the model, but the coefficient values of the others were an order of magnitude lower than those listed here.

Finally, for a direct comparison with related work, the HMLE model described here is an improved version of the one used in the XPRIZE challenge, which was ranking between the 1st and 4th place during the two month prediction period on real data for 235 territories[16].

Intervention costs

The implementation of each NPI incurs both economic and social costs. At the time of the research, no exhaustive list of economic costs was available. Thus, we surveyed the literature to compile the costs for each OxNPI. The economic cost values are scaled to represent the percentage of GDP loss incurred. For example, if the "C3: Cancel public events" NPI is active for one month and it has the cost of 1.4, then our method assumes that the GDP in this month is 1.4% lower than usual – note that this is not the yearly GDP loss but that for the predicted period.

The estimation of social costs (i.e., the decrease in wellbeing of the general population due to social isolation, restriction of freedom, and similar caused by different NPIs), is difficult as no reliable and exhaustive paper has been found on the topic. Instead, we relied on estimates made by a domain expert, making an aggregation of domain knowledge and available scientific

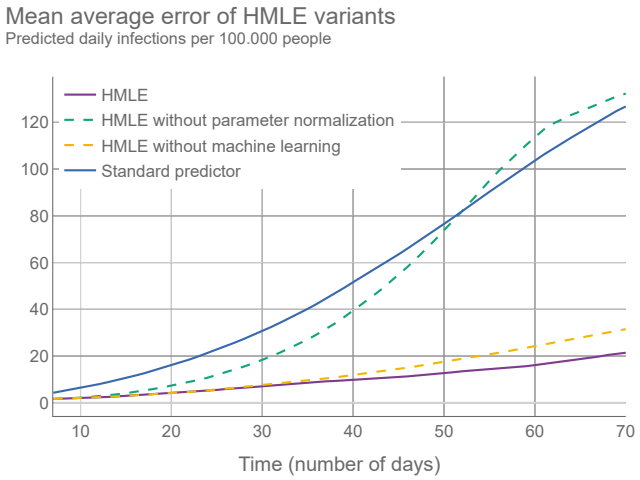

Figure 2. Different versions of the HMLE method compared to the "standard predictor"[10]. Testing was conducted on 50 random time intervals for each of the selected 194 territories.

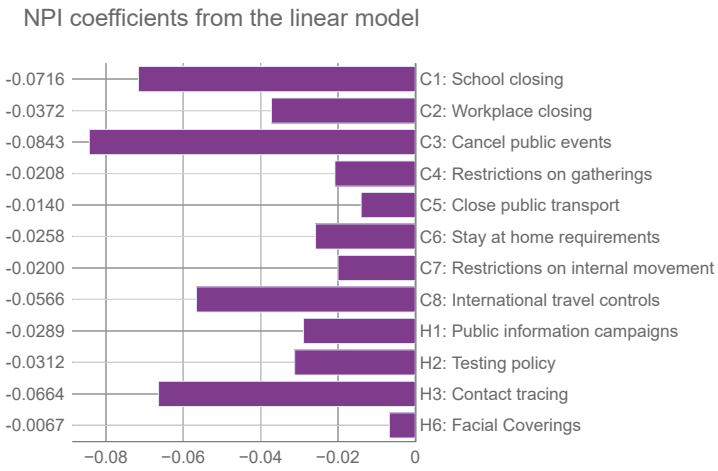

Figure 3. Coefficients from the linear model corresponding to OxNPIs. We use the terminology of Oxford's COVID-19 Government Response Tracker[13], with containment (C) and health (H) categories. Relative values of NPIs can signify their importance for reducing the number of infections – the larger the negative value, the more they suppress the infection spread.

OxNPI	Economic cost	Social cost	Combined
C1: School closing	3.9	11	0.55
C2: Workplace closing	22.0	11	0.96
C3: Cancel public events	1.4	7	0.32
C4: Restrictions on gatherings	1.4	10	0.45
C5: Close public transport	0.1	2	0.09
C6: Stay at home requirements	5.2	12	0.62
C7: Restrictions on internal movement	7.8	10	0.59
C8: International travel controls	4.9	2	0.20
H1: Public information campaigns	0.0026	1	0.04
H2: Testing policy	0.4	1	0.05
H3: Contact tracing	0.1	1	0.04
H6: Facial coverings	0.01	10	0.41

Table 1. Social and economic costs for OxNPIs. Economic costs are shown as % of GDP loss in the period the NPI was implemented. The social costs are based on domain knowledge and expressed on a 1–12 scale. The combined column is the average of the two costs, when both are normalized to the $[0, 1]$ range.

literature, however partial. Justifications for these estimates are given in the Methods section.

The expert-chosen costs are given for the case in which the NPI is implemented with its maximum strictness. For example, for H6, it would mean requiring to wear a mask all the time, including outdoors. For other cases, the costs were linearly scaled down (in rare cases, a custom social cost was defined and used instead of the linearly scaled value). In addition, the "C6: Stay at home requirements" NPI requires the implementation of the C1, C2, C3, C4, C5, and C8 NPIs. Thus, even if it did not have the highest cost, the overall cost implicitly includes the costs of all other listed NPIs.

The results in Table 1 show that "C2: Workplace closing" is the most significant contributor to GDP loss, followed by "C7: Restrictions on internal movement" and "C6: Stay at home requirements." The same NPIs are also listed as the most socially disruptive, together with "C1: School closing" and "H6: Facial coverings". The social cost of "H6: Facial coverings" is one of the previously mentioned exceptions to linear scaling, and it is much lower (2-5) when its strictness is not maximal and masks are not required to be worn outdoors.

One can argue that some of these weights could vary between territories because of differences in their economy, culture, or NPI implementation. In a practical setting, however, these parameters can be freely adjusted by the policy-maker. In addition, the results of our empirical tests revealed that reasonable modifications of the SEC values did not affect the performance (result quality, convergence speed, etc.) of the multi-objective optimization.

Proposing interventions

The proposed intervention plans are composed of OxNPIs which can vary over time, but are restricted to last at least g days in a row, where g is a predetermined parameter we refer to as *granularity*. An NPI, for example, "C2: Workplace closing", can be applied with different levels of strictness (0 – no policy, 1 – closure recommended, 2 – closure for specific sectors, 3 – closure for all-but-essential workplaces).

Figure 4 shows two trade-off intervention plans consisting of NPIs changing in time ($g = 14$), to provide a better intuition for the end goal of this work. They list all 12 NPIs (OxNPIs) we considered in this study, their maximum value, and some sample values. For example, the intervention plan depicted in Figure 4a proposes no restriction for workplaces, while that shown in Figure 4b suggests to close all all-but-essential workplaces from November 24, 2020, to December 20, 2020.

Such intervention plans are then evaluated based on how many COVID-19 infections are likely to occur given their implementation, as well as their SECs. These evaluations are, in turn, used by optimization to find plans that minimize both objectives simultaneously – generating intervention plans with different trade-offs between them, i.e., Pareto front approximations. See section "Intervention plan interpretation" for examples of such Pareto front approximations.

We tried to identify the best value for granularity and we compared five values: 1, 3, 7, 14, and 30. Theoretically, with a finer granularity, we can achieve at least as good intervention plans as with a coarser granularity. However, with finer granularity, aside from being impractical in real-life use, the search space of the optimization problem increases significantly, and the optimization cannot always find the best solutions. Then, we compared the two ways of representing intervention plans during optimization: full vs. condensed. The full representation describes the complete intervention plan (i.e., what NPIs to use in each time slot). The condensed representation, on the other hand, contains only the aggregate strictness for each time slot, which must then be decoded into concrete NPIs with the best benefit vs. strictness ratio before the evaluation (see Methods – Proposing interventions for more details).

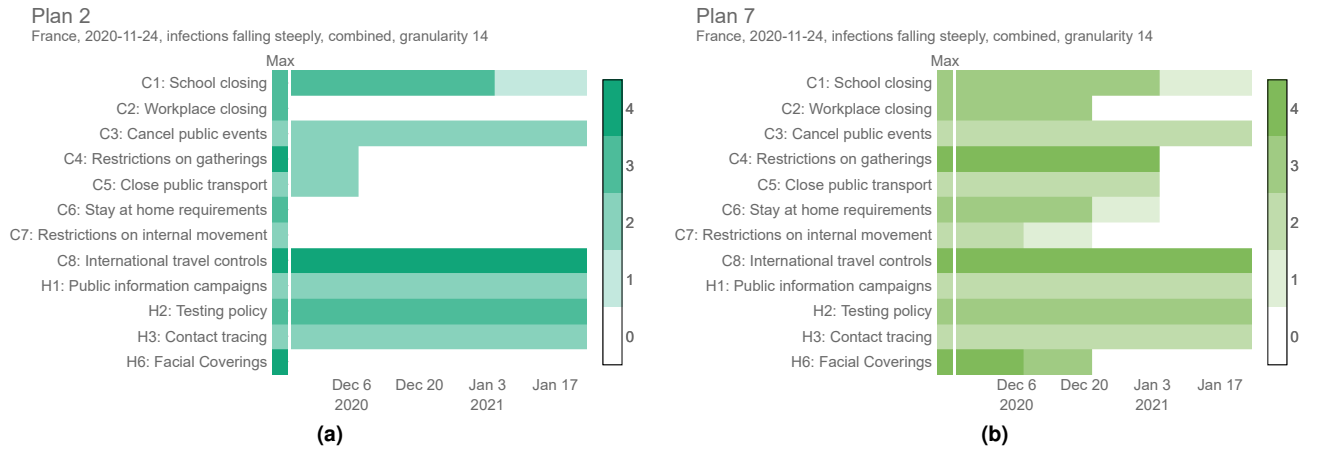

Figure 4. Sample intervention plans for France between November 24, 2020, and January 24, 2021, with a granularity value of 14 days. Refer to Figure 6f to see how these two plans compare against other proposed plans in the same period.

In all cases, the optimization was tested on 50 representative territory/time interval examples (see Methods – Dataset). Due to the stochastic nature of the employed optimization approach, the presented results were obtained after running the optimization 31 times on each example, as this is enough to obtain statistically relevant results. To measure the effectiveness of the multi-objective optimization, we used the well-known hypervolume indicator[17] – the volume of the area bounded by the Pareto front approximation and a user-defined reference point. Note that higher hypervolume values correspond to better results. The medians of the obtained hypervolumes were used for testing the statistical significance of one granularity/representation being better than the other.

We first compared different granularity values when using the condensed representation. According to the Friedman test, we observed statistically significant differences between granularity values: $\chi^2(3) \approx 150.678$ and $p < 0.01$ for social weights, $\chi^2(3) \approx 119.309$ and $p < 0.01$ for GDP weights, and $\chi^2(3) \approx 106.139$ and $p < 0.01$ for combined weights. Post hoc analysis with Wilcoxon signed-rank test and Holm's correction to adjust the p-values indicated that the granularity of 14 days was the most effective among the tested values (see Supplementary Information).

Our results confirm that the optimization algorithm struggles to find near-optimal interventions plans with fine granularity values, due to the increase in search space dimensionality. For example, Figure 5a shows the hypervolume progress – the improvement of the results during the optimization – averaged over 31 optimization runs where the number of intervention plan evaluations was experimentally increased from the default 50k to 300k. This was done to estimate the optimization behavior and convergence when using a large number of evaluations. As we can see, although the results obtained with a granularity value of 7 days eventually surpassed those results obtained with a granularity value of 14 days (at around 230k evaluations), the computational time required to obtain better results using finer granularity values was almost five times longer, and the gain in the solutions' quality was negligible compared to the additional computational resources spent (Figure 5a). In addition, the extremely small differences between the granularity value of 7 or 14 days are practically irrelevant since, in a real-world scenario, the objectives cannot be measured and predicted with such accuracy. Moreover, it is easier to implement intervention plans that change with coarse granularity values[18]; therefore, a granularity value of 14 days seems to be a reasonable choice.

A similar investigation was devoted to finding the best granularity value for the full representation. The results of the statistical analysis revealed significant differences in hypervolume values and showed that the granularity of 30 days is the best performing value for this representation. The complete results can be found in the Supplementary Information.

Finally, we compared the full and condensed representations with the best performing granularity values. According to the Wilcoxon signed-rank test, the condensed representation outperformed the full representation for all types of weights ($p < 0.01$). Moreover, Figure 5b compares the hypervolume progress between the two representations on a typical problem instance, where a much faster convergence can be observed with the condensed representation. This was not unexpected since the applied optimization approach performs significantly faster for low-dimensional search spaces (see Methods – Proposing interventions). The results provided in the following sections were obtained using the condensed representation with a granularity of 14 days since this was the best performing setting.

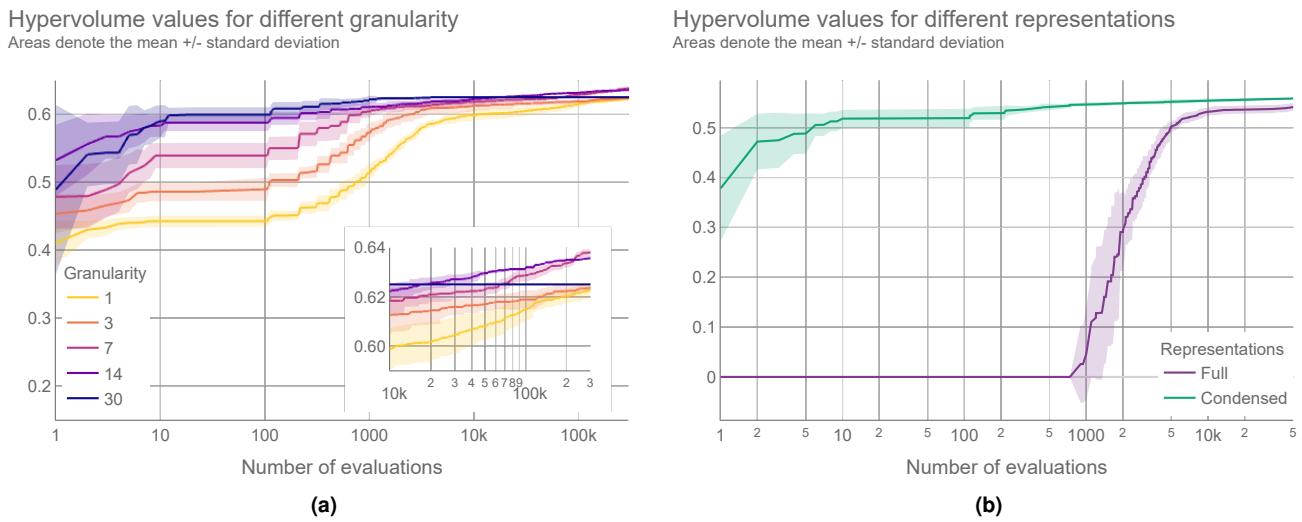

Figure 5. a) Hypervolume progress for different granularity values using condensed representation, and b) hypervolume progress for full and condensed representations with the best performing granularity values. A logarithmic scale is used for the horizontal axis (number of evaluations).

Intervention plan interpretation

To better understand how different intervention plans compare, we generated 10 different intervention plans for the same territory/time interval as that shown in Figure **4** (among all intervention plans obtained by the optimization, we selected the 10 that are the furthest from each other in the objective space). Figure 6 shows for each pan 1) the strictness of the interventions over time, 2) the resulting infection curve, and 3) the comparison of the 10 plans in terms of the number of infections and strictness. This example was done with the granularity of 14 days using the "combined" cost for the interventions. However, we generated plans using all different intervention costs and both 7 and 14 granularities for the same 50 test cases that were used for testing multi-objective optimization. This complete set of results can be found on the results webpage[19]. For a subset of these results, see Supplementary Information.

The proposed plans present a wide range of trade-offs between the two objectives, and policy-makers can choose the one most suited to their needs. In addition, they can change a portion of the plan if deemed necessary and evaluate it again. This whole framework is available as a web tool[20], currently implemented for Slovenia.

The proposed solutions were compared with the real-life solution implemented in the same territory/time. This real-life solution was estimated in two ways, *(real)* using the actual number of infections recorded and *(predicted)* using the predicted number of infections given the implemented NPIs. As the real SEC was, in most cases, unknown, we used the same estimation function for the *real* case as for the proposed plans. In all 50 test cases, the proposed solutions compared favourably against the *predicted* case, and in 47 test cases, the proposed solutions compared favourably against the *real* case. On average, we could find a solution with the same number of infections but with 47.1% lower SEC, or a solution with the same SEC but 68.8% lower number of infections (for details, see Supplementary Information – Comparison of the proposed and implemented solutions).

Trends in the proposed intervention plans

Figure 6 shows the similarities between the proposed intervention plans made in Italy and France. One can reason that – since the NPIs tend to have similar cost/benefit ratio regardless of the current epidemiological picture, and the prescriptor is designed to create solutions with a wide range of costs – the resulting plans will, in most cases, share a common structure that will be somewhat adjusted for different territories/time intervals. Another way of looking at it is to consider that reducing the number of infections when there are, for example, 1000 daily infections has the same importance to the algorithm as reducing the number when there are 3000 daily infections. It is up to the policy-maker to consider when the situation merits selecting a different proposed intervention plan with a lower/higher SEC.

To explore the trends in the structure of the intervention plans, we considered two experiments. First, we averaged the OxNPIs costs across all plans in all test examples, aggregated on a daily basis. The results in Figure 7a show that, on average, the intervention plans are the strictest at the beginning and then gradually become more relaxed. It also shows that in test intervals where the infections were falling, the overall strictness is lower than in cases where infections were raising. The difference might not be as big as expected, again due to the optimizer providing a wide range of intervention plans.

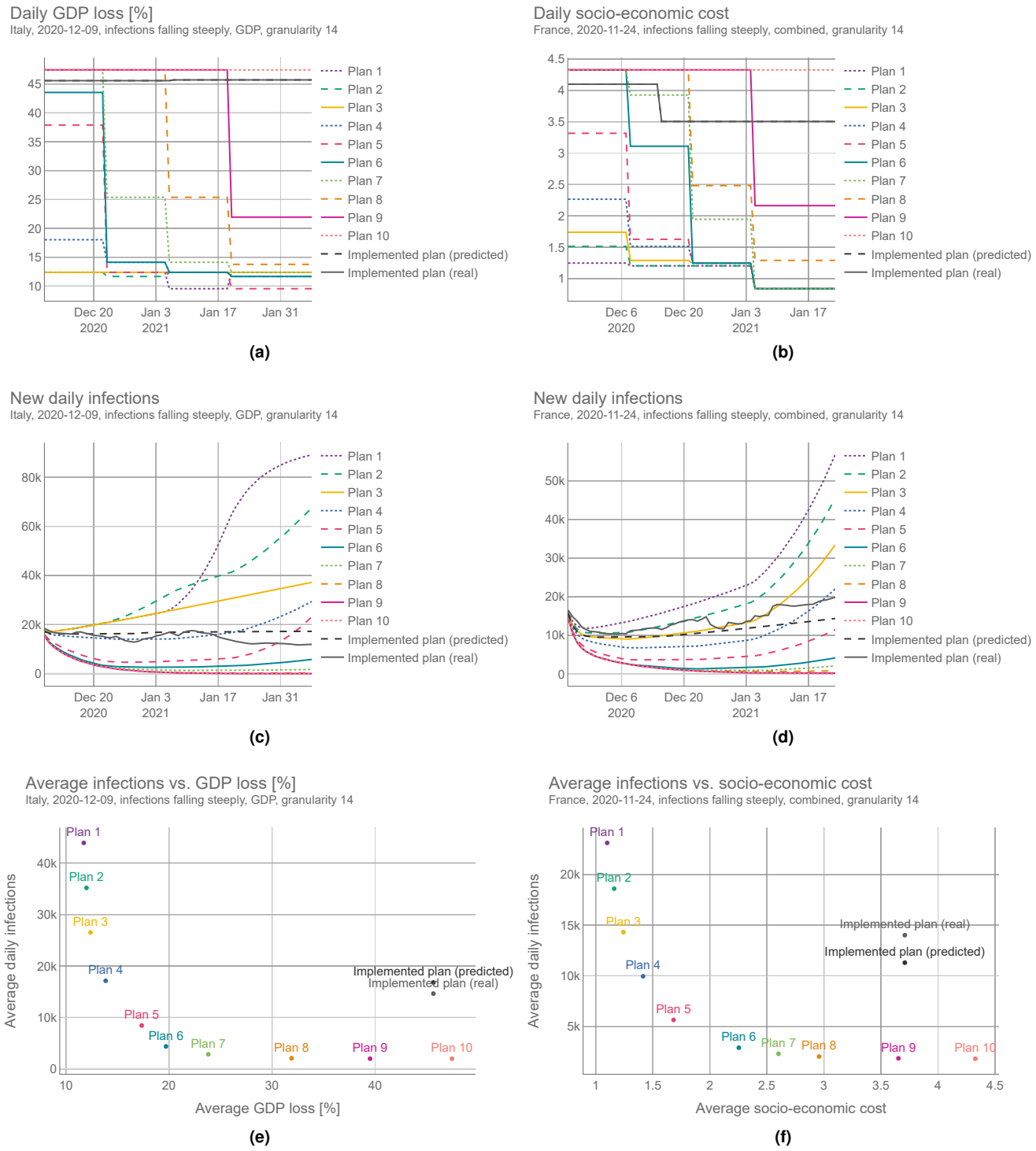

Figure 6. Comparing different intervention plans for Italy and France. a) shows the GDP loss over time, while similarly, b) shows the SEC (GDP loss + social cost) over time. c) and d) show the predicted number of infections, e) and f) show the trade-offs between the two criteria (NPI cost and the number of infections) for different plans.

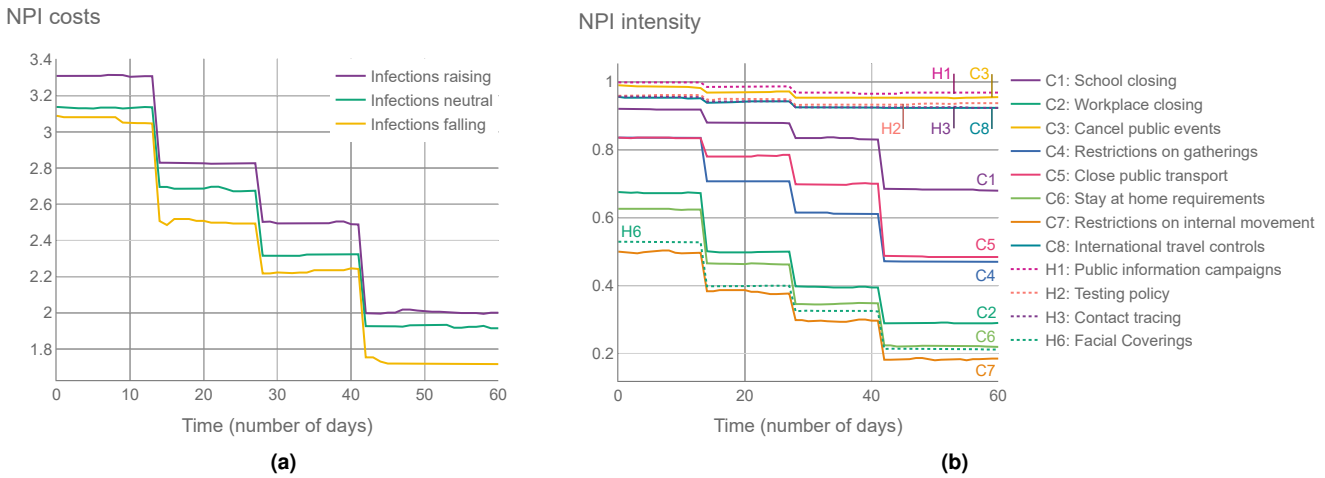

Figure 7. a) Average SEC across all 500 proposed intervention plans (50 test cases, 10 plans on each), given the number of days since the intervention has started. b) Average OxNPI strictness (normalized to 0-1 range) across all 500 proposed intervention plans, given the number of days since the intervention has started.

In the second experiment in Figure 7b, we show the average strictness of individual OxNPIs, again averaged across all intervention plans in all test cases. The NPIs with high average intensities can be considered to provide good trade-offs between their cost and effect.

Discussion

The presented framework can generate efficient intervention plans against COVID-19 and can evaluate their effect and costs. This can greatly help policy-makers to pursue sensible intervention strategies and reason about their strengths and weaknesses. It has been shown that AI-generated intervention plans – at least when evaluated with our methodology – are better than past interventions generated by policy-makers.

Intervention plan insights

In general, the most effective NPIs were school closing, canceling of public events, workplace closing, contact tracing, and international travel controls. This list is not surprising as it is similar to the findings in the literature[1,2]. When accounting for cost (which is usually not done), the most efficient NPIs were information campaigns, canceling of public events, and international travel controls, followed by school closing. The least efficient were the restrictions on internal movement, facial coverings, stay-at-home requirements and workplace closing. The latter two are on the list due to their high cost; in particular, the former can usually be substituted with a combination of other more socially acceptable NPIs. The low placement of "H6: Facial covering" was surprising. Perhaps this is due to masks being somehow inconsistently applied, which may result in bad training data – or alternatively due to "H6: Facial covering" NPI being almost always active, which made it difficult to isolate its effect. Finally, it could be the case that its social cost was overestimated in this study and it should be reduced in potential future analysis.

An additional benefit of the framework, aside from calculating the cost benefit of individual NPIs, is that it can present a timeline of NPI changes that adapts to the current epidemiological situation. In most cases, the approach "start with a strict policy and reduce it over time" seems to be the most effective. We have also shown that adapting the NPI policy every 14 days is enough to get almost ideal cost/benefit as with finer granularities (e.g., adapting every three days provides negligible benefits). Intervention plans made and changed on a monthly basis were found still acceptable; however, using a granularity value of 14 proved to be generally more robust. This could be a valuable finding as frequent changes in NPI policy make adherence difficult and can probably increase socio-economic costs (although we did not model this explicitly).

Technical advantages

The following are the key innovations introduced: 1) combining machine learning and SEIRD models in a way that allows the SEIRD parameters to be adapted to different NPIs and, thus, simulate their effect on infections; 2) using historically fitted parameters to normalize the values output by machine learning in order to adapt predictions for each territory; 3) using

multi-objective optimization for finding the best intervention plans in combination with a "condensed" solution representation – facilitating a highly efficient search.

We argue our predictor to be state-of-the-art. However, it was designed and trained for the whole world, and it is almost certain that for many specific territories, a better predictor could be developed. Similarly, while the proposed OxNPI costs are carefully considered, they can certainly be improved upon, especially for specific territories. To account for this, we have made our whole methodology highly modular, where each part can be substituted by a similar one if necessary – or one can simply adjust the parameter values of the current components.

Limitations

A drawback of the proposed framework is the negligible effect of vaccinations in the models. While we used some vaccination data, the vaccinations were not widespread at the time of data collection. This can be remedied in future work by using more recent data and probably adding another compartment that models vaccinations to the epidemiological model. Nonetheless, due to vaccination rollout in some countries being slow, NPIs may remain an important defense for a while longer.

Second, the infection predictor can sometimes become unreliable when predicting for two or more months in advance. We thus recommend that it should be mostly used for shorter periods (30–45 days in advance) and then the predictions should be updated in real time as new data become available. The predictor also becomes unreliable when the number of infections is growing very quickly. Due to the nature of exponential growth, even a small misprediction of a parameter of the SEIRD model can quickly lead the model astray. The problem is compounded by people spontaneously behaving more cautiously during severe disease breakouts, which affects the infections but is not recorded in NPI data. This effect is difficult to avoid, so it should be taken into consideration when analyzing the proposed plans.

Methods

To find intervention plans with good trade-offs between the number of COVID-19 infections and the SECs, we had to solve the following three problems: 1) how to estimate the number of infections in a specific territory, given an intervention plan; 2) how to estimate the SEC of an intervention plan; and 3) how to use both these estimators to generate good intervention plans. We start by describing the dataset used and then our solution to each of the listed problems in the following subsections.

Dataset

The NPIs used in this study are derived from the "COVID-19 Government response tracker" database, collected by Blavatnik School of Government at Oxford University[13]. This database defines the periods in which different NPIs (e.g., "C1: School closing" and travel restrictions) were implemented in each territory (entities such as countries, US states, or countries of the UK). They also define their "strictness" in the form of numbers usually ranging from 0 to 3 or 4, which can represent, for example, if all or only some schools were closed. From the NPI list available, we selected 12 for analysis in this study: H1, H2, H4, H6, C1-C8 (we denote them as OxNPIs). Their description and the reasoning for their selection can be found in the Supplementary Information – Non-Pharmaceutical Interventions.

The number of infections and deaths was queried from the same database for the period between March 1, 2020, and April 14, 2021. This database contained 235 territories, of which different subsets were used in different stages of our methodology. For fitting the epidemiological model, all 235 territories were used. Then, some territories were excluded as their data could not be accurately fit with an epidemiological model (e.g., if the number of infections were too low or data was missing). This resulted in 194 territories on which we evaluated the predictive model. For each of them, we chose fifty 70-day time intervals, thus generating 9700 test cases for the task.

In addition to the already described OxNPIs and infection numbers, the following attributes were used to train the machine learning models: vaccination[21] (one shot, two shots), strains[22,23] of concern and interest as defined by the World Health Organization[24] testing rate[25], number of hospitalized patients[26], number of patients in intensive care[26], mask use[27], mobility[28,29], weather[30], holidays[27], and 93 static features characterizing countries and regions (e.g., development, culture, and health) from our previous study[31]. "Duration" features were also constructed to capture how long each NPI had been active to date and how much time had elapsed since the first recorded infection case.

Finally, for the prescriptive model evaluation, we chose a representative sample of 50 cases, each consisting of a 60-day interval. This sample was selected by first defining the "category" for each time interval: the categories were created based on the size of the territory (small/large) and the derivative of increase/decrease in number of infections (slope). The slopes were either constant, moderately steep (falling/raising), or very steep (falling/raising). Altogether, we had 10 categories, and we randomly selected five intervals from each. An additional condition for an interval to be selected was to have at least 0.5 average number of daily new cases per 100k of population.

Hybrid machine-learning epidemiological model (HMLE)

Epidemiological model

We used the SEIRD[14] model, which originates from the SIR family of standard epidemiological models used to study the dynamics of infectious diseases. The model consists of a set of differential equations (Equation 1). Letters represent the size of a given compartment (e.g., S for the number of people in the "Susceptible" compartment), β is the infection rate, σ is the incubation period (1/days), γ is recovery rate, and μ is the mortality rate.

$$\begin{aligned} \frac{dS}{dt} &= -\beta\frac{SI}{N} & \frac{dR}{dt} &= \gamma I \\ \frac{dE}{dt} &= \beta\frac{SI}{N} - \sigma E & \frac{dD}{dt} &= \mu I \\ \frac{dI}{dt} &= \sigma E - (\gamma+\mu)I \end{aligned} \tag{1}$$

In a standard SEIRD model, the parameters (β, μ, σ) are constant. In reality – especially in the case of COVID-19 – they are highly dependant on various factors, including NPIs. In related work, there have been several attempts at modeling β as a function of interventions. In the DELPHI model developed by COVID Analytics[32], the effect of interventions is modeled using an arctan function[33]. Zou et al.[34] used machine learning to learn the epidemiological model parameter values from the number of infected and removed (deceased and recovered) cases at time t. In our model, we used machine-learning models that used several different features to achieve this task – allowing us a greater flexibility in dynamically changing the parameters, as opposed to what could be achieved with other methods from related work.

The first step of the process was to fit the parameters to different territory/time intervals. The fitted values were then used as prediction targets for the three machine-learning models (one per parameter). When trained, these models were used to predict the parameters given "alternative" intervals – the same time/territory but with different NPIs. The whole time series of data were first split into intervals based on two criteria: *NPI change* (two or more NPIs change on the same day) and *infection trend* (a 7-day moving-average number of infections that was previously raising, starts falling – or vice versa). Each of the splits was fitted separately for each of the 235 territories using the least squares method. To decide which splitting approach worked better for each territory, a relative error (Equation 2) was computed from the cumulative *true*, and *predicted* infection numbers and parameters that yielded a smaller error were considered as ground truth of the machine learning.

$$error = \left| 1 - \frac{I_{\text{true}}}{I_{\text{pred}}} \right|. \tag{2}$$

Predicting the model parameters with machine learning

The machine learning part of the pipeline consisted of three separate regression models – one for each of the three parameters from the SEIRD model. For the prediction of each parameter, we used the features described in the Dataset section, and some of their subsets.

We performed an initial feature selection on the available dataset by employing Recursive Feature Elimination (RFE) with a 10-fold cross-validation. We evaluated both 1) straightforward feature selection (i.e., running the algorithm on all available features), and 2) including the OxNPIs in the selected features and running the RFE only on the remaining features. However, the results showed no significant improvement after the RFE algorithm. For the sake of model interpretability, we next selected features presenting the strongest negative correlation with the number of infections, and ended up with OxNPIs, duration features, historical infections, COVID-19 strains, and vaccination features.

We tested linear regression[35], ridge regression[35], decision tree[35], LGBM[36], XGB[37], CatBoost[38], Elastic Net[35], Bayesian ridge[35], SVR[35], and Random Forest[35] models. The models were compared with 10-fold cross-validation where the train/test splits were performed territory-wise, meaning that all instances of a territory were in either the test or train set. Keeping all instances of one territory in the same set was important since consecutive instances were typically similar.

In the cases of linear and ridge regression, the regression coefficients for the final model were calculated as the mean values of the coefficients generated in the 10-fold cross-validation. The "H1: Public information campaigns" regression coefficient initially had an excessive value because the corresponding NPI was essentially always present (and was thus used by the model almost as the intercept). We, therefore, manually adjusted it based on Haug et al.'s study[15]. Specifically, we used the four NPIs for which there was a good match between our categorisation and the one presented by Haug et al.: "C1: School closing," "C7: Restrictions on internal movement," "C3: Cancel public events," and "C5: Close public transport." We computed the ratio between the decrease in reproduction rate (β/γ) for these four NPIs[15], and the decrease for "H1: Public information

campaigns". We then multiplied our coefficients for the same NPIs with these ratios, which yielded four possible values for the H1 coefficient. We used the average of these. We then re-ran the regression with fixed relations between the NPI coefficients, so that the relation between them and other coefficients could be readjusted.

Prediction pipeline

The input to the pipeline is an intervention plan, which prescribes OxNPIs for each day. Based on that, a feature vector is created by joining the OxNPI data with the remaining features. Then, for each day, a prediction of all three parameters is made with the three respective machine learning models.

Next, for the time interval leading to (but not including) the prediction interval, the fitted parameters are queried. We assume that the parameters at the beginning of the prediction interval should be the same as the fitted parameters at the end of the last one. Thus, the machine-learning predictions are normalized as $\beta_i = \beta_{\text{last}}/\beta_0$, where β_i is the value of the predicted parameter β on the i-th day, and β_{last} is the last known fitted value of β preceding the prediction interval. Parameters σ and γ are normalized similarly.

If the parameters for any day are such that the reproduction rate exceeds five, then the value of β is reduced until the reproduction rate falls to this threshold value. This is done because such high reproduction rates do not appear in real-life data, but they might be predicted due to some edge case in machine learning. All parameters are smoothed using weighted decay ($\alpha = 0.2$), as we assume that all parameters are changing smoothly.

When the parameters are estimated for each day, they are inserted into the SEIRD model, which can then produce the number of infections for each day. Of note, the starting value of the "Exposed" compartment is set in a way such that the predicted and reported numbers of infections match on day zero.

Socio-economic costs of different NPIs

Socio-economic costs of individual OxNPIs were derived from a set of costs from related work and from the opinion of a domain- expert. Due to the available literature, the costs are likely to contain a bias toward Western countries. Most data is based on reports and gray literature. In the absence of evidence that corresponded exactly to the definition of each NPI, some assumptions had to be made.

While economic costs were available for most OxNPIs, the literature on social costs was far more scarce. We thus placed the ranking of OxNPIs by social costs on a theoretical foundation, but we could not justify the numerical costs as solidly. In addition, according to the literature, these costs may vary across countries (e.g., collectivistic versus individualistic countries); however, we applied standard levels for all WEIRD countries (i.e., for Western, Educated, Industrialized, Rich, and Democratic, a common grouping in psychological studies).

In the study, we used the values listed in Table 1, but the methodology is rather general and a policy-maker can easily adapt it to produce a set of SECs for a specific territory – possibly also implicitly expressing their preferences on what NPIs to avoid (by assigning them higher costs). The combined SEC cost is made simply by normalizing both costs to the $[0, 1]$ range and then averaging both. While this number does not have a good interpretation, it does rank the OxNPIs according to their SECs.

GDP loss

Because the available findings differ in terms of the setting and time, they were normalized to represent the % of GDP loss caused by the NPI while it was in effect. Country-specific GDP values (US $) were used[39]. A complete overview of the cost data used can be found in Table 2, while the resulting economic costs are presented in Table 1. The strictness of the NPI corresponding most closely to each NPI cost estimate is also reported in the same table. While there is some overlap between the NPIs, we have explicitly modelled this only in case of C6: when this NPI is active, so must be C1, C2, C3, C4, C5, and C8.

Social impact

To estimate the social costs, we focused on the perceived strain, dread and loss, perceptions of restricted freedoms, and constraining behaviors (i.e., on the negative impact of each measure on behavior, attitudes, and one's well-being). We ranked interventions from the highest to the lowest based on the absolute levels of perceived dread and loss. Using the rational choice theory, we assumed that the higher the perception of dread, strain, and loss, the more negative is the impact and the higher are the social costs. Understanding human behavior and risk perception is central to effective pandemic management, and thus we apply insights from social and behavioral sciences to inform our assumptions on social impact.

Individuals make decisions by weighing the costs and benefits of NPIs, with emotions driving risk perceptions and making human behavior less predictable[51]. Using this insight, we estimate that stay-at-home requirements (when not sick) impose the highest social costs (cost of 12). Namely, humans are social beings; when introduced over a longer period, lockdown represents the ultimate human challenge. Trumping all other interventions, its harmful effects (on well-being, domestic violence, and society at large) make it the most costly NPI.

Once lockdown is lifted, school and workplace closing, as well as restrictions on internal movement and gatherings, closely follow in cost (11 and 10) because they have similar effects: they limit social relationships, including child's development,

OxNPI	Stringency	Assumption	Country	Source
C1: School closing	3	Closing of all schools	UK	[40]
C2: Workplace closing	3	Mandatory closures	US	[41]
C3: Cancel public events	2	Cancellation of public events	US	[42]
C4: Restrictions on gatherings	2	Same cost as C3	US	[42]
C5: Close public transport	1	Closing or significantly reduced	NL	[43]
C6: Stay at home requirements	3	Confinement/shelter in place	FR	[44]
C7: Restrictions on internal movement	3	Total cessation of domestic tourism	EU	[45,46]
C8: International travel controls	3	Travel & tourism restrictions	All	[47]
H1: Public information campaigns	2	Communications related to COVID-19	UK	[48]
H2: Testing policy	2	Policy of 30 million tests per week	US	[49]
H3: Contact tracing	2	Targeted to contacts of identified cases	US	[49]
H6: Facial coverings	2	EU import of face masks	EU	[50]

Table 2. Sources for GDP loss for each OxNPI. Assumptions made, the country source analyses, and the NPI stringency in the source country are also listed.

and personal and family income (i.e., household purchasing power). There is a time trade-off with these interventions, with longer and more frequent closures and travel restrictions increasing the costs. Moreover, longer workplace closures and travel restrictions are likely to result in the breakdown of general compliance. Hence, for these interventions to be effective, timing and length are key.

By contrast, public information campaigns as key to compliance have the highest social benefit. Elaboration of risks and severity of possible implications motivate cooperative behavior, which is central to managing the pandemic. Thus, the smarter the campaign, the higher the compliance and the benefit. Similar underlying logic applies to the test and trace policy; targeting high-risk individuals, it is an essential component of pandemic management. The cost of these NPIs is thus the lowest (1).

We arranged the costs of the remaining NPIs between these extremes. Cancelling of public events is a milder version of restrictions on gatherings (7). Closing of public transport was typically in force together with other NPIs, which reduced the need for public transport, so it was not a major inconvenience; and international travel controls also did not impact a large fraction of the population (both 2).

Proposing interventions

The task of proposing intervention plans can be mathematically formulated as a multi-objective optimization problem with two objectives that need to be minimized: the total number of infections (f_1) and the SECs of the proposed plan (f_2). The two objectives are conflicting since an efficient way to slow down the spread of infections requires a stringent intervention plan with expensive NPIs. The first objective is expressed as the total number of infections predicted from the HMLE model, while the second objective is the cost of NPIs averaged over the plan's duration. The problem is constrained by limiting the number of new daily infections to 1500 per 1M residents. This is done as the cases with more infections are not considered useful to policy-makers and almost never appear in real-life data.

Formally, an intervention plan – a solution to the proposed optimization problem – is represented by a $12 \times n$ integer-valued matrix, P, where its 12 rows correspond to the 12 OxNPIs and n is the number of time slots determined by the given granularity value and the whole period (e.g., Figure 4 contains $n = 4$ time slots resulting from a granularity value of 14 days and an interval length of 60 days). In detail, P_{ij} indicates the strictness of the i-th NPI in the j-th time slot. In particular, we tested five values for granularity: 1, 3, 7, 14, and 30.

Based on the multi-objective formulation of the proposed optimization problem, the experimental evaluation aimed at finding sets of trade-off intervention plans representing approximations for Pareto fronts. For this purpose, we used the Nondominated Sorting Genetic Algorithm II (NSGA-II)[52] equipped with a Constrained Dominance Principle (CDP)[52] to handle the constraint. NSGA-II belongs to the group of evolutionary algorithms, and as such, it imitates the biological evolution to search the space of possible intervention plans and find plans with good trade-offs between the two objectives.

The optimization problem was solved using two NSGA-II internal solution representations: the full representation defined by the matrix P and the condensed representation defined by a vector of length n where the j-th variable corresponds to the

overall SEC at the j-th time slot. The second representation was considered due to the significant reduction in the search space dimensionality (from $12n$ to n), allowing for much faster convergence than the high-dimensional search space for the full representation. While the full representation can be used without modifications, the condensed representation needs to be decoded to the intervention plan before evaluation. This is achieved by replacing the total SECs with OxNPI values. The OxNPI combination to replace each SEC is selected as the one with the lowest projected infections out of those within the allowed SECs. This mapping is computed in advance, by having all OxNPIs combinations sorted based on their effectiveness (by using linear model's coefficients for each NPI), and when the exact SECs of different NPIs are known, the most effective combinations that do not exceed the cost threshold is chosen from that list.

The experimental setup was defined equally for both representations and was established based on some initial experiments. NSGA-II was run with a population of 100 solutions for 500 generations (50k plan evaluations in total). This number of evaluations proved to be sufficient for convergence using coarser granularity values. Moreover, increasing function evaluations did not significantly improve the results, even for fineer granularity values. For this reason, 50k evaluations represented a good trade-off between the framework's effectiveness and efficiency. The one-point crossover was used as the crossover operator and the random resetting as the mutation operator. Additionally, the crossover probability was set to 0.9 and the mutation probability to $1/D$, where D equals $12n$ for the full representation and n for the condensed representation.

Data availability

Data used in this research was taken from public repositories[13,21–23,25,26,28–30]. All data used to generate the figures is available in our repository[19]. The same repository also contains all final results.

Code availability

All code used in the production of the results is available in our code repository[53].

References

1. Flaxman, S. *et al.* Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* (2020).
2. Moore, S., Hill, E. M., Tildesley, M. J., Dyson, L. & Keeling, M. J. Vaccination and non-pharmaceutical interventions for COVID-19: a mathematical modelling study. *The Lancet Infect. Dis.* **21**, 793–802 (2021).
3. Osterrieder, A. *et al.* Economic and social impacts of COVID-19 and public health measures: results from an anonymous online survey in Thailand, Malaysia, the UK, Italy and Slovenia. *BMJ Open* **11** (2021). https://bmjopen.bmj.com/content/11/7/e046863.full.pdf.
4. Berger, L. *et al.* Rational policymaking during a pandemic. *Proc. Natl. Acad. Sci.* **118** (2021). https://www.pnas.org/content/118/4/e2012704118.full.pdf.
5. Lazzerini, M. & Putoto, G. COVID-19 in Italy: momentous decisions and many uncertainties. *The Lancet Glob. Heal.* **8**, e641–e642 (2020).
6. World Health Organization. Considerations for implementing and adjusting public health and social measures in the context of COVID-19: interim guidance, 14 june 2021. Technical documents (2021).
7. Yousefpour, A., Hadi, J. & Stelios, B. Optimal policies for control of the novel coronavirus disease (COVID-19) outbreak. *Chaos Solitons Fractals* (2020).
8. Chen, V. C. P. *et al.* An optimization framework to study the balance between expected fatalities due to COVID-19 and the reopening of U.S. communities. *medRxiv* (2020). https://www.medrxiv.org/content/early/2020/07/20/2020.07.16.20152033.full.pdf.
9. Yaesoubi, R. *et al.* Adaptive policies to balance health benefits and economic costs of physical distancing interventions during the COVID-19 pandemic. *Med. Decis. Mak.* **41**, 386–392 (2021).
10. XPRIZE.org. Pandemic response challenge (2021).
11. Miikkulainen, R. *et al.* From prediction to prescription: Evolutionary optimization of non-pharmaceutical interventions in the COVID-19 pandemic. *IEEE Transactions on Evol. Comput.* **25**, 386–401.
12. Lozano, M. A. *et al.* Open data science to fight COVID-19: Winning the 500k XPRIZE Pandemic Response Challenge. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 384–399 (Springer International Publishing).

13. Hale, T., Webster, S., Petherick, A., Phillips, T. & Kira, B. Oxford COVID-19 government response tracker (OXCGRT) (2020).

14. Martcheva, M. *An Introduction to Mathematical Epidemiology*. Texts in Applied Mathematics (Springer US, 2015).

15. Haug, N. *et al.* Ranking the effectiveness of worldwide COVID-19 government interventions. *Nat. Hum. Behav.* **4**, 1303–1312 (2020).

16. Predictor model results as of Monday 22 February 2021. https://phase1.xprize.evolution.ml/.

17. Zitzler, E. & Thiele, L. Multiobjective evolutionary algorithms: A comparative case study and the strength Pareto approach. *IEEE Transactions on Evol. Comput.* **3**, 257–271 (1999).

18. Shen, Y. *et al.* Monitoring non-pharmaceutical public health interventions during the COVID-19 pandemic. *Sci. Data* **8** (2021).

19. Results repository. https://github.com/jsi-dis/ai-covid-interventions.

20. COVID-19 intervention plans - a web application for the Ministry of Health. http://xprize-e9.ijs.si:5555/.

21. Mathieu, E. *et al.* A global database of COVID-19 vaccinations. *Nat. Hum. Behav.* **5**, 947–953 (2021).

22. Mullen, J. L. *et al.* outbreak.info. https://outbreak.info/ (2021). Accessed: 2021-04-14.

23. Yuelong, S. & John, M. Gisaid: Global initiative on sharing all influenza data – from vision to reality. *Euro Surveillance* **22** (2017).

24. World Health Organization. Tracking SARS-CoV-2 variants. Last accessed: 2021-07-27.

25. Hasell, J. *et al.* A cross-country database of COVID-19 testing. *Sci. Data* **7**, 345 (2020).

26. Ritchie, H. *et al.* Coronavirus pandemic (COVID-19). *Our World Data* (2020). Https://ourworldindata.org/coronavirus.

27. Fan, J. *et al.* The University of Maryland Social Data Science Center Global COVID-19 Trends and Impact Survey, in partnership with Facebook. https://covidmap.umd.edu/api.html (2021). Accessed: 2021-04-14.

28. Google LLC. Google COVID-19 community mobility reports. https://www.google.com/covid19/mobility/ (2020). Accessed: 2021-04-14.

29. Apple. COVID-19 mobility trends reports. https://covid19.apple.com/mobility (2020). Accessed: 2021-04-14.

30. Visual Crossing Corporation. Visual crossing weather. https://www.visualcrossing.com/weather-data (2017-2019). Accessed: 2021-04-14.

31. Janko, V. *et al.* Machine learning for analyzing non-countermeasure factors affecting early spread of COVID-19. *Int. J. Environ. Res. Public Heal.* **18** (2021).

32. Thornburg, H. Introduction to Bayesian statistics (2001).

33. Li, M. L. *et al.* Forecasting COVID-19 and analyzing the effect of government interventions. *MedRxiv* 2020–06 (2021).

34. Zou, D. *et al.* Epidemic model guided machine learning for COVID-19 forecasts in the United States (2020).

35. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

36. Ke, G. *et al.* Lightgbm: A highly efficient gradient boosting decision tree. *Adv. neural information processing systems* **30**, 3146–3154 (2017).

37. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 785–794, 10.1145/2939672.2939785 (ACM, New York, NY, USA, 2016).

38. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. Catboost: Unbiased boosting with categorical features. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, 6639–6649 (Curran Associates Inc., Red Hook, NY, USA, 2018).

39. GDP for the US. https://data.worldbank.org/indicator/NY.GDP.MKTP.CD. Last accessed: 2021-10-20.

40. Sadique, M. Z., Adams, E. J. & Edmunds, W. J. Estimating the costs of school closure for mitigating an influenza pandemic. *BMC Public Heal.* **8**, 1–7 (2008).

41. Business closures and partial reopenings due to COVID-19 could cost the U.S. trillions. https://news.usc.edu/178979/business-closures-covid-19-pandemic-united-states-gdp-losses/. Last accessed: 2021-10-20.

42. COVID-19 triggered sports and festival cancellations will have staggering economic impact across the country. http://performanceresearch.com/covid-19-triggered-sports-and-festival-cancellations-will-have-staggering-economic-impact-across-the-countr Last accessed: 2021-10-20.

43. UITP projects €40bn hit for European public transport in 2020. https://www.railjournal.com/financial/uitp-projects-e40bn-hit-for-european-public-transport-in-2020/. Last accessed: 2021-08-16.

44. Economic flash France - INSEE estimates the impact of a month of confinement at -3 pts of annual GDP. https://www.tresor.economie.gouv.fr/Articles/2020/03/30/flash-conjoncture-france-l-insee-estime-l-impact-d-un-mois-de-confinement-a-3-pts-de-pib-annuel. Last accessed: 2021-08-16.

45. European Union: Gross domestic product (GDP) from 2016 to 2026. https://www.statista.com/statistics/527869/european-union-gross-domestic-product-forecast/. Last accessed: 2021-10-20.

46. Domestic tourism expenditure in Europe from 2012 to 2020. https://www.statista.com/statistics/617517/domestic-tourism-expenditure-europe/. Last accessed: 2021-10-20.

47. WTTC research reveals global travel & tourism sector suffered a loss of almost US$4.5 trillion in 2020 due to the impact of COVID-19. https://wttc.org/News-Article/Global-TandT-sector-suffered-a-loss-of-almost-US4-trillion-in-2020. Last accessed: 2021-10-20.

48. Costs of coronavirus advertising. https://www.whatdotheyknow.com/request/costs_of_coronavirus_advertising. Last accessed: 2021-10-20.

49. Cutler, D. M. & Summers, L. H. The COVID-19 Pandemic and the $ 16 Trillion Virus. *JAMA* **324**, 1495–1496 (2020). https://jamanetwork.com/journals/jama/articlepdf/2771764/jama_cutler_2020_vp_200215_1602876140.96328.pdf.

50. Which country imported the most face masks? https://ec.europa.eu/eurostat/web/products-eurostat-news/-/ddn-20201006-1. Last accessed: 2021-10-20.

51. Bavel, J., Baicker, K. & Boggio, P. e. a. Using social and behavioural science to support COVID-19 pandemic response. *Nat. Hum. Behav.* **4**, 460 – 471 (2020).

52. Deb, K., Pratap, A., Agarwal, S. & Meyarivan, T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evol. Comput.* **6**, 182–197 (2002).

53. Code repository. https://repo.ijs.si/covid/covid-xprize/.

54. Van Rooij, B. *et al.* Compliance with COVID-19 mitigation measures in the United States. *Amsterdam Law Sch. Res. Pap.* (2020).

55. Moya, C., Cruz y Celis Peniche, P., Kline, M. A. & Smaldino, P. E. Dynamics of behavior change in the COVID world. *Am. J. Hum. Biol.* **32**, e23485 (2020). https://onlinelibrary.wiley.com/doi/pdf/10.1002/ajhb.23485.

56. Arnot, M. *et al.* How evolutionary behavioural sciences can help us understand behaviour in a pandemic. *Evol. Medicine, Public Heal.* **2020**, 264–278 (2020). https://academic.oup.com/emph/article-pdf/2020/1/264/34824238/eoaa038.pdf.

57. Perrotta, D. *et al.* Behaviours and attitudes in response to the COVID-19 pandemic: insights from a cross-national Facebook survey. *EPJ Data Sci.* **17**.

58. Fischer, I. *et al.* The behavioural challenge of the COVID-19 pandemic: indirect measurements and personalized attitude changing treatments (IMPACT). *Royal Soc. Open Sci.* **7**, 201131 (2020). https://royalsocietypublishing.org/doi/pdf/10.1098/rsos.201131.

Acknowledgments

We acknowledge the financial support from the Slovenian Research Agency (research core funding No. P2-0209).

Author contributions statement

V.J., M.L., and M.G., conceived the experiments. N.R. and D.S. collected and prepared the dataset. N.R., A.V., D.S., and C.M., conducted the experiments. S.V., D.D.S., and J.J. provided analysis of NPI costs, T.T. provided visualizations. V.J., A.G., M.L., analyzed the results. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryinformation.pdf