A. Significance

Colorectal cancer, the third most common cancer diagnosed in both men and women in the United States excluding skin cancers, is projected to cause about 50,630 deaths during 2018 [1]. Effective treatment will require a personalized medicine approach to track disease progression and determine the major drivers for each patient to tailor an appropriate therapy. The combinations of leucovorin, fluorouracil, and irinotecan/oxaliplatin are commonly used for treating colon cancers, e.g., FOLFIRI (leucovorin, fluorouracil, and irinotecan) and FOLFOX (leucovorin, fluorouracil, and oxaliplatin). However, these combinations remain inactive in about half of the patients, and in addition, resistance to treatment appears in almost all patients who were initially responders [2]. A major clinical challenge is to obtain an effective treatment strategy for each patient or at least identify a subset of patients who could benefit from a particular treatment. Although there is a debate about the direct relation between gene expression profiles and FOLIRI response [3, 4, 5, 6], there are several studies in support of using gene expression profiles of primary colon cancer tissue to build a predictor classifier for response to the treatments [2, 7, 8, 9]. In this pilot study, we will focus on FOLFIRI and model the interaction between leucovorin, fluorouracil, irinotecan, and various cell types, including immune cells and epithelial cells.

One of the main computational tools used to discover, test, and predict dose exposure response is quantitative systems pharmacology (QSP) modeling [10, 11, 12]. These models help us to understand biological processes, identify targets and biomarkers, predict treatments' outcomes, and discover mechanisms of drug resistance [13]. They are the main tools to test the impact of drug parameters and biological variances on drug efficacy and safety [14]. Since biological process are very complex, the QSP models are often complex with many nonlinearities and parameters [15]. One of the main challenges of the QSP modeling is parameter estimation. Traditionally, it has been thought that the precision of QSP's parameter estimation is not very important [16]. Therefore, parameters are commonly estimated using the data that are often assembled from disparate sources rather than a single curated dataset. These may span multiple biological studies, in vitro, in vivo, and clinical assays, and may be qualitative, e.g., increase vs. decrease [16]. As a result, they cannot be easily validated or used for personalized treatments.

Since each tumor has its own unique characteristics, they respond differently to different treatments. Therefore, we need to develop QSP models that take into account each individual's tumor characteristics. Hence, we propose to develop a data-driven QSP software for obtaining personalized colon cancer treatment. This software can incorporate the impact of biological variance on efficacy and safety and lead to scientific decisions on choosing appropriate therapy for each patient. Furthermore, it can also help us to discover new treatment strategies. The input of the software will be the gene expression profiles of primary tumors and the output will be the optimal personalized treatment strategies, including dosages and time intervals and the efficacy of these treatments to cure colon cancer. It also has the potential to suggest new ways to design an effective drug for some patients.

The rapid advances in technology show the importance of publicly available projects and online collaborations. Therefore, we will create a user friendly environment in GitHub so that people from all over the globe could easily contribute on the project and improve the model. We will provide an environment for computational people as well as experimentalists to have a discussion and write their comments to improve the project. We will also create a user-friendly platform for experimentalists to add their data to the project. We will push the entire project to the GitHub, which has been widely used to host open-source software projects. People who want to contribute to the project can fork the repository and make the changes, and then release the revised project as a new repository. Furthermore, the contributors can add their changes to the main project by creating a pull request. We will verify their work and then accept changes into the main project if we see no issues. Additionally, we have created a project titled "Data-driven QSP software for personalized colon cancer treatment" in the National Cancer Informatics Hub (NCIP Hub, https://nciphub.org). One of the features of NCIP Hub is the ability to connect Google Drive, Dropbox, or GitHub to a Project, and we will use this feature to connect our GitHub repository with our NCIP Hub project. We will publish all files related to this project on NCIP Hub, such as seminar presentations, training material, data, and tools. NCIP Hub will also assist us to leverage community expertise and access to data and tools across the cancer research community. Additionally, to share our findings and also leverage the community experiences, at least one of the investigators of this project will participate in the ITCR program activities, including the annual meetings and working groups. PI Shahriyari plans to participate in all ITCR program activities. If she is unable not attend any of these activities, then PI Roy and/or PI Pal will participate.

B. Innovation

One of the key components in the study of QSP models is the estimation of its parameters. Existing parameter estimation methods for QSP models are done using experimental data. Very often, assembled data from various sources are utilized rather than a single curated dataset. These data sets usually span results of various biological experiments, including

in vitro and in vivo studies. For example, one might use an experimental data related to the lung study of an animal to estimate the parameters of a model for colon cancer. **The first innovative aspect of this project is that it uses patient data to estimate the values of the QSP model parameters.** Saliently, the parameters will be estimated for each patient, separately, so as to come up with a personalized treatment strategy.

A strong correlation between in situ immune reactions in tumor regions and prognosis has been observed regardless of the local extent of the tumor and of invasion of regional lymph nodes [17]. A weak in situ immune reaction in tumor regions is associated with a poor prognosis even in patients with minimal tumor invasion (stage I). Moreover, high expression of the Th17 markers predict a poor prognosis for patients with colorectal cancer, whereas patients with high expression of the Th1 markers have prolonged disease-free survival [18]. These observations show the importance of personalized treatments that consider the immune cells, their numbers and interactions. **The second innovative feature of this project is modeling an individual's immune cell variations and their interaction networks in the presence of drugs.**

Importantly, the proposed framework advances our understanding of therapeutic and toxic drug activities in individuals with diverse gene expression profiles. Our framework provides tools to discover and test new treatment strategies. Although we design this project for FOLFIRI, it is able to suggest other effective treatment strategies. **The third innovative feature of our proposed project is scalability; it can be easily updated and also modified to be used for other cancer types.**

The final innovative feature of this project is the combination of a powerful mathematical framework to estimate the model parameters and robust statistical methods to draw inference on the sensitivity of these estimated parameters. To the best of our knowledge, this novel combination has never been used before in the context of QSP models.

C. Approach

Our entire approach to develop the proposed software package is represented as an algorithm in Figure 1. The detail of the proposed approach is described in the sections below.

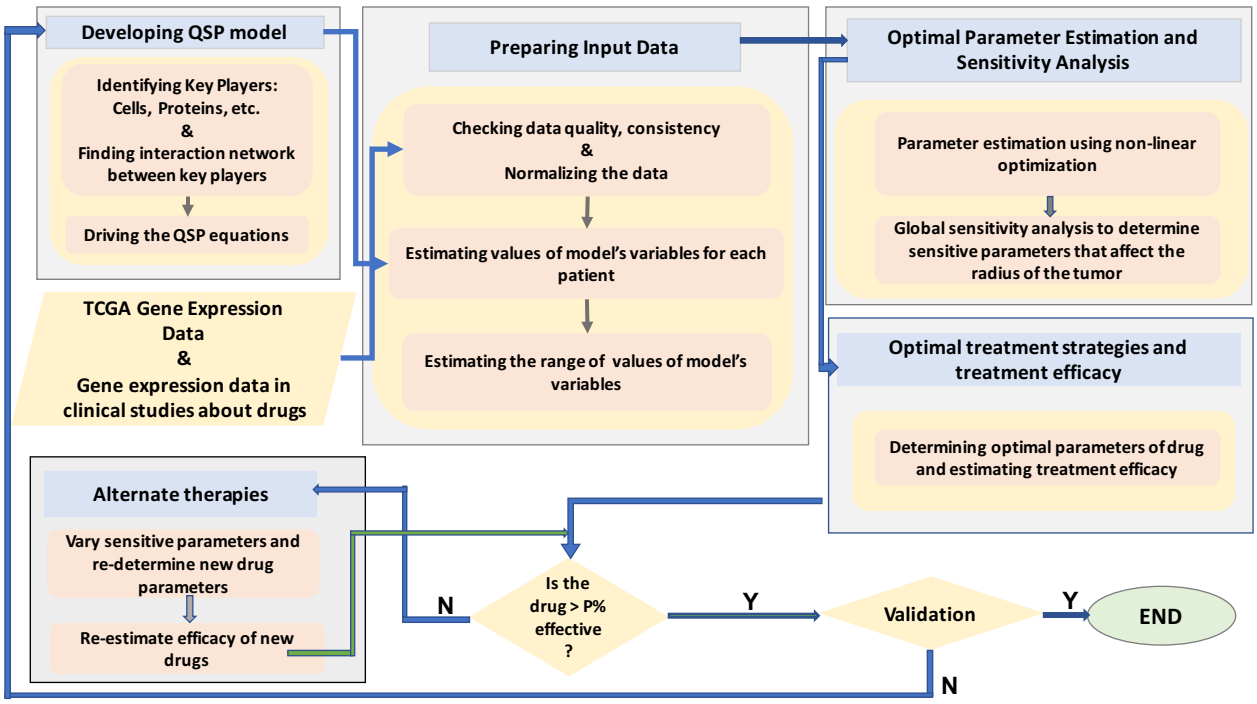

Figure 1: Flowchart. A schematic view of the algorithm for our proposed software package.

C.1. Developing the QSP model

C.1.1. Model building: To develop the data-driven QSP software, the first objective is to build the underlying QSP model. To do so, we first review published studies to obtain the key components and factors, including the interaction network between various types of immune cells and epithelial cells. We drive the kinetic reaction formulas for the interaction network based on the mass action and Michaelis-Menten laws. For biochemical processes $A + B \to C$ with

A and B unlimited, we will use the mass action law $\frac{dC}{dt} = \lambda AB$, where λ is the production rate of C. If a molecule of B ligands to a cytoplasmic-receptor on the A-cell resulting in an activated A-cell, which we designate as C-cell, instead of the mass action law, we will use the Michaelis-Menten law, $\frac{dC}{dt} = \lambda B \frac{A}{A+K_A}$, where K_A (*half-saturation* of A) and λ are constants. The same formula will be used if molecules B enhance the proliferation of the A-cells. We also consider a different limiting situation whereby molecules B are absorbed and internalized by A-cells. The capacity of A to "eat" the B molecules is limited, so the rate of loss of B through absorption by A is modeled by $-\lambda A \frac{B}{K_B+B}$. A generic QSP model can be formulated through a system of ordinary differential equations (ODE). For instance, IL-6 is produced by various cell types, however macrophages and T-cells are predominant producers of IL-6 during inflammation [19]. The dynamics of IL-6 can be modeled by

$$\frac{dI_6}{dt} = \underbrace{\lambda_{I_6T_4}T4}_{\text{Production by CD4}^+\text{ T-cells}} + \underbrace{\lambda_{I_6M}M}_{\text{Production by M}\Phi} - \underbrace{\hat{\lambda}_{CI_6}C\frac{I_6}{K_{I_6}+I_6}}_{\text{Prolif. of cancer cells by IL-6}} - \underbrace{\delta_{I_6}I_6}_{\text{Degradation}} . \tag{1}$$

The first two terms are respectively the rates of the production of IL-6 by CD4$^+$ T-cells cells and macrophages, and the last term is the degradation rate of IL-6. Cancer cells are epithelial cells with abnormally high growth and small death rates. The release of IL-6 by CD4$^+$ effector T cells and macrophages induces the proliferation of epithelial cells [20] by activating STAT3 in intestinal epithelial cells [21]. The dynamics of cancer cells will be modeled as follows:

$$\frac{dC}{dt} = \underbrace{\lambda_C C(1-\frac{C}{C_0})}_{\text{Proliferation}} - \underbrace{\delta_{CT_8}T_8C}_{\text{Death by CD8}^+\text{ T-cells}} + \underbrace{\lambda_{CI_6}C\frac{I_6}{K_{I_6}+I_6}}_{\text{Prolif. of cancer cells by IL-6}} - \underbrace{\delta_C C}_{\text{Death rate}} . \tag{2}$$

Here, the first term is the proliferation rate of cancer cells, the second term is the death rate of cancer cells by CD8$^+$ T-cells, and the third term is proliferation rate of cancer cells through IL-6. The last term is the death rate of cancer cells. In this way, we will obtain a generic system of ODEs (1-2).

Next, we denote the radius of the tumor as r. To provide effective treatments for curing cancer is equivalent to controlling the growth of the radius of the tumor. Thus, it becomes crucial to estimate the radius of the tumor. In order to do so, we assume that the tumor occupies a region $\Omega(t)$ and that the combined density of cells at each point in $\Omega(t)$ is approximately equal to $1\ g/cm^3$. Since cancer growth is abnormal, the above assumption implies that a velocity field $\mathbf{v}$, by which the cells are moving within $\Omega(t)$, is developed. Hence, for all the species of cells, represented by the vector $\boldsymbol{A}$ depending on the vector of parameter $\boldsymbol{\theta}$, the system of equations $\frac{d\boldsymbol{A}(\boldsymbol{\theta})}{dt} = F(\boldsymbol{A}, \boldsymbol{\theta})$ from the ODE model (1-2), is replaced by the system of partial differential equations (PDE) $\frac{\partial \boldsymbol{A}(\boldsymbol{\theta})}{\partial t} + div(\mathbf{v}\boldsymbol{A}(\boldsymbol{\theta})) = D_{\boldsymbol{A}}\Delta\boldsymbol{A} + F(\boldsymbol{A}, \boldsymbol{\theta})$, where $D_{\boldsymbol{A}}$ is the vector dispersion (or diffusion) coefficient of species set $\boldsymbol{A}$ and Δ is the 3-dimensional Laplace operator. The cytokines and molecules such as $IL-6$ are also diffusing with their own diffusion coefficients. However, since the diffusion coefficients are much larger than cells' diffusion coefficients, we may neglect the effect of $\mathbf{v}$, and thus replace each equation $\frac{d\boldsymbol{A}}{dt} = F$ by

$$\frac{\partial \boldsymbol{A}(\boldsymbol{\theta})}{\partial t} = D_{\boldsymbol{A}}\Delta\boldsymbol{A} + F(\boldsymbol{A}, \boldsymbol{\theta}). \tag{3}$$

The PDE in (3) is the governing system of equations in the QSP model. The radius of the tumor r depends on the vector of parameter $\boldsymbol{\theta}$, $x \in \Omega$ and $t \in [0, T]$. To control r, we first need to obtain the optimal feasible set of $\boldsymbol{\theta}$ for each patient. **The second objective of this project is to estimate $\boldsymbol{\theta}$ that appear as coefficients of the PDE system (3), using the gene expression data of each patient at a fixed time and determine the subset of this parameter vector that is sensitive to the radius of the tumor for each patient.** This is done by solving an inverse problem and performing a global sensitivity analysis. In the next sections, we discuss the type of data we plan to use in our model, solve the inverse problem to obtain the optimal feasible parameter vector and perform a sensitivity analysis of this optimal parameter vector with respect to the radius of the tumor $r = r(\boldsymbol{\theta}, x, t)$.

C.1.2. Potential pitfall and alternate solutions: In the design of QSP models, finding the right granularity, i.e. the level of detail and complexity of the model, is extremely difficult [22]. To overcome this challenge, we collaborate with Dr. Regev group at the Broad Institute (see the support letter) in developing the QSP model, and we focus on the interaction between cells and some molecules in tumors. We determine the main players and components from those that have been identified as key factors in a few independent studies. Then, we add the other key players to the model one by one, and we stop adding as soon as we get satisfactory results (see validation part in C.3.3).

C.2. Optimal parameter estimation and sensitivity analysis for each patient

C.2.1. Preparing Input Data: We use TCGA gene expression data of patients with colon cancer to estimate the values of key players, i.e. variables of the model, including the concentration of proteins and cell densities, at a fixed time, for each patient. We normalize the data to handle the possible noises in data collection. Based on expression of immune cell markers, we estimate the density of each immune cell type in the tumor. For example, the genes associated with Th17 (RORC, IL17A), Th2 (IL4, IL5, IL13), Th1 (Tbet, IRF1, IL12Rb2, STAT4), and cytotoxicity (GNLY, GZMB, PRF1) have been identified by hierarchical clustering of a correlation matrix [18]. We aim at applying all analytical tools developed to provide an estimation of the abundances of member cell types in a mixed cell population, using gene expression data, including CIBERSORT [23], DeconRNASeq [24], and the method developed by Senbabaoglu et al. [25] using single-sample GSEA (ssGSEA) score [26]. We will also apply the method proposed in [25] using singscore [27] instead of ssGSEA score. To make sure these methods work best for our QSP model, we validate and test them in collaboration with Dr. Regev's lab, which studies transcriptional circuits that control gene expression and cell phenotype within cells and in complex tissues and tumor ecosystems (see the support letter). Then, we either choose the method that works best for our model or alternatively, we could add these methods as an option for users to choose.
The estimated values of the variables of the model provide us with the initial conditions of the QSP model for each patient. Moreover, the minimum and maximum values of variables across all patients give us the acceptable ranges for values of the variables of the model. Using these ranges, we employ an inversion algorithm to estimate the values of the parameters of the model.

C.2.2. Optimal parameter estimation: The governing mathematical model for QSP is the PDE system (3). We call this system of PDEs as the state equation or the forward equation, consisting of the state variables which depend on the vector of m unknown parameters, $\boldsymbol{\theta} \in \mathbb{R}^m$, of the QSP model, which are known as the coefficients of the state equation. We want to determine the unknown parameters in the QSP model using the given range of values of variables by solving an inverse problem. This is usually known as coefficient estimation in the context of PDEs, where one tries to recover the coefficients of the PDE system using the given data as a function of the state variables [28, 29, 30]. Such methods have been used for several applications in the past that includes estimation of fluid flows [31, 32, 33, 34], medical imaging [35, 36, 37], antibiotic production [38], cellular automaton models [39], neuron modeling [40], wine fermentation processes [41], differential games [42, 43], but has not been explored in the context of QSP models. The solution of the state equation $\boldsymbol{A}$ is a function of $\boldsymbol{\theta}$. We assume m parameters in the model for each patient and, thus, $\boldsymbol{\theta} \in \mathbb{R}^m$. The given data function in the QSP model can be written in a mathematical form as $\mathcal{D}(\boldsymbol{A}, \boldsymbol{\theta})$ at a time $t^* \in (0, T)$ in Ω. We also impose feasibility constraints on the components of $\boldsymbol{\theta}$ as $0 < a_i \leq \boldsymbol{\theta}_i \leq b_i,\ a_i, b_i,\ 1 \leq i \leq m$. We denote the (possibly noisy) value of variables of the patients with colon cancer at t^* as g^δ and the corresponding dependence of the data on the state variables as $\mathcal{D}(\boldsymbol{A}, \boldsymbol{\theta})(x, t^*)$. Then, the corresponding inverse problem is:
Solve for $\boldsymbol{\theta} \in \mathbb{R}^m$ such that

$$\mathcal{D}(\boldsymbol{A}, \boldsymbol{\theta})(x, t^*) = g^\delta, \quad \forall x \in \Omega,\ t^* \in (0, T), \tag{4}$$

subject to the state equation (3). The stated inverse problem is highly ill-posed and challenging to solve using direct inversion procedures. We, thus, consider a PDE-constrained optimization problem to find $\boldsymbol{\theta}$ that minimizes an objective functional J as follows ([44, 45, 46]):

$$\min_{\boldsymbol{\theta}} J := \int_\Omega (\mathcal{D}(\boldsymbol{A}, \boldsymbol{\theta})(x, t^*) - g^\delta)^2 \, dx + F(\boldsymbol{\theta}), \tag{5}$$

subject to the state equation (3) and the feasibility constraints. In this setup, the first term in the functional J is similar to a least-squares data-fitting term where we want to find the optimal parameters $\boldsymbol{\theta}$ to get $\mathcal{D}$ as close to the observed noisy data g^δ as possible. The second term F is a regularization term that encompasses various apriori information about the variable $\boldsymbol{\theta}$. For e.g., the L^1 regularization term $F(\boldsymbol{\theta}) = |\boldsymbol{\theta}|$ promotes sparsity of $\boldsymbol{\theta}$ and using the L^2 regularization term $F(\boldsymbol{\theta}) = |\boldsymbol{\theta}|^2$ one looks for $\boldsymbol{\theta}$ with the least energy.
Solving the optimization problem (5) is a challenging task as usual gradient-based algorithms use the knowledge of convexity and smoothness of the functional J. However, our functional J is non-convex due to the presence of the PDE constraints (3). Moreover, the presence of the L^1 regularization term renders the functional to be non-differentiable in the classical sense. Thus, we use a recently developed semi-smooth algorithm, known as the variable inertial proximal (VIP) method, to solve the optimization problem (5). This method is described in detail in [47, 48]. The solution of the optimization problem gives us the optimal set of parameters which approximately solve (4).

C.2.3. Sensitivity analysis and uncertainty quantification of optimal parameter vector $\boldsymbol{\theta}$: The question of

accuracy of results from mathematical models of biological systems is an important one and such accuracies are often complicated due to the presence of uncertainties in experimental data that are used to estimate the parameter values [49]. Thus, it is important to understand the effects of model parameter values on the outcome measures or outputs of interest. Uncertainty in the chosen parameter values result in variability in the model's prediction of resulting dynamics and the significance of the variability introduced depends on the degree of uncertainty [50]. **We will perform a sensitivity analysis of the model parameters with respect to the radius of the tumor, based on which effective treatment strategies will be devised.**

Instead of carrying out a single-parameter or local sensitivity analysis, where all other parameters are kept fixed at baseline values, thereby resulting in inaccuracies in assessing uncertainties [51], we propose to study a multi-dimensional parameter space globally which facilitates simultaneous identification of all uncertainties. For this purpose, we will make use of two efficient and powerful statistical tools: Latin hypercube sampling (LHS) scheme and partial rank correlation coefficient (PRCC) analysis. We will follow the procedure discussed in [52] to carry out the LHS. The main idea is to start with m uncertain parameters corresponding to a mathematical model of interest and then use the LHS scheme to come up with N values for each of the m uncertain parameters. This will then allow us to create a $(N \times m)$ matrix. Here, N is the simulation size and is chosen such that $N > (4/3)m$. We now describe how to carry out the PRCC analysis to identify the sensitive parameters.

Partial Rank Correlation Coefficient (PRCC) Analysis: From the LHS scheme, all m values in each row of the matrix will be used as input values for the mathematical model. Since there are N rows, we have N different sets of inputs for our model that produce N different output values, noting that the output measure is a function of m parameters and is unidimensional. To find the parameters that contribute the most uncertainty to model prediction, we will use the non-parametric PRCC as detailed below.

Step 1: Rank all the m columns of our matrix and call the resulting matrix as $X_R = [X_{1R}, X_{2R}, \cdots, X_{mR}]$, where each X_{iR}, $i = 1, \cdots, m$, is a vector of dimension $(N \times 1)$ representing the rank transform for the i-th parameter. In a similar way, also rank the output values, which is a vector of dimension $(N \times 1)$, and call the resulting ranked vector of output values as Y_R.

Step 2: For the i-th parameter ($i = 1, 2, \cdots, m$), run two multiple linear regression (MLR) models. The first one is the MLR of X_{iR} on all $\{X_{jR} : j = 1, 2, \cdots, m \text{ and } j \neq i\}$ and the second one is the MLR of Y_R on all $\{X_{jR} : j = 1, 2, \cdots, m \text{ and } j \neq i\}$.

Step 3: Calculate the residuals from both MLR models. The PRCC value for the i-th parameter is the Pearson's correlation coefficient between these two sets of residuals. In this way, we calculate the PRCC values for all k model parameters.

From Step 3 above, we perform tests of significance to assess if a PRCC value is significantly different from zero. This will be done using the student's t statistic described in [53] and then calculating the corresponding p-value. The parameters with large PRCC values (> 0.5 or < -0.5) and corresponding small p-values (< 0.05 or < 0.01) will be identified as the ones that contribute to uncertainty and hence model's prediction imprecision. A negative sign of a PRCC value would imply that the corresponding parameter is inversely proportional to the outcome measure. Finally, the order in which the identified parameters contribute to uncertainty will be judged based on the magnitude of their PRCC values. We, thus, obtain the sensitive parameters, which govern the radius of the tumor r for each patient. **Using this information, the third objective of the project is to obtain the optimal dosages and design effective drugs for treatment of colon cancer for each individual patient, which is described in the next section.**

C.2.4. Potential pitfall and alternate solutions: Although PRCC is a robust sensitivity measure, it actually measures monotonic relationship between the output measure and the model parameters. So, to justify the use of PRCC, we first examine the monotonicity plots for each parameter and the output measure. If for a certain parameter the monotonicity fails, we use the method proposed by [52], where we split the parameter range into several intervals such that monotonicity holds true for each interval.

C.3. Optimal treatment strategies

C.3.1. Finding the optimal treatment strategy: From the global sensitivity analysis, for each patient, we obtain the sensitive parameter set. Let us denote the j sensitive parameter vector as $\boldsymbol{\theta}_s$ and the rest $m-j$ parameter vector as $\boldsymbol{\theta}_{ns}$. We now induce a drug with $\eta(t)$ that models the interaction between the leucovorin, fluorouracil, and irinotecan drugs and key players. Then, the underlying QSP model (3) for each patient would also be dependent on η in the following

way:

$$\frac{\partial \boldsymbol{A}(\boldsymbol{\theta})}{\partial t} = D_{\boldsymbol{A}} \Delta \boldsymbol{A} + F(\boldsymbol{A}, \boldsymbol{\theta}) + \eta(t). \tag{6}$$

To determine the optimal dosages for the drug administered, we need to find η such that the radius of the tumor r decreases to a steady state value r^* at the final time T. We note here that the optimal η would depend on the sensitive parameter vector $\boldsymbol{\theta}_s$ in a coupled way and r depends on η. Thus, we are led to solving the following optimal control problem to obtain $\eta(t)$

$$\min_{(\boldsymbol{\theta}_s, \eta)} \tilde{J} := \int_{\Omega} (r(\boldsymbol{\theta}_{ns}, \boldsymbol{\theta}_s, \eta(T), x, T) - r_T^*)^2 \, dx + \int_0^T \tilde{F}(\eta(t), \boldsymbol{\theta}_s) \, dt, \tag{7}$$

subject to the QSP constraint (6), where $\eta(t)$ can be thought of as an open-loop control for r. We again use the VIP algorithm to solve this optimal control problem and obtain the optimal $\eta(t)$ along with the values of the sensitive parameters $\boldsymbol{\theta}_s$ at the optimum. **This $\eta(t)$ is the suggested therapy for curing the patient, assuming that the other parameters in the patient's cell dynamics are fixed at their optimum feasible values.** But since the radius of the tumor r is highly sensitive to the parameter vector $\boldsymbol{\theta}_s$, we need to predict the efficacy of the recommended therapy and suggest new robust therapies for cure, if required. This is done through a probability analysis in the next part.

C.3.1.1. Potential pitfall and alternate solutions: Though the implementation of the optimization algorithm to solve for the parameters are robust and yield accurate results, there are some cases where the optimal solution might not be obtained. This is because the optimization algorithm is iterative and requires a proper initial guess to initiate the algorithm. In case of such failure, we use different initial guesses to determine the feasible one.

C.3.2. Efficacy of the optimal treatment strategy and suggested new therapies: We predict the efficacy of the suggested optimal treatment strategy from the previous part and also recommend alternate therapies for curing colon cancer in each patient. We perform this through the following steps:

(a) For each patient, keeping η fixed, we find out the range of values of $|r - r^*|$ at the final time T with respect to the feasible range of values of $\boldsymbol{\theta}_s$ to determine the confidence level $100(1-\alpha)\%$ of the radius of the tumor r such that $|r - r^*| < tol$. Here tol is a desired tolerance level chosen for defining cure and α is the error probability. We then say that the drug with parameter η is $100(1-\alpha)\%$ effective for curing the patient. For example, if α is 0.1, then, we say that the drug is 90% effective.

(b) If $\alpha > \alpha^*$, for a maximum acceptable probability of failure α^*, we suggest new therapies by choosing appropriate η. For this, we consider the range of feasible values of $\boldsymbol{\theta}_s$ as in Step (a) and solve a modified optimal control problem to obtain the new range of values of η.

(c) For each such η obtained in Step (b), we determine the confidence level $100(1-\alpha)\%$. We classify those η whose confidence level is greater than $100(1-\alpha^*)\%$ and say that these alternate η will cure the patient with probability bigger than $100(1-\alpha^*)\%$.

(d) If there exists no η such that $100(1-\alpha) > 100(1-\alpha^*)$, then we say that the chances of the patient being cured with any drug is less than $100(1-\alpha^*)\%$.

C.3.3. Validation: Del Rio et al. [2] performed a study to identify a pattern of gene expression able to predict response to FOLFIRI in colorectal cancer patients with synchronous and unresectable liver metastases. From this study, we get the gene expression data of primary tumor of 21 patients and the tumor response, the size of the metastatic lesions from bidimensional measurements (the product of the longest diameter and the longest perpendicular diameter) using computed tomography scanning, to FOLFIRI. We get 10 more cases from another study by Matinez-Garci et al. [5], which challenges the fact that gene expression is not a direct predictor of FOLIRI response.

We will perform 3-fold cross validation; each time we randomly select 21 patients from the combination of the above-mentioned studies to train the model. For each individual patient, we first estimate the values of our model's variables based on their gene expression data. Then, this will give us the initial conditions of the inverse problem. Utilizing the range of acceptable values for the variables of the model (obtained using TCGA data), we estimate model's parameter for each patient using inverse problem techniques. Then, we perform sensitivity analysis and estimate the probability of treatment efficacy. If the estimated probability of more than four (20%) patients do not match with the clinical observation, we first change the methods of estimating model's variables from gene expression data, and then improve QSP model. If we still find the false discovery rate (FDR) > 20%, we will modify the inverse problem and sensitivity analysis methods accordingly. We repeat this process until the FDR $\leq$ 20%.