INFORMATION OPTIMAL EXPERIMENT DESIGN OF HIV 2-LTR CLINICAL TRIALS BY EXPECTED KULLBACK-LEIBLER DIVERGENCE

by

LaMont C. Cannon II

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Biomedical Engineering

Winter 2018

© 2018 LaMont C. Cannon II
All Rights Reserved
INFORMATION OPTIMAL EXPERIMENT DESIGN OF HIV 2-LTR
CLINICAL TRIALS BY EXPECTED KULLBACK-LEIBLER
DIVERGENCE

by

LaMont C. Cannon II

Approved: _____________________________________________________________
Dawn Elliott, Ph.D.
Chair of the Department of Biomedical Engineering

Approved: _____________________________________________________________
Babatunde Ogunnaike, Ph.D.
Dean of the College of Engineering

Approved: _____________________________________________________________
Ann L. Ardis, Ph.D.
Senior Vice Provost for Graduate and Professional Education
I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: ________________________________________
Ryan Zurakowski, Ph.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: ________________________________________
Abhyudai Singh, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: ________________________________________
Nii Attoh-Okine, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: ________________________________________
Fabrizio Sergi, Ph.D.
Member of dissertation committee
ACKNOWLEDGEMENTS

First and foremost I would like to thank my wonderful wife Joy. I am grateful to have had such a loving and supportive partner throughout my entire journey through graduate school. I would also like to thank my friends and family for their unconditional support. In particular, I would like to thank my late grandmother, Lorraine Graves for always believing in me and encouraging me to continue my education.

Words are powerless to express my gratitude to my graduate advisor Dr. Ryan Zurakowski. I am not only grateful for exceptional guidance academically, I am also thankful for friendship that has developed between our families over the years. I would also like to thank Dr. Michael Piovoso for taking the time to provide his excellent insight and tutelage.

I would also like to express my sincerest gratitude to my colleges from Dr. Zurakowski’s lab. I would like to especially thank Aditya Jagarapu and Dr. Cesar Garcia-Vargas for their friendship and support throughout the years.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>ix</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>x</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>xxi</td>
</tr>
<tr>
<td>Chapter</td>
<td></td>
</tr>
<tr>
<td>1 INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Human Immunodeficiency Virus</td>
<td>1</td>
</tr>
<tr>
<td>1.2 HIV Treatment</td>
<td>3</td>
</tr>
<tr>
<td>1.3 Integrase Inhibition and 2-LTR Formation</td>
<td>5</td>
</tr>
<tr>
<td>1.4 Preliminary Data and 2-LTR Dynamics</td>
<td>6</td>
</tr>
<tr>
<td>1.5 Goals and Organization</td>
<td>8</td>
</tr>
<tr>
<td>1.6 Impact and Significance</td>
<td>10</td>
</tr>
<tr>
<td>2 HIV 2-LTR MODEL AND DATA ANALYSIS</td>
<td>11</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>11</td>
</tr>
<tr>
<td>2.2 2-LTR Model</td>
<td>11</td>
</tr>
<tr>
<td>2.3 Measurement Error</td>
<td>14</td>
</tr>
<tr>
<td>2.3.1 Real-Time Polymerase Chain Reaction Error</td>
<td>15</td>
</tr>
<tr>
<td>2.3.2 Droplet Digital Polymerase Chain Reaction Error</td>
<td>15</td>
</tr>
<tr>
<td>2.4 Data Analysis</td>
<td>16</td>
</tr>
<tr>
<td>2.5 Posterior Parameter Distributions</td>
<td>16</td>
</tr>
<tr>
<td>3 EXPERIMENT DESIGN FACTORS</td>
<td>18</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>18</td>
</tr>
<tr>
<td>3.2 Simple Problem</td>
<td>18</td>
</tr>
<tr>
<td>3.3 Sample Schedule Optimization</td>
<td>20</td>
</tr>
<tr>
<td>3.4 Volume Optimization</td>
<td>20</td>
</tr>
</tbody>
</table>
4 EXPERIMENT DESIGN METHODS ........................................... 24

4.1 Introduction ......................................................... 24
4.2 Traditional Methods ............................................... 24
  4.2.1 Fisher Information Matrix ..................................... 25
  4.2.2 D-optimality Criterion ......................................... 25
  4.2.3 T-optimality Criterion ......................................... 26
  4.2.4 A-optimality Criterion ......................................... 26
  4.2.5 E-optimality Criterion ......................................... 26

4.3 Expected Kullback-Leibler Divergence ........................... 27
  4.3.1 Simulated Patient Pool ......................................... 27
  4.3.2 Calculation ...................................................... 28
    4.3.2.1 Markov Chain Monte Carlo (MCMC) Methodology .... 28
    4.3.2.2 KLD Calculation .......................................... 30
  4.3.3 Parameter Transformations .................................... 31

4.4 Optimization ....................................................... 32
  4.4.1 Genetic Algorithm ............................................. 32
  4.4.2 Constrained Optimization .................................... 33

5 METHOD VALIDATION .................................................. 34

5.1 Introduction ....................................................... 34
5.2 Hatano et al. Experiment ......................................... 34
5.3 1000 Monte Carlo vs Unscented Transform EKLD ............... 35
5.4 Predicted Information Gain in Hatano Experiment .............. 40

6 SAMPLE SCHEDULE OPTIMIZATION ................................. 43

6.1 Introduction ....................................................... 43
6.2 Four Sample Optimization ....................................... 43
  6.2.1 Real-Time Polymerase Chain Reaction Design ............ 43
    6.2.1.1 D-optimality Criterion ................................ 44
6.2.1.2 T-optimality Criterion .......................... 44
6.2.1.3 A-optimality Criterion .......................... 47
6.2.1.4 E-optimality Criterion .......................... 47
6.2.1.5 EKLD Optimal ................................. 49
6.2.1.6 Result Comparison ............................. 49

6.2.2 Droplet Digital Polymerase Chain Reaction Design .... 53
6.2.2.1 D-optimality Criterion ......................... 54
6.2.2.2 T-optimality Criterion ......................... 54
6.2.2.3 A-optimality Criterion ......................... 57
6.2.2.4 E-optimality Criterion ......................... 57
6.2.2.5 EKLD Optimal ................................ 57
6.2.2.6 Result Comparison ............................. 61

6.2.3 Schedule Robustness ............................. 61

6.3 Six Sample Optimization ........................... 66
6.3.1 Real-Time Polymerase Chain Reaction Design ........ 66
6.3.1.1 D-optimality Criterion ......................... 66
6.3.1.2 T-optimality Criterion ......................... 68
6.3.1.3 A-optimality Criterion ......................... 68
6.3.1.4 E-optimality Criterion ......................... 71
6.3.1.5 EKLD Optimal ................................ 71
6.3.1.6 Result Comparison ............................. 74

6.3.2 Droplet Digital Polymerase Chain Reaction Design .... 76
6.3.2.1 D-optimality Criterion ......................... 76
6.3.2.2 T-optimality Criterion ......................... 78
6.3.2.3 A-optimality Criterion ......................... 78
6.3.2.4 E-optimality Criterion ......................... 81
6.3.2.5 EKLD Optimal ................................ 81
6.3.2.6 Result Comparison ............................. 81

6.3.3 Schedule Robustness ............................. 84

6.4 Sample Days vs. Information Gain .......................... 87
7 SAMPLE VOLUME OPTIMIZATION ........................................ 91

7.1 Introduction .......................................................... 91
7.2 Volume Optimal Four Sample Schedule .......................... 91

7.2.1 Real-Time Polymerase Chain Reaction Design ............ 91
7.2.2 Droplet Digital Polymerase Chain Reaction Design ...... 92

7.3 Volume Optimal Six Sample Schedule ............................ 93

7.3.1 Real-Time Polymerase Chain Reaction Design ............ 94
7.3.2 Droplet Digital Polymerase Chain Reaction Design ...... 94

8 CONCLUSIONS .......................................................... 96

8.1 EKLD Method and Experimental Design Criteria ............ 97
8.2 Experiment Design Implications .................................... 100
8.3 Future Work .......................................................... 101

BIBLIOGRAPHY ............................................................. 102
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td><strong>Model Parameter Definitions</strong></td>
<td>Definitions and units for all parameters from the 2-LTR model.</td>
<td>13</td>
</tr>
<tr>
<td>6.1</td>
<td><strong>Four Point qPCR Optimized Sample Schedules</strong></td>
<td>Optimal four point sample schedules optimized for a qPCR measurement assay for each optimality criterion schedule, the EKLD method schedule and the experimental schedule as well as their associated EKLD.</td>
<td>54</td>
</tr>
<tr>
<td>6.2</td>
<td><strong>Four Point ddPCR Optimized Sample Schedules</strong></td>
<td>Optimal four point sample schedules optimized for a ddPCR measurement assay for each optimality criterion schedule, the EKLD method schedule and the experimental schedule as well as their associated EKLD.</td>
<td>62</td>
</tr>
<tr>
<td>6.3</td>
<td><strong>Six Point qPCR Optimized Sample Schedules</strong></td>
<td>Optimal six point sample schedules optimized for a qPCR measurement assay for each optimality criterion schedule, the EKLD method schedule and the experimental schedule as well as their associated EKLD.</td>
<td>76</td>
</tr>
<tr>
<td>6.4</td>
<td><strong>Six Point ddPCR Optimized Sample Schedules</strong></td>
<td>Optimal four point sample schedules optimized for a ddPCR measurement assay for each optimality criterion schedule, the EKLD method schedule and the experimental schedule as well as their associated EKLD.</td>
<td>86</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

1.1 HIV Replication Cycle Replication begins when a free virus attaches itself to a host cell. Its genetic material is then injected into the host in the form of viral RNA. The cellular enzyme reverse transcriptase then creates viral complementary DNA which is then transported into the nucleus. Once in the nucleus, the viral DNA is then integrated into the Host genome, mediated by the host enzyme integrase. The host cell then transcribes new viral RNA from integrated the DNA. Some of this DNA is then translated into structural proteins to create new virions. Once the new virus is constructed and viral RNA is packaged a new virion buds from the host cell and matures, ready to infect a new host. 2

1.2 Latent HIV Cell When an infected cell enters into a latent state, the cell becomes transcriptionally silent. The virus has integrated its DNA into the host genome waiting for the cell to become active again. Once the cell becomes active, it will begin to transcribe new viral RNA. Some of this RNA will get translated into proteins to create new virons. These virions, once constructed will then bud off the surface of the host cell and mature. 4

1.3 HIV 2-LTR Production Cycle If during the process of integrating the viral DNA into the host genome, integration is blocked, host nuclear enzymes will fuse the 2 Long Terminal Repeat ends of the viral DNA. This creates circularized DNA known as “2-LTR circles”, measured and used as a marker of replication. 6

1.4 Experimental Data From Buzon et al. Data from treatment group of the integration inhibitor study done by Buzon et al. Six blood samples were taken and 2-LTR concentration was quantified at day 0, day 14, day 28, day 84, day 168, day 336 or week 0, week 2, week 4, week 12, week 24, week 48. The transient behavior of the 2-LTR dynamics is evident from the data. 7
2.1 Concatenation of Priors for Patients in Buzon et al. study. Based on the data from the experiment done by Buzon et al., Luo et al., performed a Bayesian Markov Chain Monte Carlo analysis to estimate 2-LTR mode posterior parameter distributions for the 13 patients in the study that exhibited 2-LTR dynamics. These parameter distributions were then concatenated and used as the prior knowledge about the parameters for the experiment optimization.

3.1 Experiment Design (a) The graph in the figure shows the trajectories of the exponential growth model based on different values of the model parameter $\alpha$. The figure shows a sample taken at time seven ($t = 7$) with an associated confidence interval denoted by the solid black circle marker and the black error bars. The trajectories that pass through the confidence region are colored red and the trajectories that do not pass through the confidence region are colored blue. (b) The histograms show the prior and posterior distributions of the model exponential growth parameter $\alpha$. Prior distribution is colored blue and corresponds to the range of blue trajectories in the above plot. The posterior distribution is colored red and corresponds to the trajectories which pass through the region of confidence of the measurement in the above plot.

3.2 Experiment Design (Changing Sample Day) (a) The graph shows the sample moved from time point seven ($t = 7$) to time point five ($t = 5$). More of the trajectories now pass through the region of confidence for the measurement. (b) The histograms show the prior and posterior distributions based on the measurement at time point five ($t = 5$). Because the measurement has been taken sooner and more trajectories pass through the region of confidence, the posterior distribution is broader when compared to the posterior distribution from the measurement at time point seven ($t = 7$). This leads to less information gained from taking the measurement sooner.
3.3 **Experiment Design (Changing Sample Volume)**. (a) The graph shows the sample again at time point seven \((t = 7)\); however, a large volume is taken during the sample. This creates a more accurate measurement marked by the reduced size of the error bars. Now, fewer trajectories pass through the region of confidence for the measurement. (b) The histograms show the prior and posterior distributions based on the measurement at time point seven \((t = 7)\) but with the increased accuracy due to the larger sample volume. Since fewer of the trajectories pass through the region of confidence of the measurement the posterior distribution of the model exponential growth parameter \(\alpha\) is significantly more narrow. This leads to a larger gain in information from taking this sample when compared to the other samples.

4.1 **Sigma Point Patient Dynamics.** Trajectories of the unscented transform patients plotted over the eight week experiment window. Note: trajectories for only 10 out of the 11 unscented patients is shown. The patient with greatest increase in concentration was removed to so that the remaining trajectories can be displayed.

4.2 **Transformation of Prior and Posterior Distributions to Normal Distributions and Back.** Uniform prior distribution results in a Beta posterior distribution. Both are transformed to normal distributions an KLD is calculated. KLD is conserved through all transformations. Distributions are transformed to normal distributions to reduce computational burden.

4.3 **Formulation of the Genes and Chromosomes in the Genetic Algorithm.** Genetic algorithm shown with three genes per chromosome and 19 basepairs per gene. Each basepair corresponds to a sample day with 1 being a day when sample is taken and 0 being a day when no sample is taken. Arrows denote different types of crossovers that can occur.

5.1 **Experimental Data From Hatano et al.** Data from the follow up integrase inhibitor intensification experiment performed by Hatano et al. Four samples were taken at day 0, day 7, day 14, and day 56 (week 0, week 1, week 2, and week 56). The 2-LTR concentrations were quantified with a ddPCR assay.
5.2 Measurement Error ddPCR vs qPCR in Quantifying 2-LTR dynamics. A 2-LTR trajectory is plotted with 95 percent confidence regions shaded for ddPCR assay (red) and qPCR assay (green). The ddPCR assay is much more accurate at all concentrations which leads to more information from the experiment.

5.3 1000 Monte Carlo Patients vs Unscented Transform Patients with qPCR Assay. KLD for 1000 Monte Carlo Patients Compared to EKLD From UT Patients Assuming qPCR Measurement Noise. The 1000MC patients are displayed as the dark blue bars and the UT patients are displayed as the yellow bars. KLDs are compared for four different schedules: four sample optimal schedule, Hatano et al. experiment schedule, six sample optimal schedule, and Buzon et al. experiment schedule.

5.4 1000 Monte Carlo Patients vs Unscented Transform Patients with ddPCR Assay. KLD for 1000 Monte Carlo Patients Compared to EKLD From UT Patients Assuming ddPCR Measurement Noise. The 1000MC patients are displayed as the dark blue bars and the UT patients are displayed as the yellow bars. KLDs are compared for four different schedules: four sample optimal schedule, Hatano et al. experiment schedule, six sample optimal schedule, and Buzon et al. experiment schedule.

5.5 UT vs 1000 MC Correlation Plot (qPCR). Plot comparing the KLD for 100 different sample schedules comparing when using 1000 randomly Monte Carlo sampled patients vs. when using the 11 unscented transform patients assuming qPCR measurement noise. Correlation coefficient \( \rho = 0.927 \).

5.6 UT vs 1000 MC Correlation Plot (ddPCR). Plot comparing the KLD for 100 different sample schedules comparing when using 1000 randomly Monte Carlo sampled patients vs. when using the 11 unscented transform patients assuming ddPCR measurement noise. Correlation coefficient \( \rho = 0.950 \).

5.7 EKLD comparison of Hatano et al. Experimental Patient Data KLD calculated from the actual Hatano et al. experimental data (green). Results are compared to EKLD of the Hatano et al. schedule (blue), and the optimal four point schedule (red).
5.8 EKLD comparison of Hatano et al. Experimental Patient Data compared by individual patient KLD KLD calculated from the actual Hatano et al. experimental data and displayed for each of the 15 patients in the treatment group (green). Results are compared to EKLD of the Hatano et al. schedule (blue), and the optimal four point schedule (red).

6.1 Four point optimal D-optimality criterion with qPCR measurement noise (a) four optimal points selected by the D-optimality criterion plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the D-optimality criterion.

6.2 Four point optimal T-optimality criterion with qPCR measurement noise (a) four optimal points selected by the T-optimality criterion plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the T-optimality criterion.

6.3 Four point optimal A-optimality criterion with qPCR measurement noise (a) four optimal points selected by the A-optimality criterion plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the A-optimality criterion.

6.4 Four point optimal E-optimality criterion with qPCR measurement noise (a) four optimal points selected by the E-optimality criterion plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the E-optimality criterion.
6.5 **Four point optimal EKLD schedule with qPCR measurement noise** (a) four optimal points selected by the EKLD method plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the EKLD method.

6.6 **Sampling Schedule used by Hatano et al.** Sample schedule used in the experiment performed by Hatano et al. plotted on a 2-LTR concentration trajectory.

6.7 **EKLD comparison of all sample schedules with qPCR noise** EKLD is calculated for all of the optimal schedules as well as the Hatano et al. experiment schedule and compared.

6.8 **Four point optimal D-optimality criterion with ddPCR measurement noise** (a) four optimal points selected by the D-optimality criterion plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the D-optimality criterion.

6.9 **Four point optimal T-optimality criterion with ddPCR measurement noise** (a) four optimal points selected by the T-optimality criterion plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the T-optimality criterion.

6.10 **Four point optimal A-optimality criterion with ddPCR measurement noise** (a) four optimal points selected by the A-optimality criterion plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the A-optimality criterion.
6.11 **Four point optimal E-optimality criterion with ddPCR measurement noise** (a) Four optimal points selected by the E-optimality criterion plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the E-optimality criterion.

6.12 **Four point optimal EKLD sample schedule with ddPCR measurement noise** (a) Four optimal points selected by the EKLD method plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the EKLD method.

6.13 **EKLD comparison of all sample schedules with ddPCR noise.** EKLD is calculated for all of the optimal schedules as well as the Hatano et al. experiment schedule and compared.

6.14 **Comparing Results Using qPCR Assay vs Using ddPCR Assay for an Experiment Optimized for a qPCR Assay.** (a) EKLD is calculated for all qPCR optimized schedules and Hatano et al. schedule assuming qPCR measurement noise. (b) EKLD is calculated for all qPCR optimized schedules and Hatano et al. schedule assuming ddPCR measurement noise.

6.15 **Comparing Results Using ddPCR Assay vs Using qPCR Assay for an Experiment Optimized for a ddPCR Assay.** (a) EKLD is calculated for all ddPCR optimized schedules and Hatano et al. schedule assuming ddPCR measurement noise. (b) EKLD is calculated for all ddPCR optimized schedules and Hatano et al. schedule assuming qPCR measurement noise.

6.16 **Six point optimal D-optimality criterion with qPCR measurement noise.** (a) Four optimal points selected by the D-optimality criterion plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the D-optimality criterion.
Six point optimal T-optimality criterion with qPCR measurement noise. (a) four optimal points selected by the T-optimality criterion plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the T-optimality criterion.

Six point optimal A-optimality criterion with qPCR measurement noise. (a) four optimal points selected by the A-optimality criterion plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the A-optimality criterion.

Six point optimal E-optimality criterion with qPCR measurement noise. (a) four optimal points selected by the E-optimality criterion plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the E-optimality criterion.

Six point optimal EKLD sample schedule with qPCR measurement noise. (a) four optimal points selected by the EKLD method plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the EKLD Method.

Buzon et al Sample Schedule. Sampling Schedule used by Buzon et al. Note: six samples were taken in the Buzon study however, only three samples taken during the 8 week window.

EKLD comparison of all sample schedules with qPCR noise. EKLD is calculated for all of the optimal schedules as well as the Buzon et al. experiment schedule and compared.
6.23 Six point optimal D-optimality criterion with ddPCR measurement noise. (a) four optimal points selected by the D-optimality criterion plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the D-optimality criterion.

6.24 Six point optimal T-optimality criterion with ddPCR measurement noise. (a) four optimal points selected by the T-optimality criterion plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the T-optimality criterion.

6.25 Six point optimal A-optimality criterion with ddPCR measurement noise. (a) four optimal points selected by the A-optimality criterion plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the A-optimality criterion.

6.26 Six point optimal E-optimality criterion with ddPCR measurement noise. (a) four optimal points selected by the E-optimality criterion plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the E-optimality criterion.

6.27 Six point optimal EKLD sample schedule with ddPCR measurement noise. (a) four optimal points selected by the EKLD method plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the EKLD method.
6.28 **EKLD comparison of all sample schedules with ddPCR noise.** EKLD is calculated for all of the optimal schedules as well as the Buzon et al. experiment schedule and compared. 84

6.29 **Comparing Results Using qPCR Assay vs Using ddPCR Assay for an Experiment Optimized for a qPCR Assay.** (a) EKLD is calculated for all qPCR optimized schedules and Buzon et al. schedule assuming qPCR measurement noise. (b) EKLD is calculated for all qPCR optimized schedules and Buzon et al. schedule assuming ddPCR measurement noise. 85

6.30 **Comparing Results Using ddPCR Assay vs Using qPCR Assay for an Experiment Optimized for a ddPCR Assay.** (a) EKLD is calculated for all ddPCR optimized schedules and Buzon et al. schedule assuming ddPCR measurement noise. (b) EKLD is calculated for all ddPCR optimized schedules and Buzon et al. schedule assuming qPCR measurement noise. 88

6.31 **Number of Samples vs EKLD** Number of samples vs EKLD with qPCR noise for EKLD optimal schedules and Buzon et al experiment schedule. 89

6.32 **Number of Samples vs EKLD** Number of samples vs EKLD with ddPCR noise for EKLD optimal schedules and Hatano et al experiment schedule. 90

7.1 **Constrained optimization for 4 samples with qPCR noise.** Sample volume optimization for four point optimal schedule and Hatano et al. experiment schedule. EKLDs shown for volume optimized schedules and uniform volume schedules. 92

7.2 **Constrained optimization for 4 samples with ddPCR noise.** Sample volume optimization for four point optimal schedule and Hatano et al. experiment schedule. EKLDs shown for volume optimized schedules and uniform volume schedules. 93

7.3 **Constrained optimization for 6 samples with qPCR noise.** Sample volume optimization for four point optimal schedule and Buzon et al. experiment schedule. EKLDs shown for volume optimized schedules and uniform volume schedules. 94
7.4 Constrained optimization for 6 samples with ddPCR noise.
Sample volume optimization for four point optimal schedule and
Buzon et al. experiment schedule. EKLDs shown for volume
optimized schedules and uniform volume schedules.
ABSTRACT

Finding a cure for individuals infected with the Human Immunodeficiency Virus (HIV) has proved to be a challenging task. This is primarily due to the fact that conventional treatment has not been able to adequately disrupt the replication process in order to eradicate the virus. One of the possible explanations for this lack of treatment efficacy is that there are low levels of ongoing replication occurring in locations of reduced drug concentration called sanctuary sites. In order to effectively treat the disease, it would be advantageous to clinicians to know how much on-going replication is occurring. This knowledge would then help to guide patient specific treatment for the disease. A novel method to quantify the level on going replication has been suggested. This method entails taking blood samples and measuring biomarkers of on-going. In order to be identified as a valid method clinical trials must be carried out; however, they can often be costly, time consuming and demanding to the patients. For these reasons, meticulous effort should be applied to make sure that these trials are as efficient and informative as possible.

This thesis summarizes several common methods used for optimal design that can be used to address these issues. A mathematical model is first employed to demonstrate the dynamics of the HIV and on-going replication biomarker system. Using this model in conjunction with preliminary laboratory data, Bayesian Markov Chain Monte Carlo Methods are applied to estimate model parameter distributions under a variety of different experiment assumptions. We then calculate the Expected Kullback-Leibler Divergence (EKLD) between the a priori parameter distributions and the a posteriori distributions for each experiment regimen. This value is taken to indicate the amount of information we can expect to gain from performing the experiment under each particular design. Through the use of genetic algorithms we then locate the experiment
design that optimizes the expected gain in information. In doing so, this thesis shows that the EKLD optimization method is robust and performs equally well if not better than traditional optimal experiment design techniques under multiple experiment design criteria. Due to the increased capability provided by the EKLD optimization method in the design of experiments, it should be used in on-going replication quantification experiments in order to maximize information gain and to minimize costs.
1.1 Human Immunodeficiency Virus

The Human Immunodeficiency Virus (HIV) is a retrovirus that targets human immune cells as its host for replication. The virus primarily infects helper T cells for replication; however, both macrophages and dendritic cells have been shown to be viable hosts for HIV replication [16, 35, 88].

Replication of HIV begins with free virions as shown in Figure 1.1. HIV virion surface proteins, gp120 and gp41, facilitate binding of the virion to the target cell primarily via the CD4+ glycoprotein and CCR5 receptor on the surface of the CD4+ T cells [88]. The viral RNA is then reverse transcribed into complementary DNA (cDNA) mediated by the viral enzyme reverse transcriptase. During the reverse transcription process the entire Reverse Transcriptase Complex (RTC) is transported to the nucleus via cellular transport mechanisms. Once the RTC reaches the nucleus the viral cDNA is integrated into the host cell genome carried out by the viral enzyme integrase [89]. If the cDNA is successfully integrated in the host genome transcription of viral RNA can occur, a portion of which is spliced into messenger RNA (mRNA). The mRNA is then translated into regulatory proteins that serve both as a positive feedback mechanism by promoting transcription and bind to unspliced RNA to allow them to be transported out of the nucleus. An allotment of the unspliced RNA serves as mRNA which is translated into packaging and structural proteins used to create new virions. The remaining unspliced RNA is packaged into the budding virion and serve as the new viral RNA copies. Once the new virus is packaged and enveloped the virus then buds off of the host cell and continues to assemble structural proteins and mature.
Once the assembly process is complete the virion is ready to infect a new target cell [16, 35, 88].

Figure 1.1: **HIV Replication Cycle** Replication begins when a free virus attaches itself to a host cell. Its genetic material is then injected into the host in the form of viral RNA. The cellular enzyme reverse transcriptase then creates viral complementary DNA which is then transported into the nucleus. Once in the nucleus, the viral DNA is then integrated into the Host genome, mediated by the host enzyme integrase. The host cell then transcribes new viral RNA from integrated the DNA. Some of this DNA is then translated into structural proteins to create new virions. Once the new virus is constructed and viral RNA is packaged a new virion buds from the host cell and matures, ready to infect a new host.

In vivo viral infection is characterized by three distinct stages of infection: acute infection, chronic infection, and Acquired Immunodeficiency Syndrome (AIDS). During the acute infection phase, which usually lasts for around 9-12 weeks after initial infection, the viral load rapidly increases as the virus undergoes uncontrolled replication [85, 86]. As viral host cells, CD4+ T cells, are depleted and a cytotoxic CD8+ T cell response commences, the viral load levels off and then drops slightly. The virus then enters the chronic infection phase wherein, when the viral load remains relatively stable under the control of the immune system. The chronic infection phase can last for
several years, until eventually the immune system is exhausted and the virus is able to escape the body’s natural immune response. This usually occurs because HIV patients develop opportunistic infections that ultimately cause the viral load to spike and CD4 T+ cell levels to dwindle. When CD4+ T cell concentration levels drop below 200 cells/mm$^3$ the patient is considered to have entered the final stage of infection, AIDS. Individuals diagnosed with HIV/AIDS are expected to survive for three years, if the condition is left untreated [85, 86].

1.2 HIV Treatment

The standard treatment for HIV is Combination Antiretroviral Therapy (cART). This term refers to the use of a combination of two or more antiretroviral drugs from the six classes of HIV medication: fusion inhibitor (FI), Nucleoside Reverse Transcriptase Inhibitor (NRTI), Non-nucleoside Reverse Transcriptase Inhibitor (NNRTI), Integrase Inhibitor (II), and Protease Inhibitor (PI) [76, 77]. Fusion Inhibitors prevent the binding of the HIV virion to host cell receptors. Nucleoside Reverse Transcriptase Inhibitors mimic nucleosides and act to prevent the action of the reverse transcriptase enzyme from creating cDNA from the viral RNA. Similar to NRTIs, Non-nucleoside Reverse Transcriptase Inhibitors prevent the action of the reverse transcriptase enzyme; however, instead of mimicking a nucleoside the NNRTI binds to an active site on the reverse transcriptase enzyme to disrupt the cDNA creation process. Integrase Inhibitors compete with the integrase enzyme to prevent it from integrating viral cDNA into the host cell genome. Protease Inhibitors act to restrict the protease enzyme from producing mature virions. When a subset of these antiretrovirals are used together they are able to effectively control the viral load and maintain viral concentrations well below the level of detection [76, 77].

Although cART is able to control the virus and drastically reduce the viral load, the treatment is not able to completely eradicate the virus from the body. During periods of treatment cessation, viral load is observed to rebound quickly [17]-[25]. There are two leading theories that explain why the drugs are not able to completely clear
the virus: 1. The infected cells are in a dormant state that begin to produce virus upon activation and 2. There are low levels of ongoing replication occurring in sanctuary sites where antiretroviral drugs exist in low concentrations [32, 33, 34, 43, 44]. For the first case when an infected cell has entered a dormant state, also known as latent state, the cell becomes transcriptionally silent [27, 30]. For this reason, it will not produce any viral RNA and subsequently no new HIV virions. There is some disagreement about whether latent HIV cells are infected and then become dormant, or if they are infected while already in the dormant stage. In both cases a free virion binds to the cell, is reverse transcribed, and integrated into the cell’s genome, but the cycle halts at that stage as shown in Figure 1.2. Upon antigenic stimulation the cells become active and being to produce virus [15, 16, 17].

![Latent HIV Cell Diagram](image)

**Figure 1.2: Latent HIV Cell** When an infected cell enters into a latent state, the cell becomes transcriptionally silent. The virus has integrated its DNA into the host genome waiting for the cell to become active again. Once the cell becomes active, it will begin to transcribe new viral RNA. Some of this RNA will get translated into proteins to create new virions. These virions, once constructed will then bud off the surface of the host cell and mature.

The second theory as to the reason why the antiviral drugs are not able to
completely eradicate the virus is due to the low levels of on-going replication. Both Pre-Exposure Prophylaxis (PrEP) and Post Exposure Prophylaxis (PEP), treatment with a heavy dose antiretrovirals before exposure and immediately after exposure respectively, have been demonstrated to be effective in preventing the spread of HIV in individuals who have been exposed to the disease [78, 79, 80]. During the acute phase of infection, viral reservoirs which can lead to on-going replication are seeded. In the first few days following initial infection the virus primarily spreads throughout the mucosa. Afterwards it drains into the lymphatic system where it seeds the viral lymph nodes, a potential viral reservoir. If treatment is administered before this time the virus can be prevented from spreading. Additionally, increased immune activity has been consistently detected in many patients on cART treatment with suppressed viral levels [13, 66, 67]. This suggest that the virus is still actively replicating somewhere in the body. If on-going replication is occurring, it would benefit the clinician to know how much replication is present. This knowledge could help guide treatment and could eventually help completely clear the virus. Consequently, previous studies have sought to quantify the level of on-going replication by detecting HIV 2-LTR circles, viral DNA artifacts which can serve as biomarkers for replication [22, 37, 38, 39, 40, 41, 42].

1.3 Integrase Inhibition and 2-LTR Formation

The viral enzyme, integrase, is responsible for catalyzing two reactions which lead to the integration of viral cDNA into the host cell’s DNA [88]. The first reaction pertains to the preparation of the viral cDNA for integration by cleaving the 3’ ends to expose the binding sites. The second reaction serves to tie cDNA into the host cell’s DNA. Nuclear repair enzymes complete the integration process by sealing the cDNA in place. The class of antiretroviral drugs known as Integrase Inhibitors block this integration by impinging on the binding process [76]. The cDNA is linear when it is created during the reverse transcriptase process; however, when the integration of the cDNA is obstructed cellular DNA repair enzymes combine the two Long Terminal Repeat (LTR) ends to form episomes known as 2-LTR circles as shown in Figure 1.3.
When in a resting state, the metabolic activity of T cells is drastically diminished. Accordingly, production of 2-LTR circles is largely associated with replication of HIV within active T cells, which makes them an excellent surrogate measure of on-going replication [22, 37, 38, 39, 40, 41, 42].

![Diagram of HIV 2-LTR Production Cycle](image)

**Figure 1.3: HIV 2-LTR Production Cycle** If during the process of integrating the viral DNA into the host genome, integration is blocked, host nuclear enzymes will fuse the 2 Long Terminal Repeat ends of the viral DNA. This creates circularized DNA known as “2-LTR circles”, measured and used as a marker of replication.

1.4 Preliminary Data and 2-LTR Dynamics

Previous studies have sought to detect on-going viral replication by measuring 2-LTR concentration after intensifying antiretroviral therapy with an integrase inhibitor [13, 66, 67]. In their study, Buzon et al. randomly selected 69 HIV patients who were on cART treatment and demonstrated a plasma viral load of less than 50 HIV-1 RNA copies/ml for more than one year. The treatment group consisted of 45 patients whose treatment was intensified with the integrase inhibitor Raltegravir. The control group consisted of the remaining 24 patients who continued their treatment without
intensification. The trial was conducted in this manner for a 48 week time span. Patient blood samples were taken at the inception of the trial, and again at weeks 2, 4, 12, 24 and 48. The study was able to detect 2-LTR circles in 13 out of the 45 patients in the treatment group and 1 out of the 22 patients in the control group [1, 66, 67].

Of the 13 patients in the treatment group with detectable 2-LTR circles, the time series dynamics were consistent within the group. The group observed a significant increase in 2-LTR circles at both 2 weeks and 4 weeks when compared to the baseline level at the beginning of the trial. The increase in 2-LTR circle concentration was followed by a decrease in the remaining sample points. This transient behavior provides conclusive evidence that integrase inhibitor blocked active replication events. If a high level of on-going replication is present, a transient increase in the 2-LTR concentration is expected [1, 66, 67]. There will initially be a sharp increase in production of 2-LTR circles as the new infections are inhibited, but the production will then decrease since the success rate of infection events is drastically decreased [1].

Figure 1.4: Experimental Data From Buzon et al. Data from treatment group of the integration inhibitor study done by Buzon et al. Six blood samples were taken and 2-LTR concentration was quantified at day 0, day 14, day 28, day 84, day 168, day 336 or week 0, week 2, week 4, week 12, week 24, week 48. The transient behavior of the 2-LTR dynamics is evident from the data.
1.5 Goals and Organization

The preliminary analysis done by Buzon et al. demonstrates the feasibility of using 2-LTR concentration dynamics post treatment intensification with an integrase inhibitor as a surrogate measure of on-going replication [22, 37, 38, 39, 40, 41, 42, 66, 67]. This study shows tremendous potential to be used as a clinical tool to guide patient specific HIV treatment. In order for the method to be implemented efficiently the trial must be optimized to provide as much information as possible given certain design constraints such as number of samples, volume of blood, and method of quantification [70, 73, 74]. The goal of this thesis is to explore these various design constraints with the intention of optimizing the experiment design in order to yield the maximum amount of information possible from the clinical data.

In pursuance of examining the optimal experiment design for using time series 2-LTR measurements post HIV treatment intensification with an integrase inhibitor to quantify the patient specific magnitude of on-going replication, this dissertation is organized in the following manner. In chapter 2 we introduce a mathematical model for 2-LTR production developed to analyze the data from the Buzon study. In addition, we discuss two different types of DNA quantification assays and their respective accuracy. We also discuss the preliminary analysis of the Buzon data done by Luo et al. [1, 66] We also present the parameter model estimates that came out of the analysis.

In chapter 3 we introduce the design factors that serve as the control variables that we will use to maximize information content in the resulting experiment. We maximize information content, subject to both sample schedule and sample volume. In this chapter we introduce and explain the effect that those variables have on information content in the experiment. We illustrate exactly how and why this occurs by applying them to a simple model.

In chapter 4 we introduce and discuss traditional optimal experiment design methods. Four traditional optimality methods based on the Fisher Information Matrix (FIM) are introduced and described in detail. The purpose and benefits of each method are discussed as well as any disadvantages. We then introduce our method, a Bayesian
Expected Kullback-Leibler Divergence (EKLD) design method and the assumptions and simplifications that were made to calculate it. Next we discuss the optimization methods that were used to find the design that maximizes the information gain for the particular factors for which we are controlling.

In chapter 5 we validate our method and the underlying assumptions used for simplification. We first compare the EKLD for our simplified method to the resulting EKLD from a more exhaustive method to ensure that our simplifying assumption is valid. Next we use our method to predict the amount of information gain that we should expect from a new experiment based on our method. We then compare the actual amount of information that was gained in the new experiment to the amount of information that was expected to be gained to show that our method is able to accurately predict the expected gain in information within a reasonable bound.

In chapter 6 we optimize the sample schedule for the experiment based on the constraints of four samples per experiment and six samples per experiment. This is done using two methods of DNA quantification, quantitative Polymerase Chain Reaction (qPCR) and droplet digital Polymerase Chain Reaction (ddPCR), to compare how the accuracy affects the optimal sample schedule and the amount of information gained for each optimality criteria. In addition we note and illustrate the effect of how increasing the amount of days that samples are taken in the experiment has on expected information gain.

In chapter 7 we solve a constrained optimization problem to find an optimal volume distribution for the samples in the experiment. Generally, blood samples are taken at a consistent volume throughout the experiment for each sample. However, since sample accuracy tends to increase as sample volume grows, large samples are preferred. Unfortunately, there are strict limits which regulate how much blood can be drawn in a given time frame, so we investigate if we can remain within the limits but vary the volume of certain measurements which may require more accuracy in order to maximize information content [72, 81].

Finally, in chapter 8 we summarize the conclusions of our work and discuss the
future implications for what we have presented in this dissertation.

1.6 Impact and Significance

Both the monetary cost and the overall burden to the patient are two of the biggest concerns when designing a clinical trial. The latter is meticulously regulated under regulations imposed by the Institutional Review Board [72, 81]. The monetary costs of clinical trials are not only of utmost concern to the principal investigator of the clinical trial but also to those in the various other echelons involved in financing clinical research, such as funding agencies and public policymakers [72]. Certain fundamental questions must be considered when designing a clinical trial, such as: “What is the expected cost associated with carrying out this trial?”, “what information will be gained from this trial?”, and “how much information will be gained from this trial, and how is this information beneficial?”.

Answers to these questions allow for the establishment of a cost-effectiveness metric or a Value of Information [90]. Government and private agencies use such metrics to help assess and to make decisions about various clinical trials and future medical treatment methods [68, 69]. Our work will maximize the value of information by optimizing the amount of information we gain from performing HIV 2-LTR clinical trials used to quantify the patient specific level of on-going replication.

From a broader clinical data perspective, there are several other applications of the methods that we are proposing. For example, in the hospital setting, insurance will often refuse to pay for certain assessments to be done at high frequency due to the high costs. As a result, physicians have to be judicious in how often the test should be performed [68, 69]. Under these circumstances our method will be able to give the doctors a better idea as to when to perform the tests so that they garner the maximum information about the patient.
Chapter 2

HIV 2-LTR MODEL AND DATA ANALYSIS

2.1 Introduction

Given the significant implications and potential impact of the results from the Buzon et al. integrase inhibitor intensification study, a more thorough understanding of the experiment and the results can give greater insight into the mechanisms which lead to on-going replication. In addition, a better characterization of the temporal dynamics would help to more accurately estimate the level on-going replication in the body’s sanctuary sites. If there are significant replication events occurring due to de novo infection of CD4+ T cells this means that there is a persistent source of infection available to introduce viral mutants into the blood. This occurs because with each round of replication the virus mutates slightly, specifically during the reverse transcription stage of replication [18, 24, 25, 30]. The higher the levels of on-going replication the greater the chance that the virus will mutationally escape the antiretroviral therapy, which can ultimately lead to viral rebound. To gain further insights into the dynamics of the 2-LTR production process post treatment intensification with an integrase inhibitor, Luo et al. developed a mathematical model to analyze the data from Buzon et al. [1].

2.2 2-LTR Model

In order to analyze the data from the Buzon study, Luo et al. developed a mathematical model of 2-LTR circle production following treatment intensification with an integrase inhibitor [1, 2]. Their model is comprised of a two state ordinary differential equation. The two states represent the concentration of 2-LTR circles and
the concentration of actively infected CD4+ T-Cells in the blood. The model takes the form:

\[
\dot{y} = -(1 - (1 - \eta II u II) R)ay + y e \\
\dot{c} = \phi k II (1 - \eta II u II) R ay + k II \eta II u II Ray - \delta c 
\]

(2.1)

where \( y \) is the concentration of actively infected CD4+ T-Cells and \( c \) is the concentration of 2-LTR circles. Parameter definitions and units are defined in Table 2.1. The term \( y e \), is an input into the system and represents entry of actively infected cells from exogenous sources, such as activation of latently infected cells. In both the integrase intensification study done by Buzon et al. and the similar study done by Hatano et al., the patients’ viral load had been ART-suppressed for at least six months [9, 66, 67]. From this, it is safe to assume that the dynamics have reached steady state at the beginning of the experiment when the integrase inhibitor is first administered. Based on this assumption, the solution to 2-LTR concentration dynamics in equation (2.1) can be simplified as follows:

\[
c(t) = c(\infty) + (c(0) - c(\infty))e^{-\delta t} \\
+ c(\infty) \frac{\delta \eta II R (e^{-\delta t} - e^{-a(1 - (1 - \eta II) R)t})}{(1 - R)(a(1 - (1 - \eta II) R) - \delta)} 
\]

(2.2)

with a steady state initial value of

\[
c(0) = \frac{k II y e \phi R}{\delta(1 - R)}
\]

(2.3)

and the final value of

\[
c(\infty) = \frac{k II y e (\phi + \eta II - \phi \eta II) R}{\delta(1 - (1 - \eta II) R)}
\]

(2.4)

The parameter \( R \) is known as the effective reproductive ratio. It is defined in the context of this model as the number of new infections that will stem from one infected cell over the course of its lifetime [57, 58, 59, 60]. The infected cell concentration \( y \) is modeled as the infected cell concentration in the blood. We know that all of
Table 2.1: **Model Parameter Definitions** Definitions and units for all parameters from the 2-LTR model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Definition</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>concentration of infected cells</td>
<td>cells/10^6PBMC</td>
</tr>
<tr>
<td>$c$</td>
<td>concentration of 2-LTR circles</td>
<td>2LTR/10^6PBMC</td>
</tr>
<tr>
<td>$R$</td>
<td>probability infected cell infects a target cell in a generation</td>
<td>unitless</td>
</tr>
<tr>
<td>$a$</td>
<td>death rate of actively infected cells</td>
<td>day$^{-1}$</td>
</tr>
<tr>
<td>$y_e$</td>
<td>rate of exogenous production of infected</td>
<td>infected cells/10^6PBMCxDay</td>
</tr>
<tr>
<td>$\eta_{II}$</td>
<td>Ratio-reduction in R following integrase inhibitor intensification</td>
<td>unitless</td>
</tr>
<tr>
<td>$u_{II}$</td>
<td>binary variable: 1 when integrase inhibitor is applied and 0 when it is not</td>
<td>unitless</td>
</tr>
<tr>
<td>$\phi$</td>
<td>Ratio of probability of 2-LTR formation with integrase inhibitor vs without</td>
<td>unitless</td>
</tr>
<tr>
<td>$k_{II}$</td>
<td>The probability of 2-LTR circle formation when integrase inhibitor is present</td>
<td>2LTR/infected cells</td>
</tr>
<tr>
<td>$\delta$</td>
<td>decay rate of 2-LTR circles</td>
<td>day$^{-1}$</td>
</tr>
</tbody>
</table>

The patients in the study had controlled and undetectable viral levels which implies that the reproductive ratio has an upper bound of one and a lower bound of zero. If there is active replication in sanctuary sites, the local reproductive ratio in the area of the sanctuary site must be greater than one. The parameter $\eta_{II}$ represents the effectiveness of the integrase inhibitor to block integration of the cDNA into the host genome. Parameter $\phi$ is a 2-LTR production parameter that represents the ratio of the natural rate of 2-LTR to the integrase inhibitor enhanced rate of 2-LTR formation. The 2-LTR circles that form decay at a rate represented by the parameter $\delta$. The mass action probability of 2-LTR circle formation when integrase inhibitors are present is represented by parameter $k_{II}$. Unfortunately, $k_{II}$ is not itself identifiable solely based
on the measurement of 2-LTR circle concentration. To make it identifiable we multiply
by \( y_e \) to create the term \( k_{II} y_e \). Upon analysis the parameter is then re-parameterized
by multiplying by \( R/\delta \) to reduce the level of covariance among the parameters yielding
equation 2.5 \[1, 59, 60\].

\[
A = \frac{k_{II} y_e R}{\delta} \tag{2.5}
\]

Together the set of parameters \( \Theta \equiv (A, \phi, R, \eta_{II}, \delta) \) constitute the framework
necessary to estimate the level of on-going replication from the experiment. The optimal
design is one that yields the most information about this set of parameters.

### 2.3 Measurement Error

The model presented in equation 2.2 is useful to estimate the form of the 2-LTR
response following treatment intensification with the integrase inhibitor; however, it
does not take into account measurement noise inherent to the DNA quantification pro-
cess. The experimental data contains this noise so the true model of the measurements
should be

\[
m_i(t_k) = \max \{c(t_k, \Theta_i) + e_k(c)\} \tag{2.6}
\]

where \( c(c_i, \Theta) \) is the 2-LTR concentration for the \( ith \) patient at time \( t \) and sample \( k \). The
\( e_k \) term captures the 2-LTR concentration dependent measurement error introduced
during DNA quantification at each time point. The exact characterization of the noise
\( e_k \) is dependent on the type of assay used; however, there is noise common to all assays
due error presented when taking a sample from the blood.

The first source of noise in a PCR assay is the noise due to sampling from the
blood [51]. Given a small sample size, the number of copies of HIV 2-LTR particles
follows a Poisson distribution with a probability mass function (PMF) as follows

\[
P(n|vc) = \frac{(vc) \cdot e^{-\(vc\)}}{n!} \tag{2.7}
\]
where \( v \) is the volume sampled, \( c \) is the particle concentration in the blood, and \( n \) is the total number of 2-LTR copies present in the sample. After the sample is collected from the blood the second source of error is from the assay used to quantify the HIV DNA. In this dissertation we consider two different types of assay for use in the design of optimal 2-LTR trials: Real-Time Polymerase Chain Reaction and Droplet Digital Polymerase Chain Reaction.

2.3.1 Real-Time Polymerase Chain Reaction Error

Real-Time Polymerase Chain Reaction, also referred to as quantitative Polymerase Chain Reaction (qPCR), is the most common assay used in the quantification of DNA. This assay consists of adding DNA specific probes to a solution with the target DNA \([22, 29, 53, 63]\). The DNA is then amplified and the probe reporters are detected in real time. The quantification cycle is then determined based on the properties of the amplification curve. This results in a qPCR assay with the probability mass function

\[
P(m|n) = \begin{cases} 
\exp \left( -\frac{(\ln m - \ln n)^2}{2\sigma(n)^2} \right), & m > 0 \\
1, & m = 0 \\
0, & Otherwise 
\end{cases}
\]

(2.8)

where \( \sigma(n) = 10^{-0.21 - (0.24 \times \log_{10}(n))} \) is the equation for the log normal standard deviation of the qPCR growth process as a function of viral concentration with the measured concentration denoted by \( m \) and the viral particles present in the samples denoted by \( n \) \([29, 51, 53]\).

2.3.2 Droplet Digital Polymerase Chain Reaction Error

Droplet Digital Polymerase Chain Reaction (ddPCR) quantification is very similar to the qPCR with one primary difference. After the sample is taken from the blood it is separated into thousands of nanoliter sized droplets prior to amplification and labeling. Using microfluidic technology the droplets are then classified as either positive
for containing viral particles or negative for containing viral particles. This method is much more accurate because the probability that there are multiple copies per droplet is very low, therefore the total positive droplet concentration is able to provide a good estimate of the true sample concentration [54, 91]. For the ddPCR assay the probability of d positive droplets given a blood concentration can be modeled as a simple binomial function

$$B(N, p)$$

where N is the total number of droplets and $$p = 1 - e^{(\frac{c \times v_d}{N})}$$. The concentration of viral particles in the blood is c and the droplet volume is represented by $$v_d$$.

A further analysis of these equations shows us that the noise in the qPCR measurement is largely due to the noise in the qPCR measurement process whereas the noise in the ddPCR measurement assay is largely dominated by the noise of the sampling process. For this reason the ddPCR assay yields a much more accurate estimate of the true concentration than the qPCR assay [54, 91].

2.4 Data Analysis

In their experiment, Buzon et al used a qPCR assay to measure the 2-LTR circle concentration, therefore in the analysis of the data measurement error consistent with a qPCR assay was adopted. Patient specific parameter estimates were calculated for parameters $$A$$, $$R$$, and $$\eta_{II}$$. The 2-LTR decay rate $$\delta$$ and the mass action ratio probability of 2-LTR formation with integrase inhibitor vs without $$\phi$$ are not considered to vary within the patient group so they were considered group specific parameters [1, 2, 66, 67].

2.5 Posterior Parameter Distributions

Posterior distributions were calculated using a Bayesian Markov Chain Monte Carlo (MCMC) method with Gibbs Sampling by Luo et al. [1, 2]. We have taken these posterior parameter distributions and used them as our prior parameter knowledge. In doing so, the parameters were concatenated to form a global parameter distribution set
across all patients. The parameters are illustrated in a concatenated fashion in Figure 2.1 for the patients in the treatment group.

Figure 2.1: Concatenation of Priors for Patients in Buzon et al. study. Based on the data from the experiment done by Buzon et al., Luo et al., performed a Bayesian Markov Chain Monte Carlo analysis to estimate 2-LTR mode posterior parameter distributions for the 13 patients in the study that exhibited 2-LTR dynamics. These parameter distributions were then concatenated and used as the prior knowledge about the parameters for the experiment optimization.

With the exception of parameter $R$, confidence intervals for parameters are not tightly bounded. Without satisfactory estimates for the parameters, it is difficult to estimate the level of on-going replication with certainty. Due to the lack of information in the data it also difficult to distinguish between any hypotheses. Fortunately, this shortfall of experimental information can be improved upon simply by changing design factors which allow for the collection of more informative data. This dissertation will explore those very factors and determine designs which optimize the amount of information gained from the experiment [1, 2, 66, 67].
3.1 Introduction

In order to optimize the design of the experiment to maximize information content we must vary certain factors that affect information content. We must then find the values of these factors that yield the greatest gain in information. In this thesis we focus on information content of experiment designs subject to days at which samples are taken and the volume of blood that is drawn at each sample. To get a better understanding of how these two elementary factors can affect information content in the experiment we first investigate them applied to a straightforward problem.

3.2 Simple Problem

Our model is non-linear and contains a set of five parameters so to investigate the effect that sample schedule and sample volume have on information content we look at how altering them effects the information gain for a simple exponential growth model as in equation 3.1.

\[
\dot{C} = \alpha C \\
\dot{C}(t) = C(0)e^{\alpha t}
\]  

(3.1)

We intend to optimize the experiment to maximize the information content about the exponential growth parameter \( \alpha \).

Figure 3.1a illustrates trajectories for various different values of \( \alpha \) as constituted by the prior distribution (blue) in figure 3.1b. A measurement is taken at time seven (\( t = 7 \)) with the uncertainty expressed by the error bars. Trajectories that fall within
(a) Concentration trajectories based on different values of parameter $\alpha$ and a sample taken at time seven.

(b) Prior and posterior distributions for parameter $\alpha$.

Figure 3.1: **Experiment Design** (a) The graph in the figure shows the trajectories of the exponential growth model based on different values of the model parameter $\alpha$. The figure shows a sample taken at time seven ($t = 7$) with an associated confidence interval denoted by the solid black circle marker and the black error bars. The trajectories that pass through the confidence region are colored red and the trajectories that do not pass through the confidence region are colored blue. (b) The histograms show the prior and posterior distributions of the model exponential growth parameter $\alpha$. Prior distribution is colored blue and corresponds to the range of blue trajectories in the above plot. The posterior distribution is colored red and corresponds to the trajectories which pass through the region of confidence of the measurement in the above plot.
the error bars constitute values of $\alpha$ that would make up the posterior distribution (red) in Figure 3.1b. The divergence between the two distributions quantifies the gain in information due to taking the sample.

### 3.3 Sample Schedule Optimization

First we investigate the effect that changing the time at which a sample is taken has on the amount of information gained in the experiment. In Figure 3.1a the sample is taken at time seven ($t = 7$). We now take a sample at time five and observe how the information gain changes.

Figure 3.2a again shows the concentration trajectories for different values of the parameter $\alpha$; however, the sample is taken at time five ($t = 5$) this time. We observe that when the sample is taken earlier in the time schedule more of the trajectories fall within the region of uncertainty based on the measurement error as defined by the error bars. As a result the posterior distribution in figure 3.2b is more broad than the posterior distribution in figure 3.1b when the sample was taken at time seven ($t = 7$). The divergence between the prior and posterior distributions is smaller when the sample is taken at time five ($t = 5$) when compared to the divergence between the distributions when the sample is taken at time seven ($t = 7$). This means there is less information gained in the experiment design where a sample taken at time five ($t = 5$) vs. time seven ($t = 7$).

### 3.4 Volume Optimization

Next we investigate the effect that changing the sample volume has on the amount of information gained for the simple model. For the case in Figure 3.3a we again take a sample at time seven ($t = 7$) as in Figure 6.1; however, in this case we simulate a greater sampling volume. Increasing the volume sampled will escalate the accuracy of the measurement characterized by a decrease in uncertainty and reduced bounds for the error bars.
(a) Concentration trajectories based on different values of parameter $\alpha$ and a sample taken at time five.

(b) Prior and posterior distributions for parameter $\alpha$.

Figure 3.2: **Experiment Design (Changing Sample Day)**  (a) The graph shows the sample moved from time point seven ($t = 7$) to time point five ($t = 5$). More of the trajectories now pass through the region of confidence for the measurement. (b) The histograms show the prior and posterior distributions based on the measurement at time point five ($t = 5$). Because the measurement has been taken sooner and more trajectories pass through the region of confidence, the posterior distribution is broader when compared to the posterior distribution from the measurement at time point seven ($t = 7$). This leads to less information gained from taking the measurement sooner.
(a) Concentration trajectories based on different values of parameter $\alpha$, a sample taken at time seven with increased accuracy.

(b) Prior and posterior distributions for parameter $\alpha$.

Figure 3.3: **Experiment Design (Changing Sample Volume)**. (a) The graph shows the sample again at time point seven ($t = 7$); however, a large volume is taken during the sample. This creates a more accurate measurement marked by the reduced size of the error bars. Now, fewer trajectories pass through the region of confidence for the measurement. (b) The histograms show the prior and posterior distributions based on the measurement at time point seven ($t = 7$) but with the increased accuracy due to the larger sample volume. Since fewer of the trajectories pass through the region of confidence of the measurement the posterior distribution of the model exponential growth parameter $\alpha$ is significantly more narrow. This leads to a larger gain in information from taking this sample when compared to the other samples.
Figure 3.3a again shows the concentration trajectories for varying values of parameter $\alpha$. The sample is taken at time seven ($t = 7$) as in Figure 3.1a, but error bars are noticeably smaller in magnitude. As a result, fewer trajectories past through the region of confidence for the measurement. This leads to the narrow posterior distribution shown in figure 3.3b. Consequently, the divergence between the prior distribution and the posterior distribution for the sample at time seven ($t = 7$) with increased volume is greater than that of the original sample taken at time seven ($t = 7$). This would mean that there would be a greater gain in information for this experiment design.
4.1 Introduction

Experiment design has been carried out since man has been conducting experiments. Early evidence of experiment design can be found in the Old Testament of the Bible and the earliest evidence of clinical trial design can be found in *The Canon of Medicine*, a medical encyclopedia written by Persian scientist Avincenna in 1025 [87]. Since then scientists and mathematicians have continuously improved upon the techniques used in the design of experiments [73]. In this section we will introduce some modern, statistically based, methods used for design of experiments including traditional methods as well as our proposed method.

4.2 Traditional Methods

Early methods of experiment design focused on fundamental aspects of design such as establishing controls, encouraging replication, minimizing confounding effects, and making observations for several different factors [87]. These factors continue to be the basis for adequate experiment design. The factors were initially established by inspection based on previous knowledge about the system under experiment; however, at the dawn of the 20th century mathematicians began to apply statistics to aid in the design of experiments. This movement was pioneered by British statistician Ronald Fisher, who established a popular method to measure the amount of information in a given experiment design. This method uses a measure which bares his name, known as the Fisher Information Matrix [71, 73, 74].
4.2.1 Fisher Information Matrix

The Fisher Information Matrix (FIM) is a statistical method which produces a measure of the amount of information about a parameter set from a given set of observations. The majority of the common measures of optimality focus on maximizing the information content in a given experiment. This is usually done by optimizing some criteria of the Fisher’s information [45, 70, 71, 74, 75]:

\[
M(\Theta) = \int \left( \frac{\partial}{\partial \Theta} \log P(x; \Theta) \right)^2 f(x; \Theta) dx \tag{4.1}
\]

which can be expressed as a matrix

\[
M(\Theta)_{kl} = \mathbb{E} \left[ \left( \frac{\partial}{\partial \Theta_k} \log P(x; \Theta) \right) \left( \frac{\partial}{\partial \Theta_l} \log P(x; \Theta) \right) \right] \tag{4.2}
\]

where \( \Theta \) is a vector containing the model parameters, \( x \) is data observed in the experiment and \( P(x; \Theta) \) is probability density function for the observed data conditioned on the model parameters.

The choice of which criteria to optimize can be highly subjective. In this thesis we analyze the outcome of four frequently used Fisher Information Matrix based experiment design optimization paradigms: A-optimal, D-optimal, T-optimal, and E-optimal. We then compare them to our novel Expected Kullback-Leibler method.

4.2.2 D-optimality Criterion

The D-optimal method is perhaps the most commonly used optimality criteria and seeks to minimize determinant of the Fisher Information Matrix

\[
\max \{ \text{det} (M(\Theta)^{-1}) \} \Rightarrow T = [t_1, t_2, ..., t_n] \tag{4.3}
\]

subject to the experiment design regimen \( T \), a vector of \( n \) sample points in the study. In doing so it minimizes the \( m \) dimensional ellipsoid of the maximum region of confidence for the maximum likelihood estimate of the parameters \( \Theta \) [73, 74]. This in essence minimizes the covariance of the parameter estimates.
4.2.3 T-optimality Criterion

The T-optimal method is another criterion often used due to its relative mathematical simplicity. This method simply seeks to maximize the trace of the information matrix $M(\Theta)$.

$$\max \{ \text{trace}(M(\Theta)) \} \Rightarrow T \quad (4.4)$$

While this method is straightforward mathematically and computationally simple, it is also considered to be a very weak and unsatisfactory optimization mechanism [73, 74].

4.2.4 A-optimality Criterion

The A-optimal method seeks to minimize the trace of $M(\Theta)^{-1}$ and in doing so minimizes the sum of the variances [73, 74].

$$\min \{ \frac{1}{m} \text{trace}(M(\Theta)^{-1}) \} \Rightarrow T \quad (4.5)$$

This optimization paradigm is commonly chosen due to its relative mathematical simplicity. In the case of simple linear models, closed form A-optimal solutions can often be found.

4.2.5 E-optimality Criterion

The fourth and final traditional method of optimization that is considered in this paper is the E-optimal design approach. Similar to the D-optimal design, the E-optimal technique is a bit more computationally intensive. It seeks to maximize the minimum eigenvalue of the information matrix $M(\Theta)$.

$$\max \lambda_{\min} \{ (M(\Theta)) \} \Rightarrow T \quad (4.6)$$

This method is often used as an alternative to the D-optimal method when one more of the parameters has a relatively large variance in comparison to the other
parameters. Graphically, the E-optimal method minimizes the maximum diameter of the m dimensional ellipsoid [71, 73, 74].

There are numerous other traditional optimality criteria; however, these four were chosen to demonstrate how the various methods result in different optimal designs and to compare common methods to our novel approach.

4.3 Expected Kullback-Leibler Divergence

The Kullback-Leibler Divergence was developed as way to measure the difference, or divergence, of two probability distributions, which in the context of experiment design can be interpreted as the information gained from performing the experiment with a given set of experimental conditions [50, 61, 61].

\[
D_{KL}(P(\Theta; x)||P(\Theta)) = \int_{-\infty}^{\infty} P(\Theta; x) \ln \frac{P(\Theta; x)}{P(\Theta)} dx
\]  
(4.7)

Equation 4.7 denotes calculation of the KLD between two probability distributions \( P(\Theta; x) \) and \( P(\Theta) \).

4.3.1 Simulated Patient Pool

To evaluate a sample schedule we must evaluate it against a number of possible patients. The posterior distributions developed by the Luo et al. analysis of the Buzon data represents the range of possible parameter values possible patients may have [1, 66].

The multivariate distribution is constructed from a set of five system parameters \( \Theta(A, \phi, R, \eta_{II}, \delta) \). Parameters \( R, \phi, \eta_{II} \), and \( \delta \) are exactly established from equation 2.1. Parameter \( A \) was derived as an observable parameter which reduces the covariance between other parameters [1].

\[
A \equiv \frac{k_{II} \eta_{II} R}{\delta}
\]  
(4.8)
In order to evaluate a sampling schedule’s performance across multidimensional multivariate distribution while maintaining computational tractability, we apply an unscented transform to obtain $2N + 1$ simulated sigma point patients which maintain the same first and second moment characteristics as the initial multivariate distribution [49, 61, 62].

$$X_i = \begin{cases} 
\mu, & i = 0 \\
\mu + \sqrt{N}\Sigma_i, & 1 \leq i \leq N \\
\mu - \sqrt{N}\Sigma_i, & N < i \leq 2N 
\end{cases}$$

(4.9)

where each $X_i$ is the separate set of parameters for each sigma point patient. The $\mu$ and $\Sigma_i$ terms are the mean of the prior distribution and the $i$th column of the covariance matrix of the prior distribution respectively. $N$ is the total number of dimensions in our prior distribution, which in this case is a five dimensional distribution.

The five dimensional parameter space results in 11 sigma point patients that exhibit a range of 2-LTR concentration dynamics as illustrated in figure 4.1. The transient nature of the dynamics is evident from the various trajectories.

4.3.2 Calculation

4.3.2.1 Markov Chain Monte Carlo (MCMC) Methodology

For each candidate schedule we construct simulated data based on our models and measurement noise. The posterior distributions for parameter set $\Theta_i$ are constructed for each patient $i$ using a Markov Chain Monte Carlo technique [1, 2, 55, 56, 60]. We define $c(t_k, \Theta_i)$ as the true concentration measured at sample point $k$ using parameter set $\Theta_i$. 
With a qPCR assay we assume log normal measurement noise consistent with the assay, based on work done in [51], which leads to measurements as

\[ m_{ik} = \ln N(\ln(c(t_k, \Theta_i)), \ln(10)\sigma(n)) \]  (4.10)

Applying Bayes theorem we arrive at the equation

\[ P(\Theta_i|m_{ik}) = \frac{\mathcal{L}(\Theta_i|m_{ik})P(\Theta_i)}{\int_0^\infty P(m_{ik}|\Theta_i)P(\Theta)d\Theta} \]  (4.11)

However, \( \int_0^\infty P(m_{ik}|\Theta_i)P(\Theta)d\Theta \) is a constant scaling factor of the posterior distribution [51]. For computational simplicity we simplify and arrive at the equation

\[ P(\Theta_i|m_{ik}) \propto \mathcal{L}(\Theta_i|m_{ik})P(\Theta_i) \]  (4.12)

which has the same form and conserves the KLD [71].
4.3.2.2 KLD Calculation

Calculation of the Kullback Leibler Divergence (KLD) between the five dimensional multivariate prior and posterior distributions is done using equation

\[
KLD(T) = \frac{1}{2} \left( \log_2 \left( \frac{\det \Sigma_2}{\det \Sigma_1} \right) - n - tr(\Sigma_2^{-1} \Sigma_1) + (\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) \right) (4.13)
\]

where \((\mu_1, \Sigma_1)\) and \((\mu_2, \Sigma_2)\) are the mean vector and covariance matrices of the prior and posterior multivariate distributions respectively and \(n\) is the number of dimensions in the distribution [71]. Because \(\log_2\) is used in equation 4.13, the unit of the resulting Kullback Leibler Divergence will be bits.

We should also note that equation 4.13 is applicable when all of the parameters are normally distributed. \(\log(A), \log(\phi), \text{and } \log(\delta)\) are normally distributed. Parameters \(\eta_{II}\) and \(R\) are transformed using the normal distribution quantile function. The KLD between distributions is conserved through all transformations [61, 62, 82].

The Expected Kullback Leibler Divergence (EKLD) is estimated by calculating the KLD for each sigma point patient and multiplying by the probability of the patient occurring based on the parameter distributions calculated by Luo et al.

\[
EKLD(T) = \sum_{i=1}^{11} KLD(T)_i P(\Theta_i) (4.14)
\]

where \(i\) represents a patient from the simulated sigma point patient pool sampled from the Prior Distribution. Calculating the KLD for only the 11 sigma point patients is a simplification made to make the calculation more computationally feasible. The true EKLD would require that the KLD is calculated and integrated over the entire prior distribution.

\[
EKLD(T) = \int_i KLD(T)_i P(\Theta_i) (4.15)
\]

It has been shown the sample schedule order is preserved when KLD is calculated using the simplified method in lieu of integrating over the entire distribution [61, 62].
4.3.3 Parameter Transformations

Equation 4.13 only holds if all of the parameters are normally distributed. Fortunately, $\log(A)$, $\log(\phi)$, and $\log(\delta)$ are normally distributed. Parameters $\eta_{II}$ and $R$ follow a beta distribution and must be transformed to a normal distribution before calculating the KLD. This transformation is done using a normal quantile function [82]. Even though the parameters are transformed there is no effect on the information gain as the KLD is conserved through all transformations.

![Figure 4.2: Transformation of Prior and Posterior Distributions to Normal Distributions and Back.](image)

The prior and posterior can be transformed to a normal distribution and a fully restored back to its original distribution as shown in figure 4.2. In this figure the prior (red) uniform distribution is transformed to a normal distribution and then back to uniform, while the posterior (blue) is transformed from a beta distribution to a normal distribution and then back to a beta distribution. The KLD between the two distributions remained consistent through out all of the transformations.
4.4 Optimization

4.4.1 Genetic Algorithm

The inherent binary nature of time series measurements, taking a sample on a given day or not in this case, lends itself well to a genetic algorithm optimization method. To construct the GA, candidate sample schedules are represented by a chromosome as shown in figure 4.3. Each chromosome consists of genes and each gene is further broken down into base pairs. Each base pair represents a potential sample day and takes on a binary value, 0 for days at which no sample was taken, and 1 for days at which a sample is taken. The genes combine to form a chromosome with total base pairs equal to the total number of possible sample days.

![Diagram of Genetic Algorithm](image)

Figure 4.3: **Formulation of the Genes and Chromosomes in the Genetic Algorithm.** Genetic algorithm shown with three genes per chromosome and 19 basepairs per gene. Each basepair corresponds to a sample day with 1 being a day when sample is taken and 0 being a day when no sample is taken. Arrows denote different types of crossovers that can occur.

The algorithm is run with 20 child chromosomes per generation. The first generation is created by randomly selecting the appropriate number of base pairs (sample days) per chromosome. The corresponding information content is then calculated for each chromosome by calculating its associated KLD. The chromosomes are then ranked in terms of the relative fitness by assigning chromosomes yielding higher KLD values a greater fitness level. Chromosomes with the highest fitness are then used as the parent solutions to create the children for the next generation. Children are created through a process of genetic crossovers and mutations [46, 47, 48].
Genes are able to crossover to different locations or to the same location between parents. Point mutations occur after the crossovers to ensure that the chromosome has exactly number of sample points desired. These mutations occur by bit inversion. This is to ensure that the algorithm is able to escape local minima [46, 47, 48].

### 4.4.2 Constrained Optimization

To optimize our experiment design based on sample volume we formulated a constrained optimization problem. Because the amount of blood which can be drawn in a given amount of time is heavily regulated by the Institutional Review Board, there is a upper bound the total amount of blood drawn throughout the experiment [72, 81]. We set this upper bound and allowed the volume of each sample to vary subject to the requirement that the total volume be less than or equal to the upper bound. This is done for both the four sample and the six sample design.

For the four sample case, we perform the constrained optimization to find the best volume distribution for both the four sample optimal schedule and the Hatano et al. experiment schedule. Approximately 20 ml of blood were drawn per sample in the Hatano et al study so as a fair comparison we set the upper bound for the total volume drawn throughout the experiment at 80 ml.

For the six sample case, we again perform the constrained optimization to find the best volume distribution for both the six sample optimal schedule and the Buzon et al. experiment schedule. Approximately 20 ml of blood are also drawn per sample in this study so as a fair comparison we set the upper bound for the total volume drawn throughout the experiment at 120 ml.
Chapter 5
METHOD VALIDATION

5.1 Introduction

We have established the factors that will be altered to find the optimal schedule as well as the some of the various criterion that is used for optimal experiment design. The Fisher based optimality criterion methods have been established as valid methods for optimal experiment design; although, which criterion is used is heavily dependent on which of the characteristics the investigators wish to optimize [73, 74]. In this chapter we seek to validate our method and its associated assumptions. We start by validating our patient pool assumption and then we compare information content of actual 2-LTR data from a follow on experiment vs the information gain predicted based on our EKLD method.

5.2 Hatano et al. Experiment

While the experiment was not as informative as we would have wished, the data presented from the Buzon et al. trial did prove to be promising as 2-LTR circles were detectable in 13 out of the 45 patients post treatment intensification [13, 66]. Consequently, a follow on experiment was conducted by Hatano et al to gather time series 2-LTR concentration data post integrase inhibitor intensification [9]. In their experiment, Hatano et al., attempted to improve upon the previous experiment design used in the Buzon study. The experiment was shortened to 8 weeks with time samples closer together in the beginning weeks of the trial, as opposed to the 48 week long Buzon study. Unfortunately, they did not take as many samples as only four samples were taken at week 0 (beginning of the experiment), week 1, week 2 and week 8.
Hatano et al. also used a droplet digital polymerase chain reaction (ddPCR) assay instead of the real time polymerase chain reaction (qPCR) assay used in Buzon et al. The ddPCR assay will yield much more accurate measurements as shown in Figure 5.2. In the figure we can see that the 95 percent confidence interval of the measurement assay is much more narrow at the various 2-LTR concentrations throughout the duration of the experiment for the ddPCR assay. Due to the transient nature of the 2-LTR dynamics, it is of the utmost importance that the details regarding the peak of the curve are captured accurately [1, 57, 61]. The ddPCR assay is able to better ensure that these details are captured and will ultimately lead to a greater gain in information [54].

5.3 1000 Monte Carlo vs Unscented Transform EKLD

To make our EKLD method more computationally tractable we used the unscented transform to create the minimum amount of simulated patients that would still represent the original patient distribution [49]. In this section we validate our assumption that the 11 unscented transform patients are a legitimate representation of the entire five dimensional sample space, with respect to information gain. To do so we
Figure 5.2: Measurement Error ddPCR vs qPCR in Quantifying 2-LTR dynamics. A 2-LTR trajectory is plotted with 95 percent confidence regions shaded for ddPCR assay (red) and qPCR assay (green). The ddPCR assay is much more accurate at all concentrations which leads to more information from the experiment.

take 1000 Monte Carlo (MC) random samples from our possible patient distribution and calculate the mean Kullback-Leibler Divergence. We then compare the result to the EKLD calculated using the 11 unscented transform patients. This is done for four different schedules, the EKLD optimal four point schedule, the Hatano et al. experiment schedule, the EKLD optimal six point schedule, and the Buzon et al. experiment schedule. We also calculate the EKLD for both DNA quantification assays, qPCR and ddPCR, and compare the results.

First the KLDs are compared assuming qPCR measurement error as illustrated in Figure 5.3. For all of the sampling schedules the MC patients yielded a higher gain in information than the UT patients. Most importantly, the order is preserved for each schedule relative to the amount of information gained. The four sample optimal schedule out performs the schedule from the Hatano et al. experiment and the six
Figure 5.3: **1000 Monte Carlo Patients vs Unscented Transform Patients with qPCR Assay.** KLD for 1000 Monte Carlo Patients Compared to EKLD From UT Patients Assuming qPCR Measurement Noise. The 1000MC patients are displayed as the dark blue bars and the UT patients are displayed as the yellow bars. KLDs are compared for four different schedules: four sample optimal schedule, Hatano et al. experiment schedule, six sample optimal schedule, and Buzon et al. experiment schedule.

The four sample optimal out performs the schedule used in the Buzon et al. experiment for both the MC and UT KLD analyses. Additionally, the six sample optimal schedule out performs the four sample optimal and the Buzon et al. and Hatano et al. schedules perform similarly for both the MC and UT analyses. This shows that our simplification is valid in locating the design that will provide the highest gain in information.

Next the KLDs are compared assuming ddPCR measurement error as illustrated in Figure 5.4. Again the MC patients showed a higher gain in information for the Hatano et al. experiment schedule, the six sample optimal schedule, and the Buzon et al. experiment schedule. However, the four sample optimal schedule provided slightly more information from the UT patients compared to the MC patients. This is due to the fact this schedule has points which attempt to captures the peak of the 2-LTR curve and transient dynamics. Many of the simulated MC patients have lower
peaks and weak transient behavior, which will skew the results showing a higher expected gain in information for the UT patients. Again the order is preserved for all of the schedules. The four sample optimal schedule has a higher gain in information than the Hatano et al. experiment schedule and the six sample optimal schedule has a higher gain in information than the Buzon et al. experiment schedule. As expected, the six sample optimal is still the best performing schedule followed by the optimal four point schedule. The Hatano et al experiment schedule and the Buzon et al. experiment schedule again yield a similar amount of information gain with the Hatano et al. schedule yielding a slightly higher gain in information for both the MC and UT patients analyses. Furthermore, all of the schedules yield more information when analyzed with ddPCR measurement error vs qPCR measurement error as expected. All together, this proves our assumption to be valid and robust for both assays.
Figure 5.5: **UT vs 1000 MC Correlation Plot (qPCR)**. Plot comparing the KLD for 100 different sample schedules comparing when using 1000 randomly Monte Carlo sampled patients vs. when using the 11 unscented transform patients assuming qPCR measurement noise. Correlation coefficient $\rho = 0.927$.

To further validate the unscented transform method, we compare the EKLD resulting from UT patients vs the EKLD from the MC patients for 100 randomly chosen four point sample schedules from the eight week day sample space. The product of this analysis is displayed in figure 5.5 assuming a qPCR measurement assay. The figure shows the EKLD the 11 Unscented Transform patients vs for the 1000 Monte Carlo sampled patients for each of the 100 sample schedules. From the data in the figure, there is a strong correlation between the two methods. Indeed, the correlation coefficient calculated for the data is 0.927, which mathematically confirms the strong correlation between the MC method and the UT method [83, 84].

An identical analysis was performed comparing the UT method to the MC method for 100 randomly chosen four point sample schedules assuming a ddPCR measurement assay. The 100 schedules were the same sample schedules used in the qPCR assay analysis. The EKLDs for the two methods are plotted in figure 5.6. Again, we observe a strong correlation, of the EKLDs, between the two methods. The correlation
coefficient for the EKLD data is 0.950, proving a strong correlation exists between the two methods [83, 84]. The correlation is slightly stronger for the ddPCR assay due to the increased accuracy of the measurement.

5.4 Predicted Information Gain in Hatano Experiment

To verify the legitimacy of our EKLD method we use the experimental data from the integrase inhibitor intensification trial, conducted by Hatano et al., to calculate the actual amount of information gained in the experiment measured by Kullback-Leibler Divergence. The bar graph in Figure 5.7 compares the EKLD from the optimal four sample schedule and Hatano et al. sample schedule to the actual KLD calculated using that experimental data. The optimal sample schedule is expected to yield a 6.08 bit gain information and the sample schedule used by Hatano et al. was expected to gain 3.80 bits of information; however, based on the data there was only an average of 2.16 bits of information gained from the 15 patients in the experimental group.
Figure 5.7: **EKLD comparison of Hatano et al. Experimental Patient Data**

KLD calculated from the actual Hatano et al. experimental data (green). Results are compared to EKLD of the Hatano et al. schedule (blue), and the optimal four point schedule (red).

Based on the EKLD method, we expected to gain an additional 1.64 more bits of information that what was observed in the experimental data. We believe that the actual gain in information gained was lower than expected due to the fact that there was a very small sample size of only 15 patient and that of those patients very fewer than expected actually exhibited 2-LTR dynamics consistent with on-going replication. This theory was confirmed by fitting the data to model and a calculating the Aikake Information Criteria (AIC) in order to reject the null hypothesis that any change in 2-LTR concentration was due to random noise. Additionally, prior distributions were created from posterior distributions from the Buzon et al. data analysis. Only the 13 patients in the study with significant 2-LTR dynamics were used, which may have biased our prior distribution.

A better picture of the information content in the experimental data can be realized from Figure 5.8, which shows the Kullback-Leibler Divergence for each patient individually. The dynamics of top two most informative patients were very descriptive.
Figure 5.8: **EKLD comparison of Hatano et al. Experimental Patient Data compared by individual patient KLD** KLD calculated from the actual Hatano et al. experimental data and displayed for each of the 15 patients in the treatment group (green). Results are compared to EKLD of the Hatano et al. schedule (blue), and the optimal four point schedule (red).

yielding a 6.54 and 5.82 gain in information respectively. Unfortunately, 10 out of the 15 patients produced less than a two bit gain in information and five produced less than a one bit gain in information. This heavily skewed result caused the mean information gain to shift to a less than expected magnitude.
Chapter 6
SAMPLE SCHEDULE OPTIMIZATION

6.1 Introduction

As previously addressed, altering the sample schedule can have tremendous effects on the amount of information gained in the experiment. Given the high cost associated with taking blood samples, it is paramount that the most informative sample points are chosen [68, 69]. In this section we use our EKLD method to optimize the gain in information for 2-LTR experiments with four samples and 2-LTR experiments with six samples. Additionally we investigate the effect that the sample assay chosen to quantify the HIV 2-LTR DNA concentration has on the resulting gain in information. In all cases we compare our method to optimal sample schedules selected by traditional Fisher Information Matrix based methods.

6.2 Four Sample Optimization

First we optimize our sample schedule to find the best four sample schedule over a period of eight weeks. In doing so we can compare the optimal schedules to the actual schedule used in the experiment done by Hatano et al.

6.2.1 Real-Time Polymerase Chain Reaction Design

We start by optimizing the experiment for four samples using a qPCR assay. Optimal schedules are found for the following Fisher Information Matrix based methods, D-optimality criterion, T-optimality criterion, A-optimality criterion, and E-optimality criterion as well as for our EKLD method. The resulting optimal schedules are shown plotted on an indiscriminate 2-LTR concentration curve. The genetic algorithm converges to a family of solutions as it searches in the region of the optimal schedule.
This family of solutions is analyzed for each optimal schedule with a plot showing the frequency with which each sample appeared within the top performing family of solutions.

### 6.2.1.1 D-optimality Criterion

The first four sample solution that we will investigate is the D-optimality criterion solution. As previously stated the D-optimality criterion seeks to minimize the five dimensional ellipsoid of the maximum region of confidence for the maximum likelihood estimate of the parameters Θ in our model. To do so, it attempts to maximize the determinant of the inverse of the Fisher Information Matrix.

The four point sample schedule that optimizes the experiment design subject to the D-optimality criterion chose samples at day 0, day 1, day 7, and day 22 as shown in figure 6.1a. This criterion chose to take a sample at the inception of the experiment, another to capture the increase in 2-LTR concentration, one to capture the decrease in 2-LTR concentration and one as the dynamics are settling down to steady state. Figure 6.1b shows the frequency that each sample day was selected for in the family of solutions. The sample at time zero appears to be the most critical as it is selected in all schedules and the point at day 1 is a little less critical but still very important and is selected in almost 80 percent of the schedules. The exact location of the remaining two points appears to be less critical as they are selected for in less than 20 percent of the schedules. Instead this criterion prefers the two remaining sample points to be from a region around day 8 and day 20 respectively.

### 6.2.1.2 T-optimality Criterion

The next four sample solution that we will investigate is the T-optimality criterion based solution. To optimize the experiment design T-optimality method seeks to maximize the trace of the Fisher Information Matrix. This method is computationally simple but does not take into account any underlying relationships among the parameters [71, 73, 74].
Figure 6.1: **Four point optimal D-optimality criterion with qPCR measurement noise** (a) four optimal points selected by the D-optimality criterion plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the D-optimality criterion.
Figure 6.2: **Four point optimal T-optimality criterion with qPCR measurement noise** (a) four optimal points selected by the T-optimality criterion plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the T-optimality criterion.
The four point sample schedule that optimizes the information content subject to the T-optimality criterion are consecutive samples at day 2, day 3, day 4 and day 5 as shown in figure 6.2a. This criterion selected sample points that aid in defining the peak of the 2-LTR concentration. In the sample frequency plot in figure 6.2b, the vast majority of the sample points selected by this criterion are selected in the first 10 days of the trial, which further establishes that this criterion’s aim is to define the peak of the curve.

6.2.1.3 A-optimality Criterion

The next four sample solution that we will analyze is the A-optimality criterion based solution. To optimize the experiment design the A-optimality criterion seeks to minimize the trace of the inverse of the Fisher Information Matrix. While taking the trace is relatively simple mathematically, this method can be computationally difficult as taking the inverse of large matrixes can be cumbersome. This method also does not take into account any underlying relationships between the parameters [71, 73, 74].

The four sample schedule that maximizes the information content of the experiment subject to the FIM A-optimality criterion are samples at day 2, day 3, day 4 and day 12 as shown in figure 6.3a. Similar to the T-optimality criterion solution, the A-optimality criterion based solution elected several sample points that would help to define the peak of the curve. Instead of choosing all points near the peak this criterion selected one point out from the peak as the 2-LTR concentration was decreasing. From the sample frequency plot for this criterion in figure 6.3b, it is evident that it prefers to choose samples primarily around the peak, with additional samples chosen near day 12 and day 22.

6.2.1.4 E-optimality Criterion

The last of the FIM based optimality criterion evaluated for an optimal four sample solution was the E-optimality criterion. The E-optimality seeks to maximize
Figure 6.3: **Four point optimal A-optimality criterion with qPCR measurement noise** (a) four optimal points selected by the A-optimality criterion plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the A-optimality criterion.
the minimum eigenvalue of the Fisher Information Matrix. In doing so it minimizes
the maximum diameter of the five dimensional ellipsoid [71, 73, 74].

The four sample solution that maximizes the information content subject to the
E-optimality criterion were chosen at day 1, day 2, day 3, and day 44 as shown in figure
6.4a. This criterion elected points that define the slope of the 2-LTR concentration
increase and the peak as well as on point much further out that characterizes the steady
state value. Based on the sample frequency plot in figure 6.4b for this criterion, The
samples at day 1, day 2 and day 3 were much more critical than the other points. The
criterion seemed to prefer these three samples and one additional sample further out
in the experiment.

6.2.1.5 EKLD Optimal

Lastly, we will investigate the optimal four sample solution chosen by our
method of Expected Kullback-Leibler Divergence. This method is Bayesian and takes
into account the prior information regarding the parameters to try and maximize in-
formation gain by a measure of divergence between the prior and the posterior distri-
butions across the simulated patient space.

The four sample solution that was chosen by the EKLD method chose samples
at day 1, day 2, day 12 and day 31 as shown in figure 6.5a. This method chose two
points next to each other that would characterize the slope of the increase in the
concentration followed by two later points at the bottom of the curve. The sample
frequency plot in figure 6.5b for this method we show that the points at day 1 and
day 2 are very critical and then the method seeks points further out in the 15-25 day
range. Interestingly this method never elected samples at day 0, establishing an initial
condition for the experiment.

6.2.1.6 Result Comparison

The various methods all chose different solutions based on the criterion to be
optimized. The Hatano et al sample schedule chose samples at day 0, day 7, day 14,
Figure 6.4: Four point optimal E-optimality criterion with qPCR measurement noise (a) four optimal points selected by the E-optimality criterion plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the E-optimality criterion.
Figure 6.5: **Four point optimal EKLD schedule with qPCR measurement noise** (a) four optimal points selected by the EKLD method plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the EKLD method.
and day 56 which is illustrated in figure 6.6. They chose a sample at the beginning of
the trial; however, they did not choose any samples which would define the slope of the
increase in concentration or the peak as in several of the optimal solutions. Instead
the three remaining sample points are on the descent of the 2-LTR curve and at the
end of the sample window during steady state conditions.

![Optimal Sampling Schedule](image)

**Figure 6.6:** **Sampling Schedule used by Hatano et al.** Sample schedule used in the
experiment performed by Hatano et al. plotted on a 2-LTR concentration trajectory.

All of the optimality criterion use different metrics to measure information con-
tent, so in order to compare them all against one another we calculated the Expected
Kullback-Leibler Divergence for each optimal sample schedule. The results are shown
in the bar graph in figure 6.7. From the figure we observe that the EKLD solution
and the E-optimality criterion based solution produced the highest gains in informa-
tion. The two schedules are not significantly different from one another with respect
to information gain. Next the D-optimal and the A-optimal solution also yielded sim-
ilar gains in information; however, they performed significantly worse than the EKLD
and the E-optimality criterion methods. The T-optimality criterion method performed
significantly worse than all of the other optimal methods; however, it still provided a higher gain in information compared to the schedule used by Hatano et al.

The results are summarized in tabular form in Table 6.1. The EKLD base schedule provided the highest EKLD of 3.16. This was 0.61 more bits of information gained than the Hatano schedule. This corresponds to 1.52 times more information when using the EKLD based schedule.

6.2.2 Droplet Digital Polymerase Chain Reaction Design

Next we will investigate how the optimal sample schedules vary when using the more accurate ddPCR measurement assay instead of a qPCR assay. We again aim to find the optimal four sample solution over a time span of eight weeks for the four Fisher Information Matrix based criterion and the EKLD method. We will then compare EKLDs for the resulting optimal schedules against the experimental schedule used by Hatano et al. assuming measurement noise consistent with a ddPCR assay.
Table 6.1: Four Point qPCR Optimized Sample Schedules

Optimal four point sample schedules optimized for a qPCR measurement assay for each optimality criterion schedule, the EKLD method schedule and the experimental schedule as well as their associated EKLD.

<table>
<thead>
<tr>
<th></th>
<th>Samples</th>
<th></th>
<th></th>
<th></th>
<th>EKLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-optimal</td>
<td>0</td>
<td>1</td>
<td>7</td>
<td>22</td>
<td>3.04</td>
</tr>
<tr>
<td>T-optimal</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>2.82</td>
</tr>
<tr>
<td>A-optimal</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>12</td>
<td>2.69</td>
</tr>
<tr>
<td>E-optimal</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>44</td>
<td>2.90</td>
</tr>
<tr>
<td>EKLD</td>
<td>1</td>
<td>2</td>
<td>12</td>
<td>31</td>
<td>3.16</td>
</tr>
<tr>
<td>Hatano</td>
<td>0</td>
<td>7</td>
<td>14</td>
<td>56</td>
<td>2.43</td>
</tr>
</tbody>
</table>

6.2.2.1 D-optimality Criterion

The first method that we evaluate for an optimal four sample solution with ddPCR measurement noise is the D-optimality criterion.

The four sample ddPCR measurement solution that maximizes the information content of the D-optimality criterion takes samples at day 0, day 1, day 4, and day 20 as shown in figure 6.8a. This sample schedule is similar to the D-optimality criterion schedule that was chosen for the qPCR analysis, which was samples at day 0, day 1, day 7, and day 22. They differ only in the last two sample points. Based on sample frequency for this criterion it appears as though the sample at day 4 having been selected for in more than 70 percent of the top schedules. The final sample however seem to be less critical.

6.2.2.2 T-optimality Criterion

Next we investigate the optimal four sample solution for the T-optimality criterion assuming ddPCR measurement noise.

The four sample T-optimal criterion based solution for this analysis elected to take samples at day 2, day 3, day 4, and day 5 as shown in 6.9a. This solution is identical to the solution for the case assuming qPCR measurement noise. Similar as
Figure 6.8: **Four point optimal D-optimality criterion with ddPCR measurement noise** (a) four optimal points selected by the D-optimality criterion plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the D-optimality criterion.
Figure 6.9: **Four point optimal T-optimality criterion with ddPCR measurement noise** (a) four optimal points selected by the T-optimality criterion plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the T-optimality criterion.
well to the qPCR case, the sampling frequency graph for ddPCR measurement noise, figure 6.9b, shows that the points that define the peak are heavily selected. As the samples get further away from the peak the frequency with which they are selected quickly declines.

6.2.2.3 A-optimality Criterion

Next we investigate the optimal four sample solution for the A-optimality criterion assuming ddPCR measurement noise.

The four sample schedule that provided the highest gain in information for the A-optimality criterion assuming ddPCR measurement noise takes samples at day 3, day 6, day 47, and day 53 as shown in figure 6.10a. This solution takes two samples in the first week of the experiment similar to the A-optimality criterion schedule from the qPCR assay case, which took three samples in the first week at day 2, day 3, and day 4. The other two samples in the ddPCR case occurred much later and were heavily selected for as evident in the sample frequency plot 6.10b.

6.2.2.4 E-optimality Criterion

Next we investigate the optimal four sample solution for the E-optimality criterion assuming ddPCR measurement noise.

The four point solution that maximized the information content for the E-optimality criterion with ddPCR measurement noise takes samples at day 0, day 1, day 4, and day 34 as shown in figure 6.11a. Based on the sample frequency graph in figure 6.11b day 0 was selected for more than any other point. This is interesting because day 0 was not selected in the qPCR for this criterion. The increase in measurement accuracy drives the need to have a sample at the start of the trial based on optimizing on the E-optimality criterion.

6.2.2.5 EKLD Optimal

Lastly, we investigate the optimal four sample solution for the EKLD method assuming ddPCR measurement noise.
Figure 6.10: **Four point optimal A-optimality criterion with ddPCR measurement noise** (a) four optimal points selected by the A-optimality criterion plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the A-optimality criterion.
Figure 6.11: Four point optimal E-optimality criterion with ddPCR measurement noise (a) four optimal points selected by the E-optimality criterion plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the E-optimality criterion.
Figure 6.12: **Four point optimal EKLD sample schedule with ddPCR measurement noise** (a) four optimal points selected by the EKLD method plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the EKLD method.
6.2.2.6 Result Comparison

In order to compare the schedules against one another we again calculated the Expected Kullback-Leibler Divergence for all of them. The result show that the EKLD based schedule provides the highest gain in information content. For the qPCR analysis the EKLD based schedule and the E-optimality criterion schedule performed equally as well. When ddPCR is used the E-optimal solution provides significantly less information than the EKLD based method. The T-optimality criterion was the worst performing out of all of the optimal solutions for qPCR measurement noise; however, with the increase measurement accuracy associated with the ddPCR assay T-optimality criterion solution performs as well as the D-optimality criterion and E-optimality criterion based methods. This is due to the fact the the ddPCR accuracy is better able to define the peak, where all of the T-optimality criterion schedule points are located. All of the schedules again out performed the schedules used in the experiment done by Hatano et al. and this time by a larger measure. Dispite the fact that Hatano et al. used a ddPCR assay in their experiment, they missed out on a large amount of information due to using an inefficient sample schedule.

The results are shown in tabular form in table 6.2. From the table we see that the difference in information gain between the EKLD optimal sample schedule and the schedule used in the experiment done by Hatano et al. is 2.28 bits which means that the experiment would result in 4.84 times more information if the EKLD based schedule was used. From the table we also note that the EKLD schedule out performed the T-optimal criterion schedule, the second best schedule, by 0.36 bits. This means that the EKLD schedule would provide 1.28 times more information that the best performing FIM based method.

6.2.3 Schedule Robustness

Thus far we have shown that the EKLD based design provides the greatest gain in information when used in experiments for which it is designed. In the previous examples the sample schedules have been optimized for a certain measurement assay
Figure 6.13: **EKLD comparison of all sample schedules with ddPCR noise.** EKLD is calculated for all of the optimal schedules as well as the Hatano et al. experiment schedule and compared.

Table 6.2: **Four Point ddPCR Optimized Sample Schedules** Optimal four point sample schedules optimized for a ddPCR measurement assay for each optimality criterion schedule, the EKLD method schedule and the experimental schedule as well as their associated EKLD.

<table>
<thead>
<tr>
<th>Samples</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>EKLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-optimal</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>20</td>
<td>5.67</td>
</tr>
<tr>
<td>T-optimal</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>5.71</td>
</tr>
<tr>
<td>A-optimal</td>
<td>3</td>
<td>6</td>
<td>47</td>
<td>53</td>
<td>5.29</td>
</tr>
<tr>
<td>E-optimal</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>34</td>
<td>5.56</td>
</tr>
<tr>
<td>EKLD</td>
<td>1</td>
<td>2</td>
<td>15</td>
<td>20</td>
<td>6.07</td>
</tr>
<tr>
<td>Hatano</td>
<td>0</td>
<td>7</td>
<td>14</td>
<td>56</td>
<td>3.79</td>
</tr>
</tbody>
</table>

and then analyzed assuming that assay was used. In this section we investigate if the EKLD based optimal design remains the best performing design even if a different assay is used.

Figure 6.14a is a comparison of the five optimal sampling schedules and the
Figure 6.14: **Comparing Results Using qPCR Assay vs Using ddPCR Assay for an Experiment Optimized for a qPCR Assay.** (a) EKLD is calculated for all qPCR optimized schedules and Hatano et al. schedule assuming qPCR measurement noise. (b) EKLD is calculated for all qPCR optimized schedules and Hatano et al. schedule assuming ddPCR measurement noise.
Hatano et al. sampling schedule. The optimal schedules have all been optimized to be used in experiments with a qPCR assay and exhibit the the information gain shown in the figure when they are analyzed with measurement noise consistent with a qPCR assay. Figure 6.14b, on the other hand, shows the same schedules optimized for a qPCR assay but instead they are analyzed with noise consistent with a ddPCR measurement assay.

The first thing to note is that there is a large gain in just from switching from using a qPCR assay to ddPCR assay as expected. Each schedule gains about 3 bits of information when using a ddPCR assay vs a qPCR assay. The EKLD based method schedule, proves to be robust to the change in assay, and continues to be one of the best performing methods. The A-optimality criterion based method, the E-optimality criterion based method and the EKLD method all yield similar gains in information. Interestingly the A-optimality criterion based schedule performed better relative to the other methods when analyzed with a ddPCR assay and the D-optimality criterion based schedule performed worse relative to the other schedules. The D-optimality criterion based schedule is the worst performing schedule when optimized for a qPCR assay but a ddPCR assay is used in the experiment. The schedule used in he experiment done by Hatano et al remains the worst performing schedule in both cases; however, the more uncertain the measurement assay the less of a difference there is between the Hatano schedule and the optimal samples.

Figure 6.15a is again a comparison of the five optimal sampling schedules and the Hatano et al. sampling schedule. In this case however, the optimal schedules have all been optimized to be used in experiments with a ddPCR assay, and as a result they demonstrate the information gained shown in the figure when they are actually analyzed with measurement noise consistent with a ddPCR assay. The figure beneath it, figure 6.15b, shows the same schedules optimized for a ddPCR assay but instead they are analyzed with noise consistent with a qPCR measurement assay.

As expected there is a large reduction in information gain when a qPCR assay is used instead of a ddPCR assay. The total difference in information is again around
Figure 6.15: Comparing Results Using ddPCR Assay vs Using qPCR Assay for an Experiment Optimized for a ddPCR Assay. (a) EKLD is calculated for all ddPCR optimized schedules and Hatano et al. schedule assuming ddPCR measurement noise. (b) EKLD is calculated for all ddPCR optimized schedules and Hatano et al. schedule assuming qPCR measurement noise.
three bits for each schedule. Again we see that the EKLD base schedule is robust to the change in measurement assay, and remains the best performing schedule, although not significantly better than the D-optimality criterion base schedule and the E-optimality criterion based schedule. The T-optimality criterion schedule which was one of the best performing schedules is now among the worst performing, only yielding a slightly greater gain in information than the Hatano et al. experiment schedule, which again was the worse performing schedule.

6.3 Six Sample Optimization

In this section we introduce two more points to our analysis. We optimize our sample schedule to find the best six sample schedule over a period of eight weeks. In doing so we can compare the optimal schedules to the actual used in the experiment done by Buzon et al, which also took six samples.

6.3.1 Real-Time Polymerase Chain Reaction Design

We will again start by optimizing the experiment for six samples using a qPCR assay. Optimal schedules are found for the FIM based methods, D-optimality criterion, T-optimality criterion, A-optimality criterion, and E-optimality criterion as well as for our EKLD method. The resulting optimal schedules are shown plotted on an indiscriminate 2-LTR concentration curve. We will then analyze the sample frequency of the top performing schedules in the family of solutions from the genetic algorithm.

6.3.1.1 D-optimality Criterion

First we investigate the optimal six sample solution for the D-optimality criterion assuming qPCR measurement noise.

The six point solution that optimizes the information content for the D-optimality criterion selects samples at day 0, day 1, day 7, day 11, day 20, and day 46 as shown in figure 6.16a. The first sample points, day 0, day1, and day 7 are the same for the four point qPCR schedule. These points prove to be vital in schedule design optimization for the D-optimality criterion based method. From frequency plot shown in figure
Figure 6.16: **Six point optimal D-optimality criterion with qPCR measurement noise.** (a) Four optimal points selected by the D-optimality criterion plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the D-optimality criterion.
6.16b we see the first point at day 1 was again selected for in every schedule as was the case for the four sample qPCR optimization for this criteria. The sample at day 20 proved to be vital as it was selected in more than 90 percent of the schedules. The remaining sample points all are selected for in at least 40 percent of the schedules with the exception the sample at day 11.

6.3.1.2 T-optimality Criterion

Next we investigate the optimal six sample solution for the T-optimality criterion assuming qPCR measurement noise.

The six point solution that optimizes the information content for the T-optimality criterion selects consecutive samples at day 2, day 3, day 4, day 5, day 6, and day 7 as shown in figure 6.17a. The first four samples of this schedule are identical to those of the T-optimality criterion base optimal for both the ddPCR and qPCR four sample schedules. The six samples are again heavily selected for as demonstrated in the frequency plot in figure 6.17b. The sample at day 1 is selected for in about 50 percent of the schedules, and looks to trade off with the sample at day 7, which is also sampled in near 50 percent of the schedules. After the 7th sample, the frequency with with samples are selected for declines dramatically.

6.3.1.3 A-optimality Criterion

Next we investigate the optimal six sample solution for the A-optimality criterion assuming qPCR measurement noise.

The six point solution that optimizes the information content for the A-optimality criterion selects samples at day 0, day 1, day 7, day 9, day 21, and day 52 as shown in figure 6.18a. Interestingly the first three samples are identical to those of the D-optimality criterion based solution. From the sample frequency plot in 6.18b first two samples are selected in 100 percent of the schedules. Following those two samples the frequency plot shows evidence that the criterion selects points in the regions around day 7, around day 21 and day 52. Similar to the E-optimality criterion, five out of the
Figure 6.17: 

(a) Best Six point T-optimality Criterion Schedule for qPCR Measurements

(b) T-optimality Sample Day Frequency Among Top 100 in Family of Solutions

**Figure 6.17:** Six point optimal T-optimality criterion with qPCR measurement noise. (a) four optimal points selected by the T-optimality criterion plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the T-optimality criterion.
Figure 6.18: **Six point optimal A-optimality criterion with qPCR measurement noise.** (a) Four optimal points selected by the A-optimality criterion plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the A-optimality criterion.
six sample points were selected four with in at least 50 percent of the schedules. The remaining point seems to be somewhat ancillary in both cases and would tend to roam throughout the sampling space around day 7.

6.3.1.4 E-optimality Criterion

Concluding the FIM based methods we next investigate the optimal six sample solution for the E-optimality criterion assuming qPCR measurement noise.

The six point solution that optimizes the information content for the E-optimality criterion selects samples at day 0, day 1, day 6, day 8, day 21, and day 55 as shown in Figure 6.19a. The samples at day 0 and day 1 are again selected in every schedule as illustrated in the sample frequency plot in Figure 6.19b. Similar to the A-optimality criterion and D-optimality based criterion the next two points are chosen in the range of points around day 8, which corresponds to the descent of the 2-LTR concentration. The fifth point in the A, D and E-optimality criterion seems to be selected in the region at the bottom near the bottom of the 2-LTR curve as the dynamics start to settle towards a steady state. The last point roams around many time points in the steady state region of the dynamics.

6.3.1.5 EKLD Optimal

Lastly, we investigate the optimal six sample solution for the EKLD based method assuming qPCR measurement noise.

The six point solution that optimizes the information content for the EKLD based method selects samples at day 1, day 2, day 3, day 14, day 21, and day 51 as shown in figure 6.20a. From the sampling frequency plot in figure 6.20b we observe that the samples taken at day 1 and day 2 are selected for in 100 percent of the schedules. The sample at day three is selected in 99 percent of the sample schedules. The sample points selected at day 14 and and 21 are selected for over 60 percent and 40 percent of the schedules respective. These time points serve to classify the return of the 2-LTR dynamics back to steady state. The remaining time is again weakly selected, similar
Figure 6.19: **Six point optimal E-optimality criterion with qPCR measurement noise.** (a) Four optimal points selected by the E-optimality criterion plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the E-optimality criterion.
Figure 6.20: **Six point optimal EKLD sample schedule with qPCR measurement noise.** (a) four optimal points selected by the EKLD method plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the EKLD Method.
to the A, D, and E optimality criterion solutions, and serves to classify the steady state 2-LTR concentration.

6.3.1.6 Result Comparison

The various methods all chose solutions which varied slightly based on the criterion to be optimized. The Buzon et al. sample schedule chose samples at day 0, day 14, day 28, day 84, day 168 and day 336 which is illustrated in figure 6.6. They chose to take a sample at the beginning of the trial which captures the initial 2-LTR Concentration. There next sample is not until day 14 which would only capture the dynamics of the 2-LTR curve descending from its peak. The third point would characterized the return to steady state. The remaining three points should not prove as useful as the concentration will have almost certainly reach steady state by then and any differences in measured 2-LTR concentration at the those points would be due to random noise.

Figure 6.21: Buzon et al Sample Schedule. Sampling Schedule used by Buzon et al. Note: six samples were taken in the Buzon study however, only three samples taken during the 8 week window.

Figure 6.22 compares the expected information content of all of the six point
schedules optimized for a qPCR assay as well as the Buzon schedule. The EKLD schedule presented the greatest gain in information out of all of the schedules. It significantly out performed all of the other schedules. The D-optimality, A-optimality, and E-optimality criterion based schedules all yield a similar amount of information. The T-optimality schedule produced the least gain in information out of all of the optimal schedules and the schedule used in the Buzon et al. study yielded significantly less of a gain in information that all of the other schedules.

Figure 6.22: **EKLD comparison of all sample schedules with qPCR noise.** EKLD is calculated for all of the optimal schedules as well as the Buzon et al. experiment schedule and compared.

The results are summarized in tabular form in table 6.3. The EKLD based schedule provided the highest EKLD of 3.86 bits. The schedule used by Buzon et al. only produced a 2.49 bit gain information. In their study Buzon et al used a qPCR assay for their measurements. Based on this analysis if they had used the EKLD based schedule their data would have yielded 1.37 more bits or 2.58 times more information about the model parameters.
Table 6.3: Six Point qPCR Optimized Sample Schedules Optimal six point sample schedules optimized for a qPCR measurement assay for each optimality criterion schedule, the EKLD method schedule and the experimental schedule as well as their associated EKLD.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>EKLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-optimal</td>
<td>0</td>
<td>1</td>
<td>7</td>
<td>9</td>
<td>21</td>
<td>52</td>
<td>3.52</td>
</tr>
<tr>
<td>T-optimal</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>3.19</td>
</tr>
<tr>
<td>A-optimal</td>
<td>0</td>
<td>1</td>
<td>7</td>
<td>9</td>
<td>21</td>
<td>52</td>
<td>3.48</td>
</tr>
<tr>
<td>E-optimal</td>
<td>0</td>
<td>1</td>
<td>6</td>
<td>8</td>
<td>21</td>
<td>55</td>
<td>3.54</td>
</tr>
<tr>
<td>EKLD</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>14</td>
<td>21</td>
<td>51</td>
<td>3.85</td>
</tr>
<tr>
<td>Buzon</td>
<td>0</td>
<td>14</td>
<td>28</td>
<td>84</td>
<td>168</td>
<td>336</td>
<td>2.49</td>
</tr>
</tbody>
</table>

6.3.2 Droplet Digital Polymerase Chain Reaction Design

Next we will investigate how the optimal sample schedules vary when using the more accurate ddPCR measurement assay instead of a qPCR assay. We again aim to find the optimal six sample solution over a time span of eight weeks for the four Fisher Information Matrix based criterion and the EKLD method. We will then compare EKLDs for the resulting schedules against the EKLD of schedule used by Buzon et al. assuming measurement noise consistent with a ddPCR assay.

6.3.2.1 D-optimality Criterion

First we investigate the optimal six sample solution for the D-optimality criterion assuming ddPCR measurement noise.

The six point solution that optimizes the information content for the E-optimality criterion selects samples at day 0, day 1, day 5, day 12, day 20, and day 55 as shown in Figure 6.23a. The four point solution for the D-optimality criterion chose samples at day 0, day 1, day 4, and day 20. The six sample solution is very similar except the sample at day 4 has shifted to day 5 and there is additional point added in at day 12 to characterized the return to steady state. There is also an additional point at the end of the sampling window that would help to characterize the steady state value. Based
Figure 6.23: Six point optimal D-optimality criterion with ddPCR measurement noise. (a) four optimal points selected by the D-optimality criterion plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the D-optimality criterion.
on the sample frequency plot in Figure 6.23b the sample that was added at time day 12 was not highly selected for where all of the other points were selected in at least 40 percent of the schedules. Again we see that the points at 0 and 1 are selected for in 100 percent of the patients.

6.3.2.2 T-optimality Criterion

Next we investigate the optimal six sample solution for the T-optimality criterion assuming ddPCR measurement noise.

The six point solution that optimizes the information content for the E-optimality criterion selects samples at day 2, day 3, day 4, day 5, day 6, and day 7 as shown in figure 6.24a. Similar to the four point sample cases, the T-optimality based schedule is identical when optimize for a qPCR measurement assay as it is when optimized for a ddPCR measurement assay. From the frequency sample plot shown in figure 6.24b, we observe a similar distribution of sample points as in the ddPCR case with the majority of the points selected at the peak of the curve. After day 7 the frequency at which the points are selected for drops off drastically.

6.3.2.3 A-optimality Criterion

Next we investigate the optimal six sample solution for the A-optimality criterion assuming ddPCR measurement noise.

The six point solution that optimizes the information content for the A-optimality criterion selects samples at day 0, day 1, day 5, day 6, day 19, and day 50 as shown in Figure 6.25a. Once again the A-optimality criterion based method chose a similar schedule to that of the D-optimality based criterion. The sample frequency plot in Figure 6.25b shows that the samples at day 0 and 1 are again selected for in 100 percent of the schedules. There is very little variation in the first 5 samples of this schedule as they are selected for in 50 percent of more of the schedules. The last point in the sample schedule however, seems to roam around the end of the sampling window.
Figure 6.24: Six point optimal T-optimality criterion with ddPCR measurement noise. (a) four optimal points selected by the T-optimality criterion plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the T-optimality criterion.
Figure 6.25: Six point optimal A-optimality criterion with ddPCR measurement noise. (a) four optimal points selected by the A-optimality criterion plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the A-optimality criterion.
6.3.2.4 E-optimality Criterion

Next we investigate the optimal six sample solution for the E-optimality criterion assuming ddPCR measurement noise.

The six point solution that optimizes the information content for the E-optimality criterion selects samples at day 0, day 1, day 5, day 6, day 19, and day 50 as shown in Figure 6.26a. This sample schedule is also similar to the D-optimality criterion and the A-optimality criterion based schedules. In all of the schedules, samples day 0 and day 1 are selected for in 100 percent of the schedules. From the sample frequency plot for this criterion in Figure 6.26a we observe very little variation in the first 5 schedules and the last point roams around the end of the sampling window in search of a point that will characterize the steady state.

6.3.2.5 EKLD Optimal

Lastly we investigate the optimal six sample solution for the EKLD method assuming ddPCR measurement noise.

The six point solution that optimizes the information content for the EKLD method selects samples at day 1, day 2, day 3, day 4, day 15, and day 42 as shown in Figure 6.27a. The EKLD based schedule was similar to the EKLD schedule optimized for qPCR which takes samples at day 1, day 2 day, 3, day 14, day 21, and day 51. With the ddPCR measurement assay the optimal schedule was one that preferred to take samples that defined the peak followed by a sample at the bottom of the curve and one at at towards the end of the trial.

6.3.2.6 Result Comparison

In this section we compare the schedules expected information content for all of the schedules. From the bar graph in figure 6.28 it is clear that the EKLD schedule is the best performing schedule. The four Fisher based optimality methods all yielded similar gains in information and the schedule used in the Buzon et al. experiment again produced the lowest expected gain in information from the experiment.
Figure 6.26: **Six point optimal E-optimality criterion with ddPCR measurement noise.** (a) Four optimal points selected by the E-optimality criterion plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the E-optimality criterion.
Figure 6.27: Six point optimal EKLD sample schedule with ddPCR measurement noise. (a) Four optimal points selected by the EKLD method plotted on a 2-LTR concentration trajectory. (b) Genetic algorithm frequency plot of the top 100 schedules. Solid dark blue bars represent the frequency with which each sample day was selected among the top performing schedules. Vertical red lines represent optimal sample days selected using the EKLD method.
Figure 6.28: EKLD comparison of all sample schedules with ddPCR noise. EKLD is calculated for all of the optimal schedules as well as the Buzon et al. experiment schedule and compared.

The results are summarized in tabular form in table 6.4. The EKLD based schedule provided the highest EKLD of 6.85 bits. The schedule used by Buzon et al. only produced a 3.4 bit gain in information. This corresponds to a difference of 3.45 bits or 10.93 times more information gained in the experiment when using the EKLD based schedule vs the Buzon et al. schedule in an experiment with a ddPCR measurement assay.

6.3.3 Schedule Robustness

We have shown again that the EKLD based design provides the greatest gain in information when used with the six sample experiments for which it is designed. In the previous examples the sample schedules have been optimized for a certain measurement assay and then analyzed assuming that assay was used. In this section we investigate if the EKLD based optimal design remains the best performing design even if a different assay is used.

Figure 6.29a is a comparison of the five optimal sampling schedules and the
Figure 6.29: Comparing Results Using qPCR Assay vs Using ddPCR Assay for an Experiment Optimized for a qPCR Assay. (a) EKLD is calculated for all qPCR optimized schedules and Buzon et al. schedule assuming qPCR measurement noise. (b) EKLD is calculated for all qPCR optimized schedules and Buzon et al. schedule assuming ddPCR measurement noise.
Table 6.4: **Six Point ddPCR Optimized Sample Schedules** Optimal four point sample schedules optimized for a ddPCR measurement assay for each optimality criterion schedule, the EKLD method schedule and the experimental schedule as well as their associated EKLD.

<table>
<thead>
<tr>
<th></th>
<th>Samples</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>EKLD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>D-optimal</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>12</td>
<td>20</td>
<td>55</td>
<td>6.85</td>
</tr>
<tr>
<td>T-optimal</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>6.26</td>
</tr>
<tr>
<td>A-optimal</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>6</td>
<td>19</td>
<td>50</td>
<td>6.27</td>
</tr>
<tr>
<td>E-optimal</td>
<td>0</td>
<td>1</td>
<td>6</td>
<td>19</td>
<td>21</td>
<td>53</td>
<td>6.16</td>
</tr>
<tr>
<td>EKLD</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>15</td>
<td>42</td>
<td>6.85</td>
</tr>
<tr>
<td>Buzon</td>
<td>0</td>
<td>14</td>
<td>28</td>
<td>84</td>
<td>168</td>
<td>336</td>
<td>3.40</td>
</tr>
</tbody>
</table>

Buzon et al. sampling schedule. The optimal schedules have all been optimized to be used in experiments with a qPCR assay and exhibit the the information gain shown in the figure when they are analyzed with measurement noise consistent with a qPCR assay. Figure 6.29b, on the other hand, shows the same schedules optimized for a qPCR assay but instead they are analyzed with noise consistent with a ddPCR measurement assay.

The first thing to note is that there is a large gain in information just from switching from using a qPCR assay to using ddPCR assay as expected. Each schedule again gains about three bits of information when using a ddPCR assay vs a qPCR assay. The EKLD based method schedule proves to be robust to the change in assay, and continues to be the best performing method. The T-optimality criterion based method, which was the worst performing schedule when analyzed with a qPCR assay, increases to become the best performing out of the FIM based optimization methods. The other three, A-optimality, D-optimality, and E-optmality criterion based methods all produce more information when analyzed with ddPCR measurement noise; however, they continue to perform similarly with respect to one another. The schedule used in the experiment done by Buzon et al remains the worst performing schedule in both cases.
Figure 6.30a is again a comparison of the five optimal sampling schedules and the Buzon et al. sampling schedule. In this case however, the optimal schedules have all been optimized to be used in experiments with a ddPCR assay, and as a result they demonstrate the information gain shown in the figure when they are actually analyzed with measurement noise consistent with a ddPCR assay. The figure beneath it, Figure 6.30b, shows the same schedules optimized for a ddPCR assay but instead they are analyzed with noise consistent with a qPCR measurement assay.

As expected there is a large reduction in information gain when a qPCR assay is used instead of a ddPCR assay. The total difference in information is again around three bits for each schedule. Again we see that the EKLD based schedule is robust to the change in measurement assay, and remains the best performing schedule. The D-optimality, T-optimality, and A-optimality based methods perform similarly when analyzed with the ddPCR assay, but when the qPCR assay is used the T-optimality criterion performs worse than the other two. The schedule used in the experiment done by Buzon et al remains the worst performing schedule in both cases again.

6.4 Sample Days vs. Information Gain

From our analysis we observe that the fewer the amount of samples that are taken in the experiment, the larger the difference in experimental designs among the different criterion. As more samples are added in the experiment, it will yield more information as there are more samples from which to characterize the parameters. The optimal solutions from the various optimality criterion begin to look similar as the number of sample points taken during the experiment is increased, which means that optimal experiment design is essential for experiments with sparse samples. In this section we examine the relationship between sample points and expected information gain for experiments done with both ddPCR measurements and qPCR measurements.

Figure 6.31 shows relationship between the number of sample points in the experiment and the expected gain in information for experiments done with qPCR measurements. Each schedule has been optimized based on the EKLD method for that
Figure 6.30: Comparing Results Using ddPCR Assay vs Using qPCR Assay for an Experiment Optimized for a ddPCR Assay. (a) EKLD is calculated for all ddPCR optimized schedules and Buzon et al. schedule assuming ddPCR measurement noise. (b) EKLD is calculated for all ddPCR optimized schedules and Buzon et al. schedule assuming qPCR measurement noise.
number of points. We can see that there is a smaller and smaller gain in information as more points are added to the experiment. From the figure, it can also be seen that the three sample EKLD optimal sampling schedule would yield a higher expected gain in information than the six sample schedule used in the experiment done by Buzon et al.

Figure 6.32 shows the relationship between the number of sample points in the experiment and the expected gain in information for experiments done with ddPCR measurements. We again see that the information steadily increases as we add more sample points to the experiment and the EKLD optimized 2 sample schedule will yield more information than 4 sample method in the experiment done by Hatano et al.
Figure 6.32: **Number of Samples vs EKLD** Number of samples vs EKLD with ddPCR noise for EKLD optimal schedules and Hatano et al experiment schedule.
Chapter 7
SAMPLE VOLUME OPTIMIZATION

7.1 Introduction

Another factor that can be used in experiment design to maximize the information gain is the sample volume. In general, higher sample volumes produce more accurate measurements. The accuracy of both the ddPCR measurement error and the qPCR measurement error are functions of the volume of blood sampled [51, 91]. In this section we allow the volume of the sample to vary to perform a constrained optimization problem. We vary the volume of each sample subject to the constraint that the total volume in the experiment not exceed a predetermined amount.

7.2 Volume Optimal Four Sample Schedule

For the four sample constrained optimization case, we constrain the total sample volume to a maximum of 80 ml over the four samples. Each sample can vary from 5 ml to 30 ml in increments of 5 ml. Uniform volume sample schedules of 20ml per sample were used as the control for comparison. We use the EKLD optimal four sample schedule and the Hatano et al. sample schedule in our analyses.

7.2.1 Real-Time Polymerase Chain Reaction Design

We first perform the constrained optimization assuming qPCR measurements for the experiment. The optimal sample volume distributions for the two schedules was 25ml, 20ml, 25ml, and 10ml for the four samples in the Hatano schedule and 25ml, 10ml, 20ml, and 25 ml for the four samples in the optimal schedule. The resulting expected information gain is shown in figure 7.1.
From the results in figure 7.1, we see that there is not a significant gain in information when the volume is allowed to vary for the Hatano et al. sample schedule; however, for the EKLD optimal sample schedule, there is a significant gain in information content over the uniform volume schedule. This sample schedule elected to take higher accuracy measurements at the beginning and the end of the experiment.

### 7.2.2 Droplet Digital Polymerase Chain Reaction Design

Next we performed the constrained optimization assuming ddPCR measurements for the experiment. The optimal sample volume distributions for the two schedules was 10ml, 25ml, 20ml, and 25ml for the four samples in the Hatano schedule and 25ml, 25ml, 5ml, and 25 ml for the four samples in the optimal schedule. The Hatano schedule has its first sample at time zero, this volume distribution sacrifices the accuracy of that measurement for more accuracy later in the experiment. The optimal schedule on the other hand sacrifices the accuracy of a measurement at the middle of the experiment for more accuracy in the remaining samples.
Figure 7.2: **Constrained optimization for 4 samples with ddPCR noise.** Sample volume optimization for four point optimal schedule and Hatano et al. experiment schedule. EKLDs shown for volume optimized schedules and uniform volume schedules.

From the results in figure 7.2, we see that there is slightly more information gained in the experiment for the variable volume Hatano schedule vs the uniform volume case. The EKLD based optimal schedule again provides significantly more information than the uniform volume sample schedule

### 7.3 Volume Optimal Six Sample Schedule

We next performed the constrained optimization for the six sample schedules. We again used the EKLD optimal sample schedules for each respective assay and the sample schedule used in the Buzon et al experiment. Samples in this study were allowed to vary from 10 to 40 ml under the constraint that the total sample volume could not exceed 120 ml, which gives 20ml samples for the uniform volume schedules. As a control we again used uniform volume EKLD optimal and Buzon et al schedules for comparison.
7.3.1 Real-Time Polymerase Chain Reaction Design

We first perform the constrained optimization assuming qPCR measurements for the experiment. The optimal sample volume distributions for the two schedules was 30ml, 30ml, 20ml, 10ml, 10ml and 20 ml for the six samples in the Buzon et al. schedule and 30ml, 20ml, 10ml, 10ml, 30ml and 20ml for the six samples in the optimal schedule.

![6 Point Volume Optimization (qPCR)](image)

Figure 7.3: **Constrained optimization for 6 samples with qPCR noise.** Sample volume optimization for four point optimal schedule and Buzon et al. experiment schedule. EKLDs shown for volume optimized schedules and uniform volume schedules.

From the results in Figure 7.3 we see significant gains for both of the variable volume schedules over their uniform volume counterpart. One thing to note about the Buzon variable volume optimization is that it chose to sacrifice the accuracy of two of the three, last points, for more accuracy during the point in the beginning of the schedule where the 2-LTR dynamics were more informative.

7.3.2 Droplet Digital Polymerase Chain Reaction Design

Lastly we performed the constrained optimization assuming ddPCR measures. The optimal sample volume distributions for the two schedules for this case was 10 ml,
30ml, 30ml, 20ml, 20ml and 10 ml for the six samples in the Buzon schedule and 30ml, 20ml, 20ml, 20ml, 10ml and 20ml for the six samples in the optimal schedule.

Figure 7.4: Constrained optimization for 6 samples with ddPCR noise. Sample volume optimization for four point optimal schedule and Buzon et al. experiment schedule. EKLDs shown for volume optimized schedules and uniform volume schedules.

From the results in Figure 7.3 we notice a slight gain in information content for the Buzon schedule and a larger gain in information content for the EKLD optimal schedule. Interestingly, the Buzon schedule opted to forgo the accuracy of the day 0 measurement for more accuracy in other measurements again as it did in the ddPCR analysis for the four point schedule.
Chapter 8

CONCLUSIONS

There have been tremendous advances in the treatment of individuals infected with HIV since the early days of research into the disease. Antiretroviral drugs, in the form of Highly Active Antiretroviral Therapy (HAART), initially prescribed as monotherapy, were able to reduce and control the virus for a period of time until eventually the virus would develop a resistance to the drug and the viral load would rebound [10, 13, 21, 34, 39, 67]. The emergence of viral Drug Resistance Mutations (DRMs) proved to be a critical hindrance which would need to be overcome in order for a cure to be developed. This hurdle was overcome with the development and the prescription of Combined Antiretroviral Therapy (cART). By administering a treatment in which multiple antiretroviral drugs are given, there is a much lower chance that the virus will develop a mutation that enables it to evade the effect all of the drugs [2, 24, 85].

Although the use of cART is able to drastically reduce the development of DRMs, it is not able to completely cure the disease, due to an HIV reservoir that maintains an active viral infection in the body [2, 24, 85]. The next stumbling block to conquer in the effort towards a cure to HIV infection entails the effective treatment and elimination of these reservoirs. HIV reservoirs are comprised of resting infected memory CD4+ T cells which remain dormant until they are stimulated antigenically to become transcriptionally active [15]. While in the dormant state, these cells are not susceptible to the drugs rendering cART and ineffective treatment. HIV reservoirs are also consisting of infected cells in sanctuary sites, where cART drugs exist in low concentrations. Due to the poor antiretroviral drug dissemination in these sites, low levels of ongoing replication can occur which maintain an active infection during treatment [8, 13, 27, 30].
In this dissertation we focus on developing an accurate method to measure the amount of ongoing replication and optimizing the design of clinical trials which will incorporate this technique. Previous research has intensified cART treatment with an integrase inhibitor in order to stimulate the production of circularized HIV DNA with two long terminal repeat ends [9, 66]. These circles are aptly named 2-LTR circles and serve as a marker of ongoing replication [37, 38, 9, 40, 41]. Inferences about the level of ongoing replication present in the patient can be made by identifying the trajectory and dynamics of the 2-LTR curve post intensification [1]. Using a model of 2-LTR production we can determine the dynamics by analyzing the model parameters. Because we seek to identify model parameters, the best clinical trial design is one that will yield the most information about the parameters. To find the best design we designed a method based on maximizing information content by Expected Kullback-Leibler divergence (EKLD) [50, 61].

Our tool for optimizing experiments to quantify the level of on-going replication will aid in the treatment of HIV as it will allow clinicians to design patient specific treatment to maximize antiretroviral efficacy. The formation of the HIV reservoir is the major roadblock preventing the development of a cure for the disease [2, 17, 33, 34, 43, 65]. A sterilizing cure, which entails the full eradication of the virus, would require that reservoir be completely eliminated. The development of a functional cure, which would allow the immune system to control the virus, is much more likely in the short term. In order to control the virus, knowledge about the level of ongoing replication, is of the utmost importance, as the the amount of on-going replication must also be controlled for a functional cure.

8.1 EKLD Method and Experimental Design Criteria

Our novel method employs the use of a two state ordinary differential equation model, developed by Luo et al., which constitutes the dynamics of the concentration of infected cells and 2-LTR circles [1]. Luo et al. used this model in conjunction with data from an initial integrase inhibitor intensification experiment, done by Buzon et al, to
develop initial estimates for model parameters [1, 66]. The parameter estimates from the study done by Luo et al. serve as the prior parameter estimates for our analysis. We also used these parameter estimates to develop a pool of simulated patients based on the unscented transform method. In doing so we were able to reduce the five dimensional parameter space to 11 unscented transform patients with the same first and second moment characteristics as original distribution [50]. Using a Bayesian Markov Chain Monte Carlo (MCMC) method we estimated the posterior distributions for these patients under the various experiment designs [58, 60]. The Kullback-Leibler Divergence (KLD) is calculated for each patient and then weighted by the probability of the patient occurring. Summing the results gives the EKLD.

The first thing to consider when designing a HIV integrase inhibitor intensification clinical trial is which type of assay will be used to quantify the HIV DNA, as the amount of information provided from an experiment is heavily dependent on the accuracy of the measurement assay. A real time Quantitative Polymerase Chain Reaction (qPCR) assay is the cheapest and most common assay used to quantify DNA; however, the log normal measurement noise associated with the amplification and measurement of the DNA particles creates a great deal of uncertainty in the estimation [22, 29, 51, 53]. Droplet Digital Polymerase Chain Reaction (ddPCR) assays have been developed as a more accurate method of DNA quantification; however, they are more complex and expensive. The ddPCR assay uses microfluidic technology to do a PCR analysis on thousands of nanoliter sized droplets. The droplets are marked positive if DNA is present and the expected concentration for the sample is calculated using Poisson statistics [54, 91]. The increased accuracy of the ddPCR assay will allow for more accurate estimates of model parameters.

The next aspect of the experiment design to consider is the sample schedule. The locations of the samples play a significant role in the resulting amount of information that is gained from the experiment. The Buzon et al. experimental data only included six data points over a period of 48 weeks (337 days including day 0) [66]. Due to the sparsely sampled experimental data, the model parameters could only be
estimated to a very broad level of certainty. This indicated that the experiment design did not yield very informative data. Our goal in this dissertation was to develop a method to design the trial such that it will provide the most informative data under any given constraints. The experiment design constraints that we considered were sampling schedule, sampling volume and measurement assay. Due to the monetary cost and burden to the patient samples can only be taken sparingly throughout the experiment [68, 69, 72, 81]. In the experiment done by Buzon et al., six samples were taken so we first optimized the experiment for six samples. Our intention was to find the location of the sample points that would maximize the information content of our parameter estimates as measured by the Expected Kullback-Leibler Divergence between model prior parameter distributions and posterior distributions. We then repeated the analysis optimizing the experiment under the constraint that only four samples would be taken during the experiment. In both sample schedule optimization cases, the four sample optimization and six sample optimization, the optimal solution was found using a genetic algorithm with ELKD as the measure of fitness for each candidate schedule. The genetic algorithm converges to a family of solutions with similar properties, which upon further analysis showed the relative criticality of the chosen sample points. There are usually two or three critical points which are always selected for, and the remaining points are slightly less critical but are usually still heavily selected for.

Generally the accuracy of the measurement is a function of the volume of blood that is drawn from the patient and used for analysis as larger blood volumes lead to more accurate measurements [51, 54]. Increasing the accuracy of the measurement by sampling larger volumes also increases the amount of information gained in the experiment. The ideal scenario is one where large volumes of blood can be drawn from each patient at each sample; however, the Institutional Review Board (IRB) stringently regulates the amount of blood that can be drawn during a given clinical trial period [72, 81]. Due to this strict constraint, we set up a constrained optimization problem in which an upper bound is established for the total amount of blood that can be drawn during the experiment, but the amount of blood that can be drawn at any given sample
day is free to vary. Using the optimal schedules found in the sample schedule analysis for both the four and six sample cases the constrained optimization was done to find the optimal volume distribution schedule that maximized the EKLD. For both the four and six sample schedules the optimal volume distribution for the samples varied from the uniformly distributed volume samples. The method indicated that large volumes be taken for critical samples and smaller volumes be taken for less informative samples. When comparing EKLD for the volume optimized schedules to the uniform volume schedules, the volume optimized schedules provided a greater gain in information for all cases; however in some cases it was not a significantly greater gain in information over the uniform volume case.

8.2 Experiment Design Implications

The results of our analysis show that using our methods can be used a robust tool to optimize the design of the clinical trial to garner as much information as possible about the patients. HIV dynamics. We demonstrate the utility of our method by using it to optimize HIV 2-LTR experiments and compare the results to optimal designs using traditional methods. Our results demonstrate that all of the optimization methods provided a greater gain in information than the experimental schedules used in both the Buzon et al. and Hatano et al. experiments. The optimal experiment designs developed using the EKLD method consistently outperformed the other designs selected by the traditional experiment design optimality criterion. Our method even proved to be robust to changes in measurement assay. When the experiment was designed for a qPCR assay our method provided the highest gain in information even when a ddPCR assay is used in the experiment. The same was true when the experiment was designed to be done with a ddPCR assay and a qPCR assay was used in the experiment instead. Due to the Bayesian nature of our analysis and the inclusion of the prior parameter information, it provides a more exhaustive analysis of optimal experiment design.

HIV 2-LTR trials can serve as a critical procedure to use to determine if and to what extent ongoing viral replication is occurring with HIV(+) patients. This
knowledge is essential as it will guide the individual patient treatment of the disease. In order for the procedure to be useful, it must reap the maximum about the patients’ ongoing replication as possible. This drives the need to optimize the trial with analyses such as those presented in this thesis. The results of our analysis show that sample times, number of samples and sample assay all have a large effect on the amount of information that the experiment will provide.

In general monetary cost and patient burden are two of the main concerns when conducting any clinical trial [68, 69, 72, 81]. A recent Nature review paper of a study of clinical trials stated that each additional month for phase II clinical trials translates into a median $671,000 spent [69]. In our analysis we showed that when compared to the experimental schedules used in the Buzon et al. and Hatano et al experiments our method is able to provide a much greater gain in information over a much shorter experiment window. In the case of the Hatano experiment our analysis shows that two optimally placed points would have provided significantly more information than the four points chosen in the experiment. For the Buzon experiment our analysis shows that three optimally chosen points would yield a significantly greater gain in information than the six points chosen in the experiment. In both cases more information could have been gained with half of the effort if the sample schedule had been adjusted. We clearly see that optimizing the experiment design would lead to more informative data and at a considerably lower cost.

8.3 Future Work

In this dissertation we effectively apply our method to optimize HIV 2-LTR clinical trials post intensification with an integrase inhibitor to detect and quantify the level of on-going replication. Our method can easily be extended to other systems. We have also applied our method to design an experiment to study the intracellular dynamics associated with HIV infection and latency. Future work will focus on extending our method to design experiments for other systems both biological and non-biological.


