

SMART PATIENT DIGITAL TWIN

Bhagyashri Jadhav¹, Rohit Raut², Supriya Laykar³, Pavankumar Kale⁴,

Gauri Langde⁵, Harshada Misal⁶ Mohit Pednekar⁷

Department of Computer Science and Engineering (Data Science)

Dr. D. Y. Patil Pratishthan's College of Engineering, Kolhapur

bhagyadyp27@gmail.com, rohitraut8680@gmail.com, supriyalaykar7@gmail.com,

kalepavan298@gmail.com, gaurilangde0909@gmail.com,

misalharshada6@gmail.com, mohitpednekar959@gmail.com

Abstract

A Smart Patient Digital Twin is a copy of a patient that mixes patient data, disease details, treatment records, clinical measurements and health outcomes over time. This project proposes a Smart Patient Digital Twin framework to compare treatment options using patient-specific information. The Smart Patient Digital Twin system focuses on supporting treatment related decision-making than automatically prescribing a treatment. The proposed Smart Patient Digital Twin framework consists of two disease modules: Heart Failure and Non-Small Cell Lung Cancer (NSCLC) with a common treatment comparison module. The datasets contain patient information, treatment records, longitudinal follow-up data, clinical events, treatment response, disease progression survival-related outcomes and adverse events. The longitudinal data allows the Smart Patient Digital Twin to represent changes in a patient's condition across visits instead of using only a single patient record. The Smart Patient Digital Twin system organizes characteristics and disease information and links them with treatment and followup outcomes. Different treatment groups can then be compared using response, progression, survival-related outcomes and adverse events. The results from the dataset analysis show that the selected data provides the necessary patient, treatment, longitudinal and outcome information needed to develop the proposed Smart Patient Digital Twin framework. The Smart Patient Digital Twin system is designed as a decision-support tool that presents information to support discussion between healthcare professionals and patients. The Smart Patient Digital Twin does not independently. Prescribe a treatment. Since the initial implementation uses clinical data the results demonstrate the technical feasibility of the proposed Smart Patient Digital Twin framework rather, than clinical effectiveness. Further development will focus on implementing and evaluating treatment comparison models using the longitudinal patient data.

Keywords

Artificial intelligence, Machine Learning, Digital Twin, Healthcare, Predictive Analysis, Disease Predication, Patient Monitoring, risk Stratification, Clinical Decision Support, Deep Learning.

1 Introduction

Healthcare systems collect amounts of patient information. This includes histories, lab results, vital signs, diagnoses, treatments and follow-up visits. These records hold details about a person's health.. Often this information is stored in separate files. As a result it can be hard to see how a patient's health condition changes over time. Putting all this information together is important.

It helps in evaluating treatment choices and tracking outcomes. A Digital Twin is a copy of a real-world object. It updates itself using data from the object. In healthcare a patient digital twin can represent a patient. It uses data such as age, gender, medical history, test results, disease status, treatment records and outcomes. This virtual model gives a way to track a patient's health and how it evolves. The Smart Patient Digital Twin system focuses on comparing treatments. It does not automatically suggest treatments. Patients differ in ways. They have health conditions different stages of disease, different medical histories and different reactions to treatments. So just looking at the treatment used for a disease may not be enough. This system compares two or more available treatments. It uses patient- data and past treatment results to make the comparison. The project starts with two disease areas: Heart Failure and Non-Small Cell Lung Cancer (NSCLC). The data includes patient details along with ongoing records and clinical events. A shared treatment comparison module will be built for both diseases. This module will take in characteristics and clinical data. It will then compare the treatment options. The comparison can include expected response, to treatment, disease progression, survival-related results and any side effects seen in the data. Statistical methods and machine learning can be used to find patterns in the data. This supports the comparison process. The system is meant to help doctors and patients make decisions. It will give information. It will not decide which treatment to use. The final

treatment choice stays with the doctor and the patient. This decision is based on judgment and the person's individual situation. Since the project uses clinical data it is a prototype. It demonstrates the idea of a Smart Patient Digital Twin. How it can be used to compare treatment outcomes.

2 Methodology

The proposed study focuses on creating a decision-support system that looks at different treatment options for patients with different diseases. The main goal is to give information about treatment results how the disease changes over time check-ups and bad reactions so that doctors and patients can better understand the available treatment options. The system is made to help and not to choose a treatment for a patient automatically.

2.1 Research Design

The study uses a data-based way of working with term clinical data. Two disease parts are looked at in the system: Heart Failure and Non-Small Cell Lung Cancer (NSCLC). For each disease information about traits how bad the disease is, treatment details, check-up results how the treatment works how the disease gets worse results related to living longer and bad reactions are looked at. The overall way of working includes preparing data analyzing at the level finding out about treatment and results looking at data over time comparing treatments making machine learning models and putting the results into a decision-support system.

2.2 Dataset Preparation

The study uses the HLT-004 data sample, which has data for Heart Failure and NSCLC separately. For each disease three kinds of files are used: data, long-term check-up data and event data. The initial data gives details around the time of diagnosis including things like age how bad the disease is, what treatment was given, how the treatment worked how long the patient stayed free from disease and how long they lived. The long-term data has repeated checks during follow-up including lab results, disease stage, treatment status and how bad the side effects were. The event data gives events in the patients treatment time like when they were diagnosed, when treatment started, when it ended, when they were checked when the disease got worse and when they lived or died. The data used is made up; therefore the results from the study are seen as a way to show how the method works not as proof that treatments work in life.

2.3 Disease-Specific Data Organization

The data is split into two parts for each disease. In the Heart Failure part the treatment groups available are GDMT Core, Ivabradine and LVAD. How bad the disease is shown by the NYHA classification and other initial traits. In the NSCLC part the treatment groups available are Carboplatin plus Paclitaxel, Docetaxel, Carboplatin plus Pemetrexed plus Pembrolizumab and Pembrolizumab. How bad the cancer is shown by the TNM-based stage and other patient traits. Even though the two diseases have things to look at both parts use the same main way of handling data so that the treatment-comparison part can be used in both.

2.4 Data Preprocessing

Before looking at the data the files are checked for missing parts, wrong formats, same records and wrong types of data. Dates are made to look the same. Numbers are made to be correct. Categorical data like gender, disease stage, treatment group and response type are changed when needed for machine learning. Missing parts are handled based on what the data's how important it is not just replaced without checking. The patient ID is kept during the process so that different check-up records for the person are connected properly. So repeated check-ups are not seen as people.

2.5 Combining Initial Long-Term and Event Data

The three data parts are connected using the ID. Initial data is used to say what the patient was like at the start while long-term data is used to see how things change over time. Event data is used to make a timeline of the patients treatment. When treatment started, when it ended, when they were checked, when the disease got worse and when they lived or died are all marked with the dates they happened. This way the study keeps the time between treatment and what happened after.

2.6 Treatment and Outcome Identification

Information about treatment is taken from the treatment in the initial data and the treatment groups in the long-term data. The main results looked at include how the treatment worked how the disease got worse how long the patient stayed free from disease and how long they lived. Categories like Complete Response (CR) Partial Response (PR) Stable Disease (SD) and Progressive Disease (PD) are kept as results. For results about living the time they lived and if they died are used of saying they all had the same time to check.

2.7 Long-Term Check-Up Analysis

Long-term data is studied to see how patient results change during treatment and check-ups. Depending on the disease things like NT-proBNP for Heart Failure and CEA for NSCLC are looked at along with lab tests and disease stage. The analysis looks at the visit number and time since diagnosis so that changes are seen in the time during check-ups. This helps the system give information about what treatment was used and how the patient changed during check-ups of only looking at one initial check.

2.8 Side Effect Analysis

The long-term data has grades for side effects. These are used to say what the safety was like for each treatment group. The analysis looks at when side effects happened and how bad they were, like tiredness, nausea, anemia, infection, blood clots, kidney problems, heart problems, lung issues and other events. Side effects are shown as patterns in the data. They are not seen as risks for all patients because the data is made up.

2.9 Treatment Comparison

The main part of the system is the treatment-comparison part. Of just comparing treatment names the system compares treatment groups using the traits of the patients and what happened. For a chosen disease two or more treatment groups can be compared using things like:

- Number of people in each treatment group
- How well treatment worked
- How often the disease got worse
- How long the patient stayed free from disease
- How long the patient lived
- Check-up results
- Side effects

Differences in the traits of the people in each group are looked at because the groups may not have people with the traits. If needed ways to make differences smaller can be used. The comparison is seen as what was found or what the model guessed not as proof that one treatment is better than another.

2.10 Machine Learning Model Building

Machine learning models are made to look at how traits, what treatment they had and certain results are connected. Patient traits like age, sex, BMI, disease stage how the patient felt, other health problems, lab results and treatment group can be used as inputs. Depending on the result classification or time-to-event methods are used. For results like how the treatment worked or if the disease got worse the right ways to test are used. For time-to-event results like how the patient stayed free from disease or how long they lived time-to-event methods are used.

The data is split at the level into a training group and a testing group. This stops the patient from being in both groups and reduces the chance of data from repeated check-ups being used in the wrong way.

2.11 Model Testing

The machine learning models are tested with the ways for the type of task. For classification tasks things like accuracy, precision, recall, F1-score and ROC-AUC can be used. For time-to-event tasks things like the concordance index and survival graphs can be used. Model results are looked at together with what the data could do. A high score alone does not mean a treatment is better, in life.

2.12 Treatment Comparison Results

After looking at the data the system shows the treatment groups in a comparison. The results can include treatment response

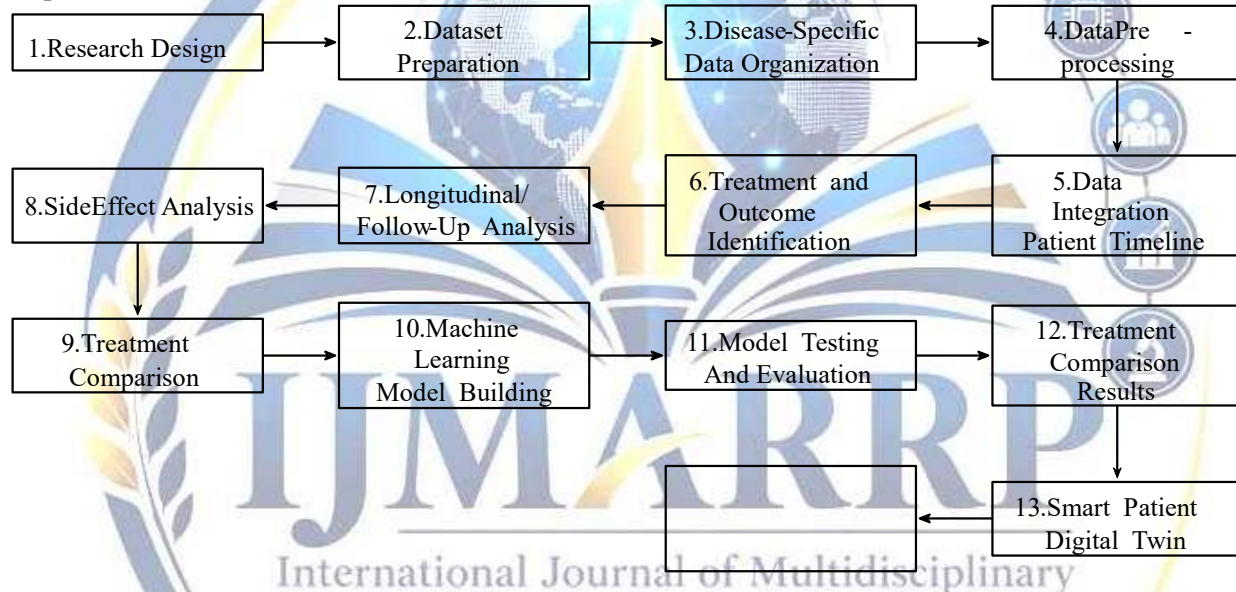


Figure 1: Methodology workflow of the proposed Smart Patient Digital Twin for treatment comparison and outcome analysis.

3 Results

The proposed Smart Patient Digital Twin was built using two disease modules: Heart Failure and Non-Small Cell Lung Cancer (NSCLC). The main goal of this stage was to see if the available data includes information to represent a patient over time and to compare different treatments.

3.1 Heart Failure Results

The Heart Failure dataset includes 400 patients. Each patient has details such as age, sex, BMI, smoking history, disease stage and other clinical characteristics. The dataset also includes treatment groups:

- LVAD
- Ivabradine
- GDMT Core

Treatment outcomes are. Include Complete Response (CR) Partial Response (PR) Stable Disease (SD) and Progressive Disease (PD). This means the Heart Failure data can be used to study how patient traits, treatments, follow-up visits and outcomes are related.

3.2 NSCLC Results

The NSCLC dataset also includes 400 patients. It has details, disease stage, treatment data, follow-up records and clinical outcomes. The main treatment groups are:

- Carbo-Taxol
- Docetaxel 2L
- Carbo-Pem-Pembro
- Pembro Mono

The dataset also includes information on treatment response, disease progression and survival related outcomes. This makes the NSCLC dataset a fit for the second disease module in the proposed system.

3.3 Longitudinal Data Results

An important finding from the data analysis is that both disease modules contain data. These files have records for the same patient at different times. For example a patients journey can be shown as: (Diagnosis → Treatment → Visit 1 → Visit 2 → Visit 3 → Response/Progression → Outcome)

Each visit can include information like treatment status, disease stage, lab results and adverse events. This is crucial for the Smart Patient Digital Twin. The digital twin must not rely on a single record. It must show how the patients condition evolves over time.

3.4 Treatment and Outcome Results

The dataset includes both treatment information and treatment outcomes. For example the data includes events such, as: Diagnosis ,Treatment started ,Treatment ended , Response assessment , Disease progression , Survival endpoint This allows the project to compare treatments based on outcomes rather than just listing which treatment was given.

3.5 Overall Result

The analysis confirms that the selected dataset has all the key information needed for the Smart Patient Digital Twin: (Patient Information + Disease Information + Treatment + Longitudinal Follow-up + Outcomes) This makes it possible to create two disease-specific modules and link them to one shared treatment comparison module. At this point the result is the identification and organization of patient, treatment, follow-up and outcome data. The next step is to build treatment comparison models and generate comparative results from the data. Since the dataset is synthetic the results produced by the system are part of a research prototype. They should not be used as world clinical treatment recommendations.

4 Discussion

The results show that the selected dataset provides the types of information required for developing a Smart Patient Digital Twin. The combination of characteristics, disease information, treatment records, longitudinal observations and clinical outcomes makes it possible to represent a patient beyond a single point in time. One important aspect of the proposed system is the use of data. A patients condition can change during treatment. Repeated clinical observations provide information about these changes. The longitudinal records in the dataset contain visits for individual patients along with treatment status, disease stage, laboratory measurements and adverse-event information. This makes the longitudinal component useful for representing the changing state of the digital twin. The two disease modules also show that the same general framework can be used for diseases. Heart Failure and NSCLC have disease stages, measurements and treatment groups but both can be represented using the same overall structure of patient information, treatment, follow-up and outcome. A common treatment comparison module can therefore be developed while keeping disease- information within each disease module. Another important point is that the dataset contains both treatment information and outcome information. This allows the proposed system to move beyond recording which treatment a patient received. Treatments can be examined in relation to response, progression, survival-related outcomes and adverse events. However differences observed between treatment groups should not automatically be interpreted as proof that one treatment caused an worse outcome. Patient characteristics and disease conditions may differ between groups and further statistical analysis is required before drawing conclusions. The proposed system is intended to provide

decisions support information than automatically prescribe a treatment. The purpose of the treatment comparison module is to organize and present information about treatment options and their observed outcomes for patients with similar characteristics. The final decision should remain with the healthcare professional and patient. The current work also has limitations. The initial implementation uses clinical data so the results cannot be directly considered evidence for real-world medical treatment decisions. In addition the current results mainly demonstrate the availability and structure of the data and the feasibility of the proposed twin framework. Actual predictive performance and patient-specific treatment comparison results require implementation and evaluation of statistical and machine learning models. Overall the findings support the feasibility of using patient data as a foundation, for a Smart Patient Digital Twin. The next stage of the project is to implement the treatment comparison models and evaluate whether the system can provide reliable comparative information from the available patient data.

5 Conclusion

The proposed Smart Patient Digital Twin framework offers a way to store patient information and compare different treatments. When we looked at the chosen datasets we found that both Heart Failure and Non-Small Cell Lung Cancer data include characteristics, disease details, records of treatments long-term follow-up information and clinical results. Having records is especially helpful for the Smart Patient Digital Twin because it lets us show how a patient's condition changes over many visits. When we have information on how patients respond to treatment how the disease progresses, survival outcomes and any sideeffects we have a base for comparing different treatment choices. The project brings together two modules and a shared treatment comparison module. With this structure the team can handle data on its own yet still use one common method to analyze treatment outcomes. The system is meant to show comparison data not to prescribe treatment. The final choice stays with the doctor and the patient. From the results the chosen data and the Smart Patient Digital Twin framework give a good base for building a prototype. This work does not prove that any treatment is better. The research team still needs statistics and machine-learning analysis. Because the dataset we used at first is synthetic the findings are like a technical demo, than real clinical proof. The next part of the project will work on building treatment comparison models checking how well they perform, looking at treatment outcomes and creating a user interface that shows patient-specific comparison data.

References

- [1] K. P. Venkatesh, M. M. Raza, and J. C. Kvedar, "Health digital twins as tools for precision medicine: Considerations for computation, implementation, and regulation," *npj Digital Medicine*, vol. 5, no. 1, p. 150, 2022.
- [2] The Cancer Genome Atlas Research Network, J. N. Weinstein, E. A. Collisson, G. B. Mills, K. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart, "The Cancer Genome Atlas Pan-Cancer analysis project," *Nature Genetics*, vol. 45, no. 10, pp. 1113–1120, 2013.
- [3] H. J. W. L. Aerts, E. Rios Velazquez, R. T. H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebens, M. M. Rietbergen, C. R. Leemans, A. Dekker, J. Quackenbush, R. J. Gillies, and P. Lambin, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature Communications*, vol. 5, p. 4006, 2014.
- [4] M. N. K. Boulos, "Health Digital Twins in Life Science and Health Care Innovation," *Annual Review of Pharmacology and Toxicology*, 2024.
- [5] A. E. W. Johnson, L. Bulgarelli, L. Shen, et al., "MIMIC-IV, a freely accessible electronic health record dataset," *Scientific Data*, vol. 10, p. 1, 2023.
- [6] T. J. Pollard, A. E. W. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi, "The eICU Collaborative Research Database, a freely available multi-center database for critical care research," *Scientific Data*, vol. 5, p. 180178, 2018.
- [7] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.
- [8] P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F. I. Lunze, W. Samek, and T. Schaeffter, "PTB-XL, a large publicly available electrocardiography dataset," *Scientific Data*, vol. 7, p. 154, 2020.

- [9] B. Gow, T. Pollard, L. A. Nathanson, A. Johnson, B. Moody, C. Fernandes, N. GreenJ. W. Waks, P. Eslami, T. Carbonati, A. Chaudhari, E. Herbst, D. Moukheiber, S. Berkowitz, R. Mark, and S. Horng, "MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset," PhysioNet, version 1.0, 2023.
- [10] T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman, "Survival analysis part I: Basic concepts and first analyses," British Journal of Cancer, vol. 89, no. 2, pp. 232–238, 2003.
- [11] J. Cao and S. Zhao, "MIMIC-IV-Ext Cardiac Disease," PhysioNet, version 1.0.0, 2025.
- [12] A. Johnson, L. Bulgarelli, T. Pollard, L. A. Celi, R. Mark, and S. Horng, "MIMIC-IV-ED," PhysioNet, version 2.2, 2023.
- [13] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," New England Journal of Medicine, vol. 380, no. 14, pp. 1347–1358, 2019.
- [14] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," Nature Medicine, vol. 25, no. 1, pp. 24–29, 2019.
- [15] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, "An overview of clinical decision support systems: benefits, risks, and strategies for success," npj Digital Medicine, vol. 3, p. 17, 2020.

