A Multiple Source Data Collection and Integration Paradigm for the Creation of a Dynamic COPD Data Mart

Giulio Pagliari[1][a], Agni Delvinioti[1][b], Nicoletta Di Giorgi[1][c], Maria Vittoria De Girolamo[1][d],
Angela Nervoso[2][e], Francesco Macagno[2][f], Carlotta Masciocchi[1][g],
Stefano Patarnello[1][h] and Alice Luraschi[1][i]

[1]Gemelli Generator RWD R&D, Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Largo Agostino Gemelli 8, 00168 Rome, Italy

[2]CEMAR, Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Largo Agostino Gemelli 8, 00168 Rome, Italy

Keywords: Digital Health, Medical Informatics, COPD, Data Collection, Heterogeneous Data Sources, Real World Data.

Abstract: The creation of dynamic data marts in a hospital environment is challenging due to the number of different data sources, the heterogeneity of data formats and the availability of structured datasets. Other than identifying the relevant pathology and related information, the interaction with the Hospital Information System requires dedicated personnel and an in-depth knowledge of the IT architecture of the Hospital. In this paper, we show an ad-hoc solution for the RE-SAMPLE project in Fondazione Policlinico Universitario Agostino Gemelli IRCCS, where the Chronic Obstructive Pulmonary Disease (COPD) is studied and a framework for managing that pathology is proposed. The final aim of this work is to provide a description of the tailored procedures of data extraction, integration and harmonization, and the final creation of a dedicated COPD data mart for research purposes that has been implemented in the hospital premises by Gemelli Generator RWD R&D.

1 INTRODUCTION

Chronic obstructive pulmonary disease (COPD) is a common lung pathology, considered as the third leading cause of death worldwide (WHO, 2023). Due to symptoms as chronic cough or breathing difficulty, this condition has a great impact on patients' daily life. COPD is considered as non-curable and is often accompanied by other comorbidities, such as chronic heart failure, anxiety and depression, and other adverse conditions that require attention and specific care for patients. For these reasons, patients often need to perform follow-up visits and to stay in touch with clinicians to report any event or exacerbation.

The RE-SAMPLE project (accessible at https://www.re-sample.eu/) has the objective of creating a digital framework to support patients in managing their complex chronic conditions deriving from the concurrent presence of COPD and other comorbidities. With the use of a mobile application and a clinical dashboard, patients can be monitored remotely and can receive tailored suggestions to manage their health condition. Leveraging on AI models, risk profiles can be delivered to health care professionals (HCPs) and, together with the clinical and self-reported data of each single patient, can be used as an additional source of information for predicting symptom worsening and quality of life.

[a] https://orcid.org/0000-0001-8481-1529
[b] https://orcid.org/0000-0002-2402-9444
[c] https://orcid.org/0000-0002-8033-5411
[d] https://orcid.org/0009-0003-4046-6454
[e] https://orcid.org/0009-0005-4976-9691
[f] https://orcid.org/0000-0001-9603-9660
[g] https://orcid.org/0000-0001-6415-7267
[h] https://orcid.org/0009-0008-2765-5935
[i] https://orcid.org/0000-0001-7400-7182

Pagliari, G., Delvinioti, A., Di Giorgi, N., Vittoria De Girolamo, M., Nervoso, A., Macagno, F., Masciocchi, C., Patarnello, S. and Luraschi, A.
A Multiple Source Data Collection and Integration Paradigm for the Creation of a Dynamic COPD Data Mart.
DOI: 10.5220/0013150100003911
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 18th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2025) - Volume 2: HEALTHINF*, pages 507-514
ISBN: 978-989-758-731-3; ISSN: 2184-4305

Through the use of the project tools, the patient can in fact easily report any exacerbation or event, moving their usual care from the hospital to their home and potentially minimizing the number of hospitalizations adverse events day-by-day.

In this paper, we detail the methodology, and the framework developed for the identification and the integration of multiple data sources in the hospital with the final release of a daily updating data mart compliant to the common project data model and specific data quality standards. The data mart serves as a data source for the ingestion of patient data to a Fast Healthcare Interoperability Resources (HL7 FHIR) based data repository for further federated learning and data visualization tasks. Details on the technical requirements and the implementation of the RE-SAMPLE platform are beyond the scope of this work.

The last years, there is growing interest in exploring the benefits from the re-utilization and the integration of Electronic Health Records (EHR) in clinical trials (Kalankesh, 2024) (Nordo, 2019). At the same time, digital tools are introduced to facilitate clinical trial management especially during patient screening and enrolment tasks (Kasahara, 2024). In this work, we explore both EHR integration for data collection along with data validation from multiple sources and dedicated screening and enrolment applications as a part of the RE-SAMPLE infrastructure.

Main challenges in data collection in a real-world setting, such as the hospital environment, are the heterogeneity of data sources – which need to be identified and mapped within the hospital, along with data availability. A description of this problem has been addressed in other works including (Kwok, 2022) and (Kerkri, 2001), and a description of different solutions were reported in (Mate, 2015), where an ontology-based solution is presented, or (Jayaratne, 2019), where the authors introduce an open data integration platform across different sources. The creation of research datasets in such context remains a challenging problem and often leads to ad-hoc solutions that are tailored on the specific Hospital. In COPD research domain, most works focus on data modelling and disease characterization problems while few ones focus on systematic data collection such as the collaborative approach for the definition of a COPD dataset in a Healthcare System reported in (Lam, 2023).

To tackle these challenges in RE-SAMPLE project, a core facility of Fondazione Policlinico Universitario Agostino Gemelli IRCCS (FPG) named Gemelli Generator RWD R&D (Damiani et al., 2021) has developed a dedicated pipeline for data extraction and data collection with the aim of retrieving all required information from the different data sources that are present in the hospital, including internal tools that support HCPs in managing the prospective study. The group has a relevant track record in the creation of research data marts for other pathologies, such as breast cancer (Marazzi, 2021), heart failure (D'Amario, 2023), dyslipidemia (Capece, 2024) or Covid-19 (Murri, 2022). FPG team actively participated in the definition of both clinical and technical requirements of the RE-SAMPLE platform. To this end, a crucial task was the definition of a common data model (Acebes et al., 2022), that includes all the clinically relevant variables for characterizing the health profile of COPD patients. In fact, several variables are required to capture the health condition of COPD patients with chronic complex conditions. Functional scores based on spirometry measurements, blood samples, along with six-minute walking tests and Patient-Reported Outcomes (PROs) on life habits (e.g. smoking) and symptoms are needed for providing to Health Care Professionals (HCPs) a complete overview of the actual health status of the patient.

Data collection in the hospital requires a shared effort between clinicians and a dedicated technical team not only for the conduction of regular outpatient visits but also for the development of data extraction procedures from the Hospital Information System (HIS) or the EHR that make hospital data available for further visualization and modeling tasks.

In the following sections, a description of the implementation of an ad-hoc solution for the creation of the RE-SAMPLE data mart in FPG is reported, along with the results of the deployment and use of the defined procedures.

2 METHODS

As shown in Figure 1, the creation of the RE-SAMPLE data mart stems from the need of collecting clinical and secondary data for all the patients included in the project. As a first step, screening is required before asking a patient to join the study. This step is made via a web-based recruitment app, where all the inclusion and exclusion criteria are standardized. Interacting with this tool, HCPs can understand whether a patient is eligible for the participation in the study and consequently being enrolled.

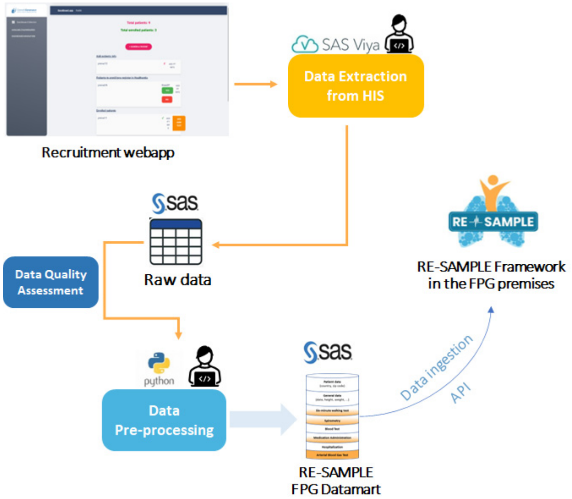


Figure 1: Data pipeline from patient screening to enrolment and creation of the data mart, including the interaction with the RE-SAMPLE framework in the FPG premises, hosted in a dedicated Virtual Machine.

Once the study cohort is identified, and patients are mapped within the HIS, for all the different patients primary and secondary data should be retrieved through dedicated queries and pipelines. In the schema shown in Figure 1, raw data from the Hospital Data Warehouse are collected and processed to create the final project dataset.

In the following sections, we present how the data model was defined and how data sources were mapped and integrated within FPG, towards the creation of the final RE-SAMPLE data mart.

2.1 HIS Dataset

The Hospital Information System (HIS) dataset as defined in RE-SAMPLE project (Acebes et al, 2022) includes several data categories organized in distinct time reference points. Namely, variables get collected during enrolment and a required baseline visit, during follow-up visits with a six-month frequency and during in-patient hospital visits. As demonstrated in Figure 2, in every time reference point named encounter, specific data is captured. Finally, additional data during different encounters such as inpatient and emergency visits are also included.

During enrolment, the clinical team consisted of pulmonologists and research nurses report general patient information including zip code and biometric data such as body mass index. A spirometry test is performed to evaluate COPD diagnosis using the Global Initiative for Chronic Obstructive Lung Disease (GOLD) classification (Agustí et al, 2023). Eligible patients are requested to perform a six-minute walking test and blood exams. Moreover, the clinical team captures information related to clinical history including exacerbations, hospitalizations, smoking, and medication plan and complete together with the patient different questionnaires related to general health, COPD and comorbidity symptoms and mental state. Similar data gets collected successively during follow-up visits apart from some questionnaires.

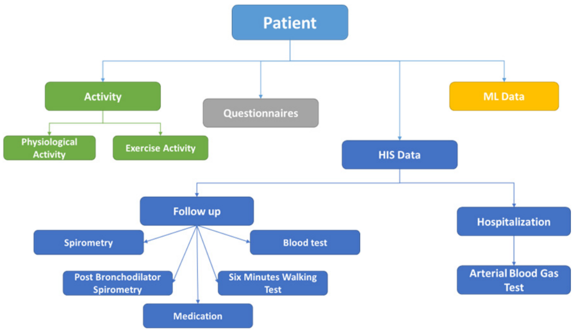


Figure 2: HIS Dataset as a subset of RE-SAMPLE data model. For simplicity we consider also baseline visits as a type of follow-up visit. Adapted from (Acebes et al, 2022).

Regarding hospitalization instead, the dataset includes additional variables related to arterial blood gas tests typically performed during inpatient visits, oxygen use, mechanical ventilation procedures, and pneumonia.

2.2 Data Sources

The different data categories required for the RE-SAMPLE project tasks, get effectively stored in different data sources in the hospital and in different formats. This requires the creation of a procedure that retrieves data from multiple systems and creates an integrated and harmonized project data mart. In this section, a description of the different data sources is provided.

2.2.1 Electronic Health Records (EHRs)

In clinical practice, all data reported by medical personnel during hospital visits, inpatient, or outpatient, gets collected in dedicated hospital systems and stored within the hospital data warehouse (DWH). Part of these sources contain information stored in a structured format, while e.g. in clinical notes and discharge letters are typically available in an unstructured text format containing a summary of the details collected during visits or medical procedures.

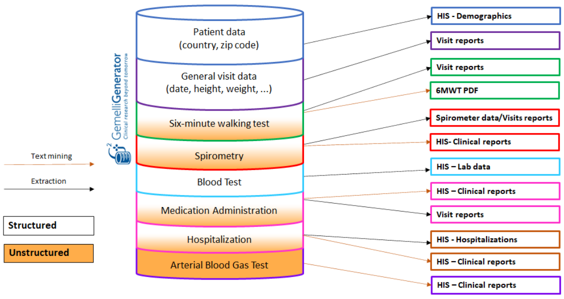


Figure 3: Data categories mapped into data sources.

Figure 3 reports a schematic view of the final RE-SAMPLE data mart in FPG, i.e. a curated patient-centered data repository that includes a subset of all the available data related to the project within the hospital's IT system and data warehouse. Administrative visit data and laboratory related data are instead stored in a structured format. Medications and arterial blood gas test data get extracted from clinical notes and discharge letters. All this data is accessible in the FPG data warehouse in a centralized manner.

2.2.2 Medical Devices

Spirometry and six-minute walking test data is stored locally in medical device dedicated application storages. During functional tests, a dedicated application accompanies each medical device and produces reports in pdf files that include all measurements performed during the tests. Typically, this data is not integrated in the hospital data warehouse and its extraction requires additional effort and dedicated extraction pipelines: in our case, data is retrieved from the devices and saved in the HIS or in the dedicated platform described in §2.3.1 by the HCP that performs the visit.

2.2.3 Self-Reported Data

As mentioned before, during baseline and follow-up visits, patients complete several questionnaires with the assistance of the clinical team. Some of these questionnaires are filled in on paper and later reported in the clinical notes, while most of them are filled in, in a tailored version of a mobile application named Healthentia (iSprint, 2021). The HIS dataset includes only information extracted from centralized hospital systems or locally from medical device storages. External data collected through the patient application during the study are out of scope of this manuscript.

2.3 Data Collection and Data Integration Framework

Data collection is performed in every phase of a prospective clinical study, from the screening of the patients to enrol until the follow-up visits. While during hospital visits data is collected by procedure in hospital systems, this is not the case during screening and enrolment phase. Furthermore, in the specific RE-SAMPLE project use case, in the beginning of the study the patient needs to be evaluated for enrolment. Successively, he or she also needs to be registered in an external mobile application for being monitored and fill in self-reported questionnaires.

To cover such need during enrolment, we developed a dedicated patient enrolment application that facilitates screening but also allows for registering new patients in the mobile application in an automated way. During enrolment, the internal hospital unique identifier gets mapped to the subject ID unique identifier assigned to the patient in the mobile application for the specific study. Such a requirement is crucial to make possible the integration of the various data from the different data domains; in hospital and external that are kept carefully separated in the entire RE-SAMPLE platform following a privacy preserving by design principle. Overall, all this heterogenous data coming from multiple data sources needs to be integrated in a unique data storage following the RE-SAMPLE data model and daily updated with new encounter data for further visualization and modelling tasks.

2.3.1 RE-SAMPLE Enrolment and Screening Applications

As previously mentioned, a custom enrolment web application has been developed within Gemelli Generator RWD R&D services to standardize and streamline patient screening and enrolment processes. The RE-SAMPLE enrolment application, based on a Django v. 3.2.16 and Python v3.11 framework, centralizes diverse features useful for clinicians in the preliminary phase of patient assessment. Patient characteristics are evaluated in a structured way, selecting different alternatives through drop-down menus. Once patient inclusion and exclusion criteria are assessed, they get automatically integrated to define patient eligibility. Eligible patients who agree to participate, get enrolled in the clinical study and a unique in-study identifier is systemically assigned to each of the one. Pre-registration in Healthentia is integrated by a simple click in the enrolment application workflow. An overview page allows clinicians to observe enrolled and drop-out patients, due and effective follow-up visit dates, and to monitor enrolled patient number over time. Patients table is daily recovered from the application SQLite database and stored in a SAS data repository included in dedicated storage areas (SAS Viya 3.5 Caslibs). RE-SAMPLE cohort definition is thus carried out to further extract patient information from HIS. Additionally, the application has been designed to collect specific clinical reports hard to retrieve from HIS with standard procedures (e.g. six-minutes walking test reports). The RE-SAMPLE enrolment app is accessible only in intranet from hospital's internal systems and through specific authentication using accounts reserved for clinical team members.

The inclusion and exclusion criteria integration algorithm, implemented within the enrolment application on the basis of the study protocol, has been further used in the development of a supporting tool for the identification of patients who might benefit from the RE-SAMPLE prospective study out of the FPG Hospital. Indeed, a completely anonymized version of the algorithm has been integrated in a screening application. The RE-SAMPLE screening application is developed to provide general practitioners (GPs) with a simple and straightforward tool for assessing RE-SAMPLE eligibility criteria, to shorten the communication pathway and facilitate collaboration between GP and FPG specialist. The screening application is externally accessible, and no patient data is collected or exposed. GPs can easily perform a screening for inclusion and exclusion criteria while a notification is generated to inform FPG clinical staff when a GP performs a positive screening. When a patient is eligible to participate to the study, it remains GP's responsibility to contact an FPG specialist to introduce the patient in the study.

2.3.2 Visit Template

Clinical notes include an overview of the data reported during hospital visits by procedure in clinical practice. Nonetheless, the unstructured format of the textual data remains a bottleneck for further data availability and exploitation. The FPG technical team introduced a visit template to be used by the clinical team during visits and to facilitate data extraction procedures from text stored in electronic health records.

The visit template includes all required information for the study organized with keywords and specific separator characters in separate lines. In this way, all necessary information is always present in text and in a specific format. At the same time, unstructured information inserted in more data sources serves for multiple source data validation and completion such as in the case of six-minute walking test data, where pdf report files must be manually uploaded to the enrolment application by the clinical team and further processed. Below, an example of the data structure retrieved from such document:

```
#DATA_RESAMPLE#
  #Comorbidities
    DIABETES: YES/NO \n
    CHF: YES/NO \n
  [...]
    OTHER_COMORBIDITY: YES/NO \n
  #Other Sublist
    var_1: value_1\n
    var_2: value_2\n
  [...]
    var_n: value_n\n
```

2.3.3 Data Extraction and Data Integration

As demonstrated in Figure 3, an Extract Transform Load (ETL) procedure runs daily on a SAS server and creates a dedicated SAS data repository accessible in SAS Viya 3.5 Caslibs. This initial SAS data repository includes all structured data such as general patient data, administrative visit data and blood tests and clinical notes, using as a reference the patient cohort stored by the enrolment application in a dedicated table. Already in this level, a rule-based text mining procedure extracts different variables: 1) medication data and standardizes it using Anatomical Therapeutic Chemical (ATC) Classification codes, 2) visit template data by analysing fixed format keywords and 3) hospitalization variables from discharge letters.

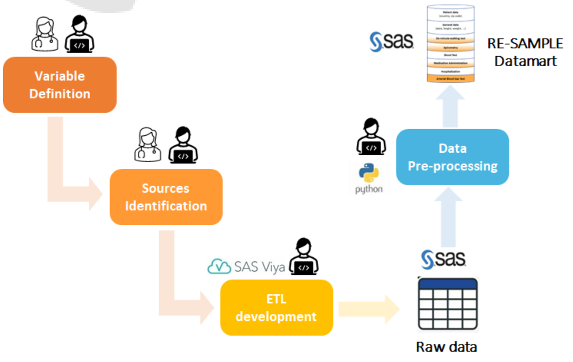


Figure 4: RE-SAMPLE data mart creation pipeline

Successively, a scheduled job in Python v3.9.2, running daily on a dedicated for the project virtual machine (OS Debian 11, 32GB RAM, 4-core Xeon

CPU) processes six-minute walking test reports stored as pdf files in the enrolment application database, processes and combines all different data according to RE-SAMPLE data model into a final SAS data repository, the RE-SAMPLE data mart.

2.3.4 Data Quality

As detailed in previous sections, data collection involves manual procedures performed by the clinical team. Data collected through the enrolment application is standardized and structured and thus of high quality. On the contrary, data extracted from text using the visit template are more liable to errors since the data is standardized by procedure but still inserted manually. To mitigate such errors, to estimate and to improve the quality of the collected data, a scheduled job in Pyttexthon v3.9.2 is running weekly on the same virtual machine and generates detailed reports that include missing rates and error tables per visit type e.g. baseline, follow up stored on the initial SAS data repository. There are two types of reports, one referring to the complete dataset and another only to newly inserted data in a week time. Both reports are automatically sent via email to the technical and clinical team for further and timely validation.

RE-SAMPLE prospective study requires a significant effort level from the clinical team but also particular training in the use of the various applications and procedures to follow. We introduce such data quality procedures as an effort to monitor the quality of the collected data but also to support clinicians to recover missing information whenever possible.

2.3.5 Data Recovery

The data quality procedure generates error tables organized per visit type. To make sure all available text information gets exploited, an additional rule-based text mining pipeline extracts relevant information from clinical notes and completes the collected visit data, where errors or missing values occur. A scheduled job in Python v3.9.2 is running daily after the RE-SAMPLE data mart update and generates tables with missing rates and data recovery rates per data category after the integration of text mining results. All data quality and data recovery results are stored in the initial SAS repository.

3 RESULTS

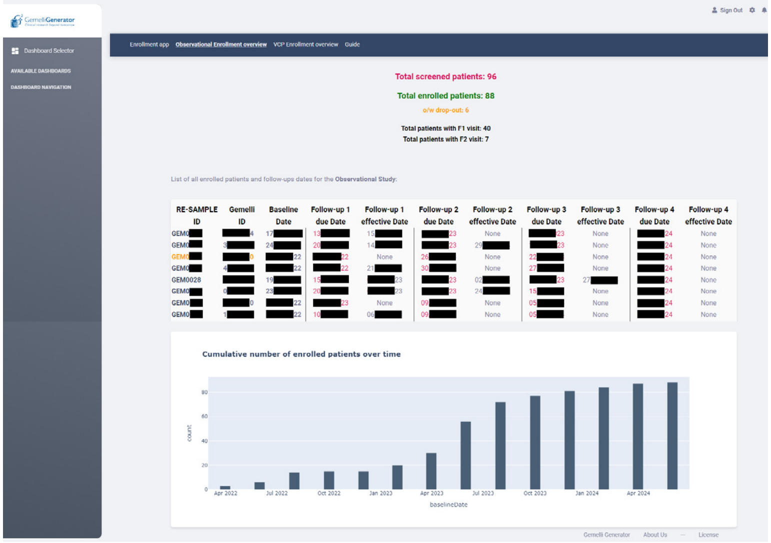
In this section, we present some use cases of the introduced framework for data collection and data integration.

Table 1: Missing rates before and after data quality and data recovery procedures for baseline visits.

Baseline visits (n=85)	Missing rate (%)	
Variable	**Initial**	**Final**
General: weight	2 (2.4%)	0 (0.0%)
SixMinuteWalkingTest: oxygenQuantity	44 (51.8%)	0 (0.0%)
General: bmi	2 (2.4%)	0 (0.0%)
SixMinuteWalkingTest: stopWalking	35 (41.2%)	0 (0.0%)
General: height	2 (2.4%)	0 (0.0%)
SixMinuteWalkingTest: stopWalking	35 (41.2%)	0 (0.0%)
General: mmseScore	6 (7.1%)	1 (1.2%)
General: smokingPackYears	27 (31.8%)	1 (1.2%)
General: mmrcDyspneaScale	4 (4.7%)	1 (1.2%)
Spirometry: fev1	11 (12.9%)	10(11.8%)
Spirometry: fev1Fvc	11 (12.9%)	10 (11.8%)
Spirometry: fev1PercentagePredicted	11 (12.9%)	10 (11.8%)
Spirometry: fvc	11 (12.9%)	10 (11.8%)
SixMinuteWalkingTest: minimumOxygenSaturation	35 (41.2%)	19 (22.4%)
SixMinuteWalkingTest: medication	37 (43.5%)	23 (27.1%)
SixMinuteWalkingTest: percentTimeBelow85	35 (41.2%)	23 (27.1%)
SixMinuteWalkingTest: diastolicPressure	35 (41.2%)	23 (27.1%)
SixMinuteWalkingTest: walkingAid	38 (44.7%)	23 (27.1%)
SixMinuteWalkingTest: systolicPressure	35 (41.2%)	23 (27.1%)
SixMinuteWalkingTest: walkedDistance	38 (44.7%)	24 (28.2%)
SixMinuteWalkingTest: oxygenUse	44 (51.8%)	24 (28.2%)
SixMinuteWalkingTest: theoreticalWalkedDistance	38 (44.7%)	29 (34.1%)

3.1 RE-SAMPLE Enrolment Process

In Figure 5, we present a couple of screenshots of the enrolment application. Here, the user has access to a



Total screened patients: 96

Total enrolled patients: 88

o/w drop-out: 6

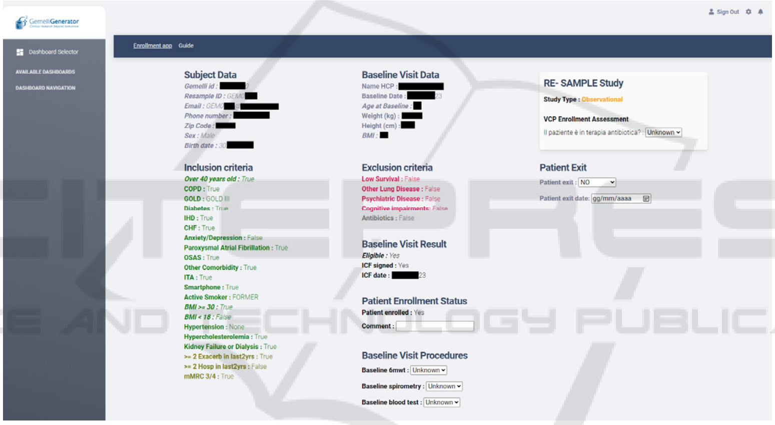
Total patients with F1 visit: 40
Total patients with F2 visit: 7

List of all enrolled patients and follow-ups dates for the **Observational Study**:

RE-SAMPLE ID	Gemelli ID	Baseline Date	Follow-up 1 due Date	Follow-up 1 effective Date	Follow-up 2 due Date	Follow-up 2 effective Date	Follow-up 3 due Date	Follow-up 3 effective Date	Follow-up 4 due Date	Follow-up 4 effective Date
GEM0[redacted]	[redacted]4	17[redacted]	13[redacted]	15[redacted]	[redacted]23	None	[redacted]23	None	[redacted]24	None
GEM0[redacted]	3[redacted]	24[redacted]	20[redacted]	14[redacted]	[redacted]23	29[redacted]	[redacted]23	None	[redacted]24	None
GEM0[redacted]	[redacted]0	[redacted]22	[redacted]22	None	26[redacted]	None	22[redacted]	None	[redacted]24	None
GEM0[redacted]	4[redacted]	[redacted]22	[redacted]22	21[redacted]	30[redacted]	None	27[redacted]	None	[redacted]24	None
GEM0028	[redacted]	19[redacted]	15[redacted]	[redacted]23	[redacted]23	02[redacted]	[redacted]23	27[redacted]	[redacted]24	None
GEM0[redacted]	0[redacted]	23[redacted]	20[redacted]	[redacted]23	[redacted]23	24[redacted]	15[redacted]	None	[redacted]24	None
GEM0[redacted]	[redacted]0	[redacted]22	[redacted]23	None	09[redacted]	None	05[redacted]	None	[redacted]24	None
GEM0[redacted]	1[redacted]	[redacted]22	10[redacted]	06[redacted]	09[redacted]	None	05[redacted]	None	[redacted]24	None

Cumulative number of enrolled patients over time

Figure 5: Enrolment application; enrolled patients overview.



Subject Data
Gemelli id : [redacted]0
Resample ID : GEM0[redacted]
Email : GEM0[redacted]@[redacted]
Phone number : [redacted]
Zip Code : [redacted]
Sex : Male
Birth date : 30[redacted]

Baseline Visit Data
Name HCP : [redacted]
Baseline Date : [redacted]23
Age at Baseline : [redacted]
Weight (kg) : [redacted]
Height (cm) : [redacted]
BMI : [redacted]

RE- SAMPLE Study
Study Type : Observational

VCP Enrollment Assessment
Il paziente è in terapia antibiotica? : Unknown

Inclusion criteria
Over 40 years old : *True*
COPD : True
GOLD : GOLD III
Diabetes : True
IHD : True
CHF : True
Anxiety/Depression : False
Paroxysmal Atrial Fibrillation : True
OSAS : True
Other Comorbidity : True
ITA : True
Smartphone : True
Active Smoker : FORMER
BMI >= 30 : *True*
BMI < 18 : *False*
Hypertension : None
Hypercholesterolemia : True
Kidney Failure or Dialysis : True
>= 2 Exacerb in last2yrs : True
>= 2 Hosp in last2yrs : False
mMRC 3/4 : True

Exclusion criteria
Low Survival : False
Other Lung Disease : False
Psychiatric Disease : False
Cognitive impairments: False
Antibiotics : False

Baseline Visit Result
Eligible : *Yes*
ICF signed : Yes
ICF date : [redacted]23

Patient Enrollment Status
Patient enrolled : Yes
Comment :

Baseline Visit Procedures
Baseline 6mwt : Unknown
Baseline spirometry : Unknown
Baseline blood test : Unknown

Patient Exit
Patient exit : NO
Patient exit date: gg/mm/aaaa

Figure 6: Enrolment application; patient data view.

comprehensive overview of all planned and completed visits as required for the specific study. The clinical team can easily spot upcoming follow-up visits in red and drop-out patients in orange. Additionally, a monthly trend plot provides a better picture of enrolment progress over time.

Instead, as shown in Figure 6, the user can access any moment the complete list of information collected per screened or enrolled patient. Additionally, in this view the user can tag the patient as drop-out and can fill in information related to conducted follow-up visits.

3.2 Data Quality and Data Recovery

In Table 1, we report the results of the data quality and data recovery pipelines on a specific variable group with respect to the missing rates. As clearly demonstrated, the overall data integration pipeline introduced in this work, contributes significantly to the reduction of missing rates. Please note that missing rates also include cases of data with wrong format where the data is available but not as expected.

4 CONCLUSIONS

In this work, we present an innovative paradigm for collecting and integrating prospective study data from multiple sources dynamically over time in a hospital setting. Such methodology and framework successfully serve RE-SAMPLE prospective study purposes up until now, that is, the creation of a daily updated COPD data mart available for data ingestion into a comprehensive platform for further data exploitation tasks.

Support applications for enrolment and screening process significantly increase data quality and facilitate study management. Nonetheless, clinicians and non-technical personnel needs to be trained properly and get progressively familiar with the proposed tools.

Visit templates guarantee easier data integration as all patient data remains in hospital data sources. Additionally, clinicians do not need to use multiple tools to insert patient data. However, they appear to be more liable to human errors especially when the study protocol to follow is complex and requires significant amount of time per visit. Thus, structured data based CRFs might be a better solution to guarantee higher data quality with the cost of required training and higher complexity for the clinical personnel.

ACKNOWLEDGEMENTS

This work is supported by the RE-SAMPLE project that has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 965315.

REFERENCES

Acebes A., et al. (2022, August 31). RE-SAMPLE D4.1: Representation of Multi-Modal Data and Disease Progression Monitoring Features. Available at https://www.re-sample.eu/resources/deliverables/ (accessed October 2024)

Agustí, A., et al. (2023). Global initiative for chronic obstructive lung disease 2023 report: GOLD executive summary. *American journal of respiratory and critical care medicine, 207(7), 819-837.*

Capece, U., et al. (2024). Real-world evidence evaluation of LDL-C in hospitalized patients: a population-based observational study in the timeframe 2021–2022. *Lipids in Health and Disease, 23(1), 224.*

Damiani A., et al. (2021). Building an artificial intelligence laboratory based on real world data: the experience of gemelli generator. *Frontiers in Computer Science, 3, 768266.*

D'Amario, D., et al. (2023). GENERATOR HEART FAILURE DataMart: An integrated framework for heart failure research. *Frontiers in cardiovascular medicine, 10, 1104699.*

Innovation Sprint (iSprint). (2021). *Healthentia: Driving Real World Evidence in Research & Patient Care.* Available at: https://innovationsprint.eu/healthentia (accessed October 2024)

Jayaratne, M., et al. (2019). A data integration platform for patient-centered e-healthcare and clinical decision support. *Future Generation Computer Systems 92 996-1008.*

Kalankesh, L. R., & Monaghesh, E. (2024). Utilization of EHRs for clinical trials: a systematic review. BMC medical research methodology, 24(1), 70.

Kasahara, A., et al. (2024). Digital technologies used in clinical trial recruitment and enrollment including application to trial diversity and inclusion: A systematic review. Digital health, 10, 20552076241242390.

Kerkri, E. M., et al. (2001) An approach for integrating heterogeneous information sources in a medical data warehouse. *Journal of Medical Systems 25, 167-176.*

Kwok, C. S. et al. (2022). Data collection theory in healthcare research: the minimum dataset in quantitative studies. *Clinics and practice, 12(6), 832-844.*

Lam, S.S.W., et al. (2023). Development of a real-world database for asthma and COPD: The SingHealth-Duke-NUS-GSK COPD and Asthma Real-World Evidence (SDG-CARE) collaboration. *BMC Med Inform Decis Mak 23, 4*

Marazzi, F., et al. (2021). Generator breast datamart—the novel breast cancer data discovery system for research and monitoring: Preliminary results and future perspectives. *Journal of Personalized Medicine, 11(2), 65.*

Mate, S., et al. (2013) "Ontology-based data integration between clinical and research systems." *PloS one 10.1, e0116656.*

Murri, R., et al. (2022). A real-time integrated framework to support clinical decision making for covid-19 patients. *Computer Methods and Programs in Biomedicine, 217, 106655.*

Nordo, A. H., et al. (2019). Use of EHRs data for clinical research: Historical progress and current applications. Learning health systems, 3(1), e10076.

World Health Organization (WHO). (2023, March 16). *Chronic obstructive pulmonary disease (COPD).* Available at https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd) (accessed October 2024).