Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Capturing Clinical Complexity in Internal Medicine: A Scalable Real-World Data Collection Model for Research and Care Innovation

Marco Montagna

montagna.marco@hsr.it

IRCCS Ospedale San Raffaele

Rebecca De Lorenzo
IRCCS Ospedale San Raffaele

Giuseppe A. Ramirez
IRCCS Ospedale San Raffaele

Enrica Bozzolo
IRCCS Ospedale San Raffaele

Laura Leoni
Vita-Salute San Raffaele University

Giulia Pata
Vita-Salute San Raffaele University

Chiara Pomaranzi
IRCCS Ospedale San Raffaele

Elena Rela
Vita-Salute San Raffaele University

Clara Soggetti
IRCCS Ospedale San Raffaele

Giovanni Benanti
IRCCS Ospedale San Raffaele

Cinzia F. Sala
IRCCS Ospedale San Raffaele

Corrado Masciullo
IRCCS Ospedale San Raffaele

Guido Scicolone
IRCCS Ospedale San Raffaele

Valentina Gugliano
IRCCS Ospedale San Raffaele

Giovanni Tonon

Vita-Salute San Raffaele University
Sarah Damanti
IRCCS Ospedale San Raffaele
Moreno Tresoldi
IRCCS Ospedale San Raffaele
Patrizia Rovere-Querini
IRCCS Ospedale San Raffaele

Research Article

Keywords: real-world data, real-world evidence, data preprocessing pipeline, quality control, internal medicine, multimorbidity, extract transform load pipeline

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11431181/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Capturing Clinical Complexity in Internal Medicine: A Scalable Real-World Data Collection Model for Research and Care Innovation

Marco Montagna[1,2], Rebecca De Lorenzo[1,2], Giuseppe A. Ramirez[3], Enrica Bozzolo[4], Laura Leoni[2], Giulia Pata[2], Chiara Pomaranzi[1], Elena Rela[2], Clara Soggetti[1], Giovanni Benanti[4], Cinzia F. Sala[5], Corrado Masciullo[5], Guido Scicolone[5], Valentina Gugliano[5], Giovanni Tonon[2,5], Sarah Damanti[1,2], Moreno Tresoldi[4], Patrizia Rovere-Querini[1,2], on behalf of the MED-Cli Study Group[‡]

[1]Unit of General Medicine 2, IRCCS San Raffaele Hospital, Milan, Italy

[2]School of Medicine, Vita-Salute San Raffaele University, Milan, Italy

[3]Division of Immunology, Transplantation and Infectious Diseases, IRCCS San Raffaele Hospital, Milan, Italy

[4]Unit of General Medicine and Advanced Care, IRCCS San Raffaele Hospital, Milan, Italy

[5]Center for Omics Sciences, IRCCS San Raffaele Hospital, Milan, Italy

[‡]Members are listed in the acknowledgements section

Corresponding author: Marco Montagna, Vita-Salute San Raffaele University, Milan, Italy. E-mail: montagna.marco@hsr.it

ABSTRACT

BACKGROUND: The increasing complexity of patients in internal medicine wards limits the applicability of randomized controlled trial-derived evidence and highlights the need for complementary real-world data (RWD). The MED-Cli study was developed to generate real-world evidence to better characterize patient complexity and support more personalized, efficient care.

METHODS: MED-Cli is a single-centre, prospective observational study enrolling all adults admitted to internal medicine and nephrological wards since 2022 at IRCCS San Raffaele Hospital. MED-Cli relies on a comprehensive RWD pipeline comprising production, collection, integration and analysis. Clinical, laboratory, and administrative RWD are collected via electronic case report forms on the dedicated on-premise data management Cohort Genomic Platform. An additional platform, the San Raffaele Ai CEnter (S-RACE), then allows for integration of collected RWD with further laboratory and imaging data directly acquired from hospital's data lakes. The large volume of generated RWD can then be analysed using statistical and machine learning techniques to generate RWE in the form of descriptive, predictive and prescriptive statistics.

RESULTS: Since the beginning of the study, RWD of 1352 admission episodes have been collected. Data completeness was high overall, except for selected variables (e.g., respiratory rate, alcohol intake), and improving over time. Mean outlier percentage was 2.8%, with evident data entry mistake happening in only few variables.

CONCLUSIONS: By establishing a robust and scalable model for RWD integration in internal medicine, MED-Cli opens new avenues for delivering more timely, personalized, and sustainable care to complex older patients, bridging the gap between clinical practice and data-driven innovation.

Keywords: real-world data, real-world evidence, data preprocessing pipeline, quality control, internal medicine, multimorbidity, extract transform load pipeline.

BACKGROUND

In the dynamic landscape of healthcare, critical and real-time evaluation of our management strategies within hospital wards could significantly influence the quality and effectiveness of patient care. This becomes particularly vital in internal medicine wards, where the complexity of cases due to multimorbidity, extensive therapeutic needs, advanced age and frequent diagnostic dilemmas make clinical decision making and patients' phenotyping particularly challenging [1–4].

Within the Evidence Based Medicine framework, randomized controlled trials (RCTs) are considered the most rigorous and classical methodology for evaluating the efficacy of interventions, as they minimize bias through random allocation and strict protocol adherence [5–7]. However, it is hard to transfer the knowledge generated by traditional RCTs to the internal medicine context, as these studies often exclude patients with complex health profiles and are therefore not representative of the real-world populations cared for in internal medicine wards [8–16]. Moreover, therapeutic goals in such patients are often multifaceted, and available resources limited, making the decision-making process even more demanding. It becomes then evident that, if informed and context-relevant decisions are to be made, there is a need for complementary evidence generated during everyday clinical activity to supplement that coming from RCTs. This evidence, referred to as real-world evidence (RWE), provides precise insights that not only reflect our current practices but also highlight potential areas for optimization in resources and everyday patient management [17–22]. On PubMed, RWD are identified with the MesH term "routinely collected health data", inserted in 2021 and defined as "Data collected for purposes other than research. Examples include health administrative data, electronic health record data, and disease or clinical registry data."[23].

The aim of the present manuscript is to describe the methodological framework of the MED-Cli study, a large institutional RWD collection initiative launched within the Internal Medicine Departments of IRCCS San Raffaele Hospital, a tertiary-care centre in Milan (Italy). The data collected will be leveraged for multiple purposes, including:

1. Characterizing the clinical phenotype and care complexity of patients admitted to internal medicine wards, to align it with current patterns of resource allocation;
2. Developing predictive models for in-hospital and post-discharge adverse outcomes (e.g., mortality, prolonged hospitalization, therapeutic complications, re-hospitalizations), to inform early preventive strategies;
3. Supporting continuous and personalized clinical decision-making during hospitalization, with the goal of improving care quality, reducing non-beneficial interventions, and promoting more efficient use of hospital resources [24,25].

MATERIALS AND METHODS

Study design, participants, and sampling method

The MED-Cli study is a single-centre, prospective, observational trial conducted in the Internal Medicine departments of IRCCS San Raffaele Hospital, a tertiary care centre in Milan, Italy. All adult patients (≥18 years of age) admitted to participating departments as inpatients or outpatients are eligible for inclusion. Those providing written informed consent are consecutively enrolled. The study was launched in 2022 and is expected to remain open for 10 years. Based on historical admission rates (e.g., over 1,900 patients admitted in 2019) and observed recruitment coverage (roughly 65%), the projected sample size is approximately 10,000 individuals. The overall workflow of data generation, harmonization, and analysis is summarized in Figure 1.

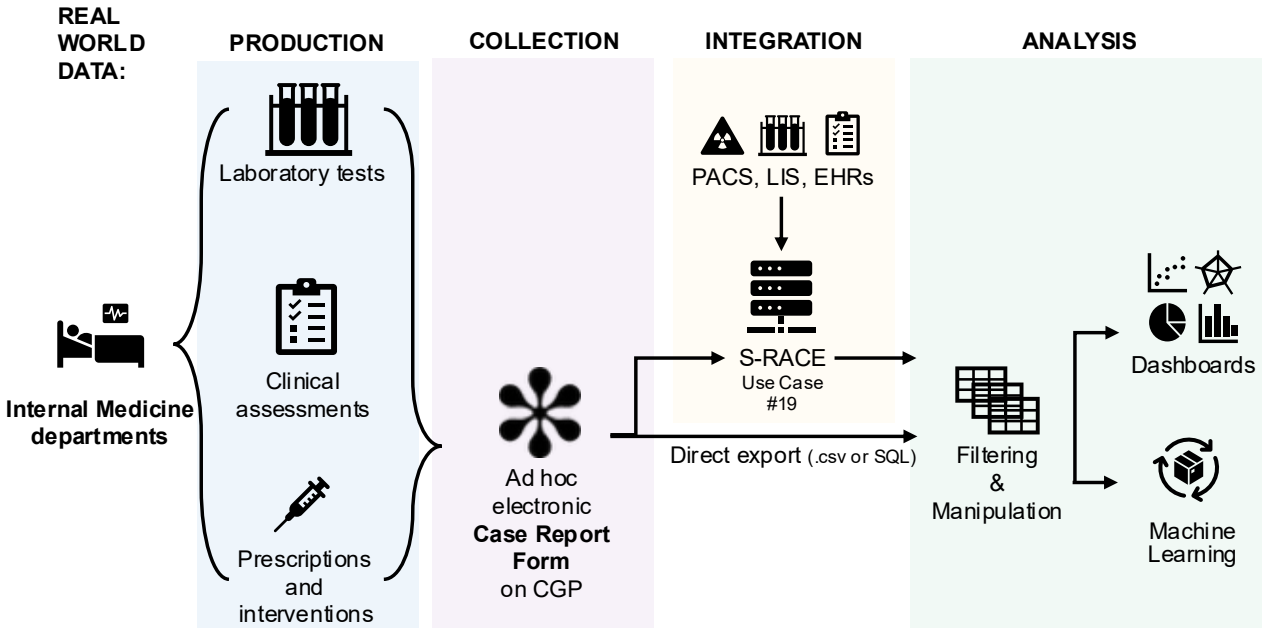

Figure 1 – ***Schematic representation of MED-Cli project's Real-World Data (RWD) pipeline from production to analysis.*** *Workflow of RWD acquisition, integration, and analysis within the MED-Cli study. RWD are produced during routine care in internal medicine wards and include clinical assessments, prescriptions, laboratory results and administrative records. These data are collected through a dedicated electronic case report form hosted on the Cohort Genomic Platform (CGP). Integration with imaging data from the hospital's Picture Archiving and Communication System (PACS), the Laboratory Information System (LIS) and the electronic health records (EHRs) is performed via the San Raffaele Ai CEnter (S-RACE) platform. Extracted datasets can be exported in structured formats and analysed through machine learning models and interactive dashboards for clinical decision support and research purposes.*

Data generation, collection, and management

The study is grounded in the principle that each interaction between patients and the healthcare system produces valuable RWD. While primarily intended to support clinical care, these data can be systematically repurposed for research, provided that appropriate governance and methodological safeguards are in place. In this study, routinely collected clinical, laboratory, and administrative data are captured to generate RWE relevant to patient management in internal medicine.

We developed a dedicated electronic case report form (eCRF) on the Cohort Genomic Platform (CGP) to standardize and structure data collection (Figure 2). To this aim, software engineers and clinicians interact in an interdisciplinary team and continuously meet to refine and edit the eCRF to accommodate new needs of researchers and new clinical settings generating RWD. Accordingly, a changelog of the eCRF is stored for reproducibility and traceability purposes. Medical students,

resident physicians, consultants, and trained data managers are involved in data entry, with regular training to ensure adherence to standard operating procedures. More experienced staff are responsible for data quality checks and periodic validation.

Data are collected through direct clinical assessment, patient interviews, review of medical records, and automated integration with the hospital's electronic health record system. Variables include demographics, anthropometrics, past medical history, symptoms at onset and admission, imaging and microbiological data, laboratory values, daily assessments, outpatient assessments, treatment interventions, complications, and outcomes.

The CGP platform provides: i) secure data segregation and user-specific access via granular authorization layers; ii) GDPR-compliant pseudonymization with multilayer ID management; iii) integration with hospital APIs for automated import of lab and demographic data; iv) JSON-based rendering of dynamic, longitudinal eCRFs via SurveyJS; v) mapping of variables to standard dictionaries (e.g., ICD-10, AJCC, ATC..); vi) built-in validation rules, completion prompts, cross-field consistency checks, and logic gating for quality assurance.

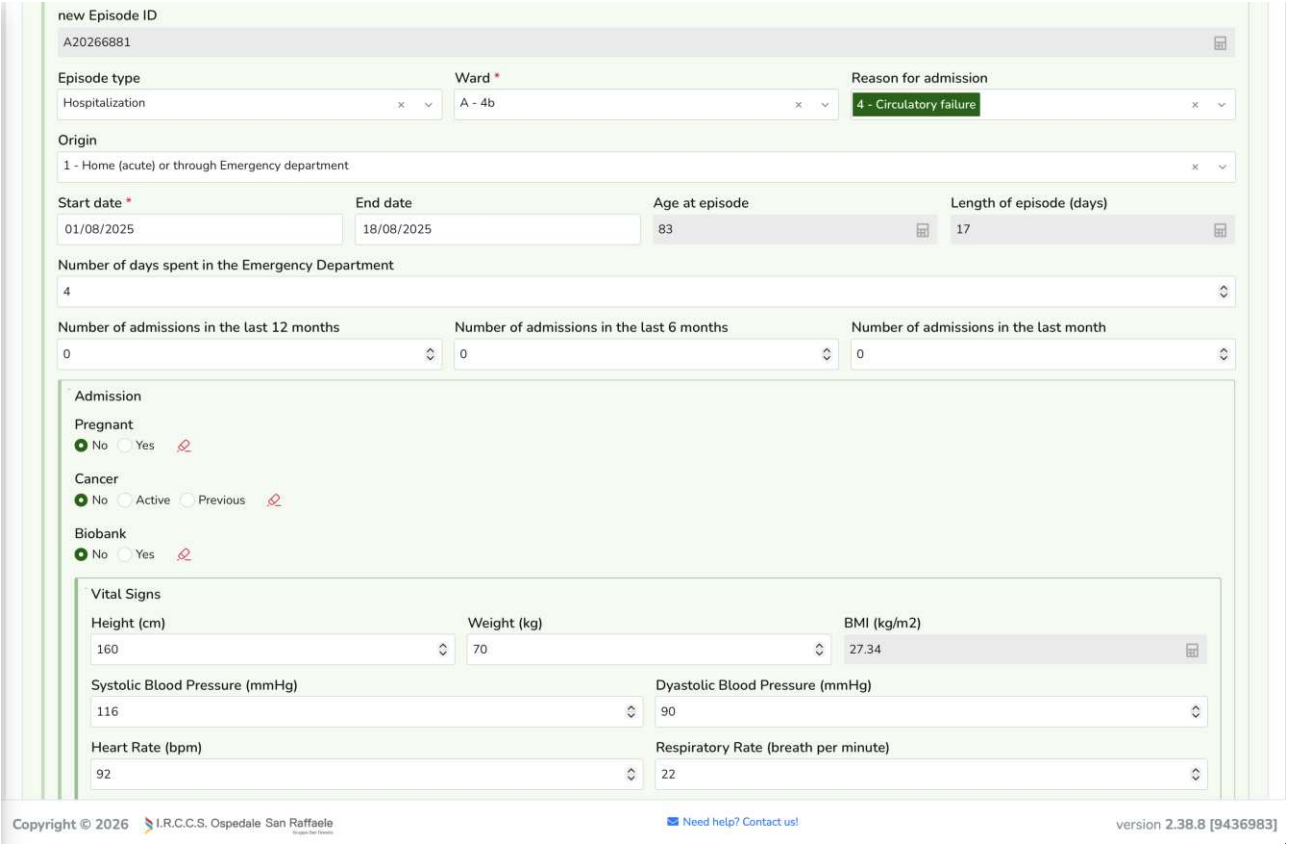

Figure 2 – ***Screenshot of the electronic Case Report Form (eCRF) developed to collect internal medicine Real-World Data and hosted on the Cohort Genomic Platform of IRCCS San Raffaele Hospital****. The JSON-based dynamic and longitudinal eCRF is rendered via SurveyJS and is mostly constituted by structured fields with validation rules and pre-defined data structures and dictionaries (e.g. ICD10, AJCC, ATC..).*

Data extraction, integration, and analysis pipeline

CGP enables flexible data export by transforming raw eCRF content into structured entity-relationship models. Users can define, save, edit and reuse customized data views, which can be exported in .csv or .xlsx formats, or integrated into SQL databases with automatically generated secure credentials. Views can be customised by choosing: 1) the portions of the eCRF (e.g. demographic information, medical history, laboratory exams, episode-level data..); 2) the cohorts of patients (e.g. “MET”: metabolic diseases; “RAR”: rare diseases; “CCF”: chronic-complex-fragile..); specific filters on variables (e.g. episode start date windows by filtering on the “startDate” column in the episode table, or patients with specific discharge diagnoses by filtering on the “ICD10 code” column in the discharge diagnosis table). Extracted datasets can be transferred to the on-premises

San Raffaele Ai CEnter platform (S-RACE), where they can be integrated with imaging data from the hospital's Picture Archiving and Communication System and with laboratory data from the Laboratory Information System[26]. Information coming from each data source for each patient is matched by mapping the pseudonymisation codes of CGP ("pubID") to those of S-RACE ("Cloud Patient Index") both of which can be in turn mapped to the original patient ID in the hospital's IT environment ("Master Patient Index").

We then conduct regular data quality checks and subsequent data analysis and modelling according to individual research questions using both traditional univariate and multivariate statistical methods and machine learning (ML) techniques through the Azure Machine Learning environment provided in S-RACE (Microsoft Corporation). To this aim, classical Python data manipulation and statistical packages including Pandas, Numpy, Matplotlib and Sci-kit learn are used.

Data visualization

We provide real-time visualization and exploratory insights through Power BI dashboards (Microsoft Corporation). PowerBI is part of the Office suite and was originally meant for business intelligence purposes, such as tracking orders, revenues and customers in the manufacturing sector. We adopted the same approach to enable a data-driven management of our clinical activity by tracking relevant indicators including monthly admissions and discharges, length of stay in the emergency department and in the wards, demographics, and clinical complexity of managed patients in terms of number of comorbidities and chronic medications. To this aim, we prepared a data view with the full-length eCRF structured as a SQL database which we interrogated from the PowerBI environment. We then generated a data model of the contents of the database to link information coming from its different tables. Last, we prepared a series of dashboards with interactive plots and cards to visualise and explore the aforementioned indicators taking advantage of ad hoc live slicing tools to filter data according to relevant variables.

Ethical considerations

The study protocol was reviewed and approved by the institutional ethics committee of IRCCS San Raffaele Hospital and is registered at clinicaltrials.gov (NCT05780099, ethics approval 09 June 2021; study start 24 June 2022; first submitted to ClinicalTrials.gov on 21 February 2023 and first posted on 22 March 2023, i.e., registered retrospectively). All participants provide written informed consent prior to inclusion. The study was conducted in accordance to the Declaration of Helsinki.

RESULTS

Data collection

The MED-Cli eCRF hosted on CGP (Fig. 2) is divided in 7 sections (Table 1). The Demographics and Baseline Patient Characteristics section collects data such as sex, age (both automatically imported from the hospital's registry), ethnicity, civil status, years of education, alive or dead status, and the Charlson Comorbidity Index[27]. The Clinical History section collects the ICD-10-coded past medical history, with the possibility to record the age at onset and age at diagnosis for each added disease. It also collects information on vaccinations and risk factors (smoking, alcohol, recreational drugs), and on drug allergies. The Family History section contains yes/no questions on a series of diseases such as cardiovascular diseases, hypertension, diabetes, obesity, and more, and a sub-section that allows to record the presence of cancer history with the ICD10 coding of the tumour type and the relative affected. The Detailed Clinical History section allows to collect further anamnestic details such as pregnancy history, autoimmunity, past or chronic infections and more. The Episodes collects information on inpatient/outpatient hospitalisation events for each patient, with multiple entries allowed. For each new event, many subsections can be filled, allowing to gather data about admission (reasons for referral, vitals, performance scores..), hospital course (complications, interventions..) and discharge (modality, ICD10 discharge diagnoses, vitals, performance scores..).

The Laboratory Exams section allows to manually select and import from the laboratory informatic system specific laboratory panels (blood, urine, other fluids) carried out on patients during the specific episode.

Table 1 – Structure summary of the electronic Case Report Form in use for data collection in the MED-Cli study.

Section	Collected data
Demographics and Baseline Patient Characteristics	Informed consent date, demographic characteristics, Charlson comorbidity index
Clinical history	Past medical history, social history, allergies, vaccinations
Family history	Family history, genetic risks
Detailed clinical history	Pregnancies and births, autoimmunity, infections, metabolic risk factors
Episodes	Episode type (inpatient/outpatient), date, length, variables at admission, hospitalization course, outpatient evaluations, variables at discharge
Laboratory exams	Blood panels, fluid (blood, urine, others) cultures or virological tests, serology
Therapies	Chemical substances, administrations (purpose, route, dose, start date)

Each record is labelled with one of eleven "timing" categories (e.g. "hospital admission", "hospital discharge", "acute event during hospitalisation", "outpatient", etc). The Therapies section collects information on ATC-coded medications currently or previously administered to the patient, together with purpose, route, dosage and reason for interruption, eventually. The resulting database is composed of 89 tables, for a total of 789 variables. The full list of the database tables is available in the supplementary materials.

Database freeze and general database characteristics

To assess the overall quality and reliability of the collected RWD, we performed a database freeze on January 31st, 2026, focusing on inpatient admission events happening in two departments (Ward "A": Unit of General Medicine, Metabolic Health and Aging; Ward "B": Unit of General Medicine and Advanced Care) and with discharge dates falling between 01/02/2024 and 31/01/2026. According to these filters, we obtained data of 1352 admissions for 1249 patients, with a median of 1 episode per patient and a maximum of 6 episodes per patient. The median episode length was 17.0 (12.0-

27.0) days comprising the days spent in the Emergency Room, when present, for a total of 31 thousand patient days. 54% of episodes belonged to Ward A. 43% of patients were female, with a median age of 76.0 (65.0–84.0) years, a median of 6.0 (4.0-9.0) chronic diseases per patient and a median Charlson Comorbidity Index of 6.0 (4.0-8.0) consistent with elderly and multimorbid patient population, as expected.

Data completeness analysis

We then analysed the frequency of missing values across a subset of manually recorded variables considered of highest importance. Missingness was not uniformly distributed across variables and across wards (Figure 3).

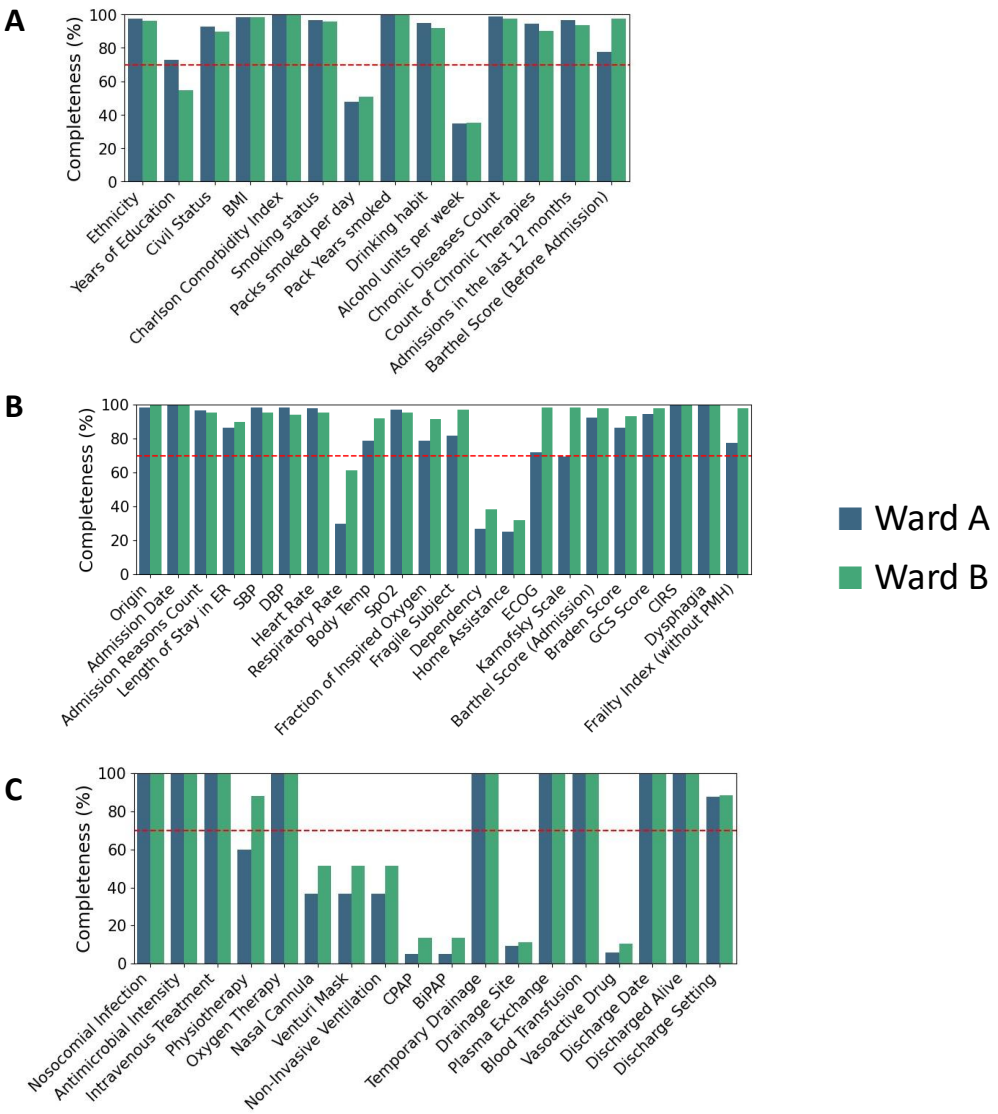

Figure 3 - ***Completeness of manually collected clinical variables across the two wards of interest.*** *Bar plots showing the percentage of data availability for manually collected variables across the study dataset. Blue: Ward A; Green: Ward B. Variables are grouped by collection timepoint: baseline characteristics (A), data on admission (B), and in-hospital or discharge-related variables (C). The red dashed line marks the 70% completeness threshold, below which data are considered of limited usability for robust statistical analysis. Most variables demonstrated high completeness, while a few—such as respiratory rate, Karnofsky performance status, and alcohol intake—fell below the threshold.*

Baseline patient characteristics (Figure 3A) exhibited high completeness, with value availability exceeding 95% for most variables. Exceptions were noted for three parameters (years of education, pack-years smoked, and number of alcohol units per week) all of which showed completeness below 70%. Admission-related data (Figure 3B) also demonstrated good availability, albeit with slightly more variability. Most variables maintained completeness above 90%, with a progressive decline observed for the Eastern Cooperative Oncology Group (ECOG) performance status, the Karnofsky

scale and particularly the respiratory rate, especially for Ward A. Finally, information collected during hospitalization and at discharge (Figure 3C) showed the lowest overall completeness.

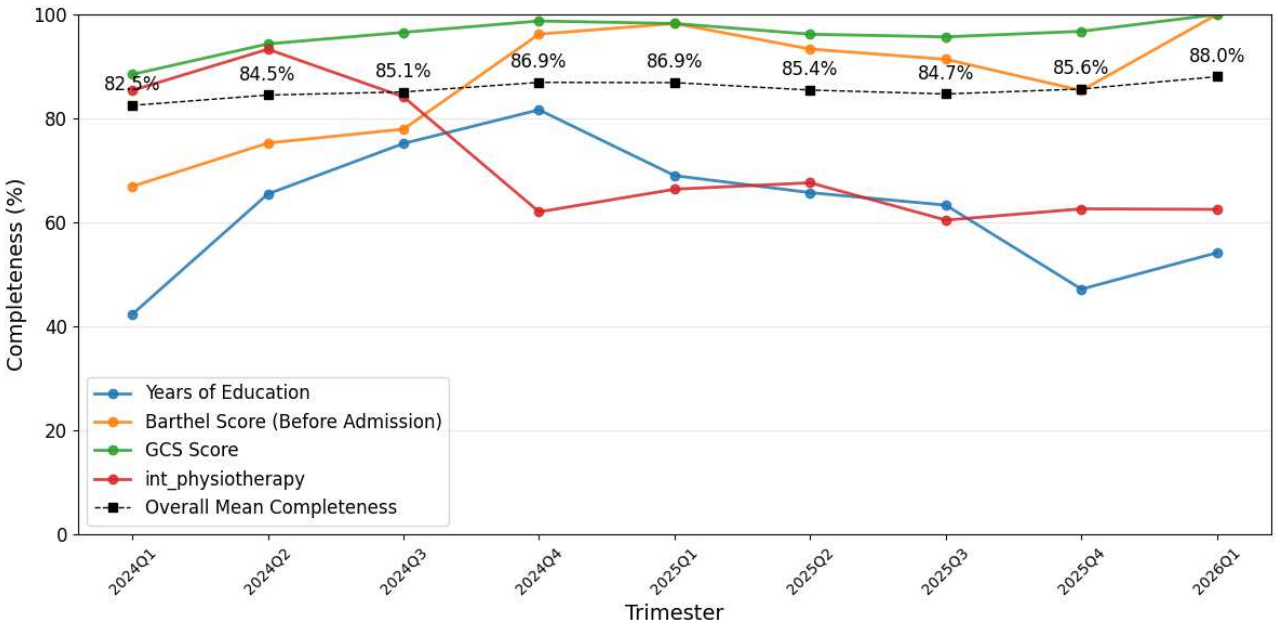

Figure 4 - ***Trimester-wise data completeness for selected variables and overall mean.*** *Completeness percentages across trimesters showing different patterns for individual variables (Blue: Years of Education, Orange: Barthel Score before admission, Green: GCS Score, Red: physiotherapy intervention), together with the overall mean completeness across all variables. Values are expressed as percentages, with higher values indicating lower missingness; point labels report trimester-specific overall mean completeness.*

It must be noted that, due to the interactive nature of the eCRF, certain fields are available for input only under specific conditions (e.g. “CPAP” and “BiPAP” fields appear only when “Non-Invasive Ventilation” is flagged as “Yes” and this latter field itself is available only when “Oxygen Therapy” is flagged as “Yes”). When assessed across trimester, mean variable completeness improved from 82.5% of the first trimester of 2024 to 88.0% of the first trimester of 2026, with individual variables showing different patterns of completeness in time (Figure 4).

Data Distribution Analysis

A further data quality control step consists in checking the distribution of variables and assessing for the presence of outliers. A commonly utilised nonparametric method to detect outliers is that of applying Tukey’s fences to the distributions. Any data point outside the fences is considered an outlier[28]. As health data

commonly presents skewed distributions and some patients can have measurements that naturally fall quite far from the median of the distribution, we selected a k = 2.5 to detect data points that could potentially constitute data entry mistakes, according to our previous experience[29]. Table 1 shows number and percentage of outlier for selected variables together with the automatically computed upper and lower bounds. Across 49 variables, the mean percentage of outliers was 2.8%.

Table 2 – Full report of the data missingness analysis for selected numerical variables of interest in the MED-Cli database with freeze as of 31st January 2026

Variable	N non missing	N of Outliers	% of Outliers	Lower Bound	Upper Bound
Admissions in the last month	1246	243	19.5	0	0
GCS Score	1297	160	12.3	15	15
GGT	516	53	10.3	0	430
proBNP	136	13	9.6	0	16551
ALT	1240	104	8.4	0	110
AST	1241	103	8.3	0	128
ALP	450	35	7.8	0	483
Alcohol units per week	473	36	7.6	0	32
Creatinine	1292	81	6.3	0	4
Glycemia	1111	44	4	0	284
Packs smoked per day	663	23	3.5	0	4
Eosinophils (%)	1231	42	3.4	0	6
Pack Years smoked	1352	41	3	0	122
LDH	1125	34	3	0	708
episodeLength	1352	37	2.7	0	68
WBC	1308	31	2.4	0	31
Barthel Score (Before Admission)	1175	28	2.4	4	169
Admissions in the last 12 months	1287	30	2.3	0	4
SpO2	1302	29	2.2	84	108
BMI	1330	27	2	6	43
Lymphocytes (%)	1231	25	2	0	50
Neutrophils (%)	1231	22	1.8	26	126
Respiratory Rate	598	10	1.7	2	36
Monocytes (%)	1231	21	1.7	0	22
Basophils (%)	1231	18	1.5	0	1
Body Temp	1144	17	1.5	34	39
Potassium	1279	18	1.4	2	7
Sodium	1283	16	1.2	120	156
Albumin (%)	227	2	0.9	22	78
Admissions in the last 6 months	1230	11	0.9	0	4
Hemoglobin	1308	10	0.8	1	22
CRP	1238	5	0.4	0	435
Platelets	1308	5	0.4	0	678
SBP	1312	3	0.2	35	215
Heart Rate	1308	3	0.2	12	162

Years of Education	875	2	0.2	0	26
Length of Stay in ER	1189	2	0.2	0	12
Admission Reasons Count	1299	3	0.2	0	4
Age	1352	1	0.1	18	132
DBP	1305	1	0.1	14	128
Barthel score (Discharge)	912	0	0	0	225
CIRS	1352	0	0	0	28
Total proteins	258	0	0	29	85
Braden Score	1210	0	0	0	40
Barthel Score (Admission)	1281	0	0	0	288
Chronic Diseases Count	1330	0	0	0	22
Charlson Comorbidity Index	1352	0	0	0	19
Count of Chronic Therapies	1252	0	0	0	25
Mean percentage of outliers			2.8		

Visual inspection via boxplots of collected data provides a quick measure of the skewness of distributions and eventually the presence and direction of outliers. Figure 4 displays the distributions of four exemplary variables (see more in the Supplementary Materials). Pack Years smoked distribution shows important outliers in the order of thousands, clearly a result of data entry mistakes: very likely the number of cigarettes was inputted in place of the number of cigarette packs. Haemoglobin distribution shows very few outliers, possibly linked to unit of measurement change in the laboratory system (from g/dL to g/L). BMI distribution shows 27 values defined as outliers, but manual revision of them reveals values that are still in plausible range and should therefore not be accounted as outliers. Age distribution reveals a 0-year-old patient, an unexpected finding given the fact that this field is automatically populated with a call from the hospital's registry.

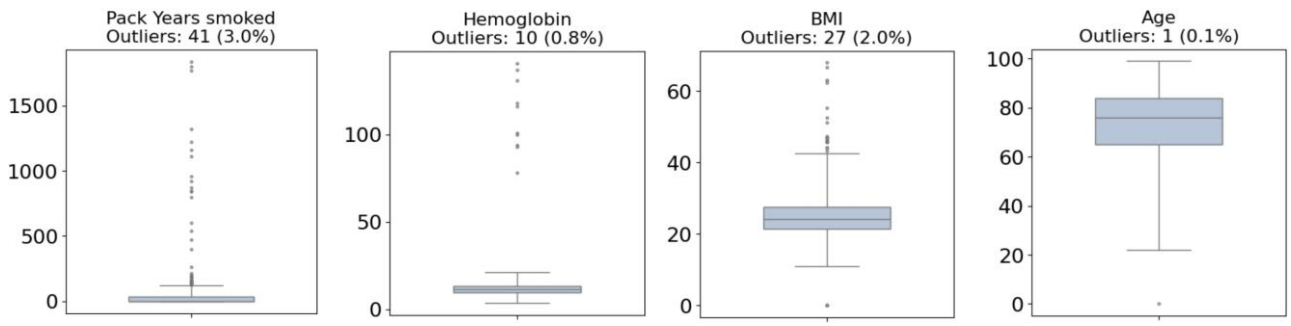

Figure 5 – ***Boxplots of data distribution of four exemplary variables.*** *Box-plots for Pack Years smoked, haemoglobin, BMI and age are reported, together with counts and percentage of automatically detected outliers via Tukey's fences method. Whiskers represent the lower and upper bounds set as 2.5 times the interquartile range. Circles represent data points labelled as outliers.*

PowerBI dashboards

Using Power BI, we developed interactive dashboards to visualize and explore real-world clinical data. Figure 6 illustrates one of the dashboards, featuring interactive bar plots, ring charts, cards, a scatter plot, and a density plot, alongside live slicers for on-the-fly filtering. Users can filter data by hospital ward (upper left corner), episode discharge date range (upper right corner), or the composite outcome condition (lower left corner), allowing for granular, user-driven analysis of patient outcomes and clinical trends. Our dashboards enable dynamic, real-time exploration of key clinical indicators, including monthly admissions and discharges, length of stay in the emergency department and wards, patient demographics, clinical complexity (e.g., comorbidities and chronic medications), distributions of vital parameters and laboratory values, and prevalence of discharge modalities and diagnoses (Supplementary Figures 1-2-3). Additionally, they enable quick and real-time data quality control checks, and per-patient or overall data missingness analysis (Supplementary Figures 4-5).

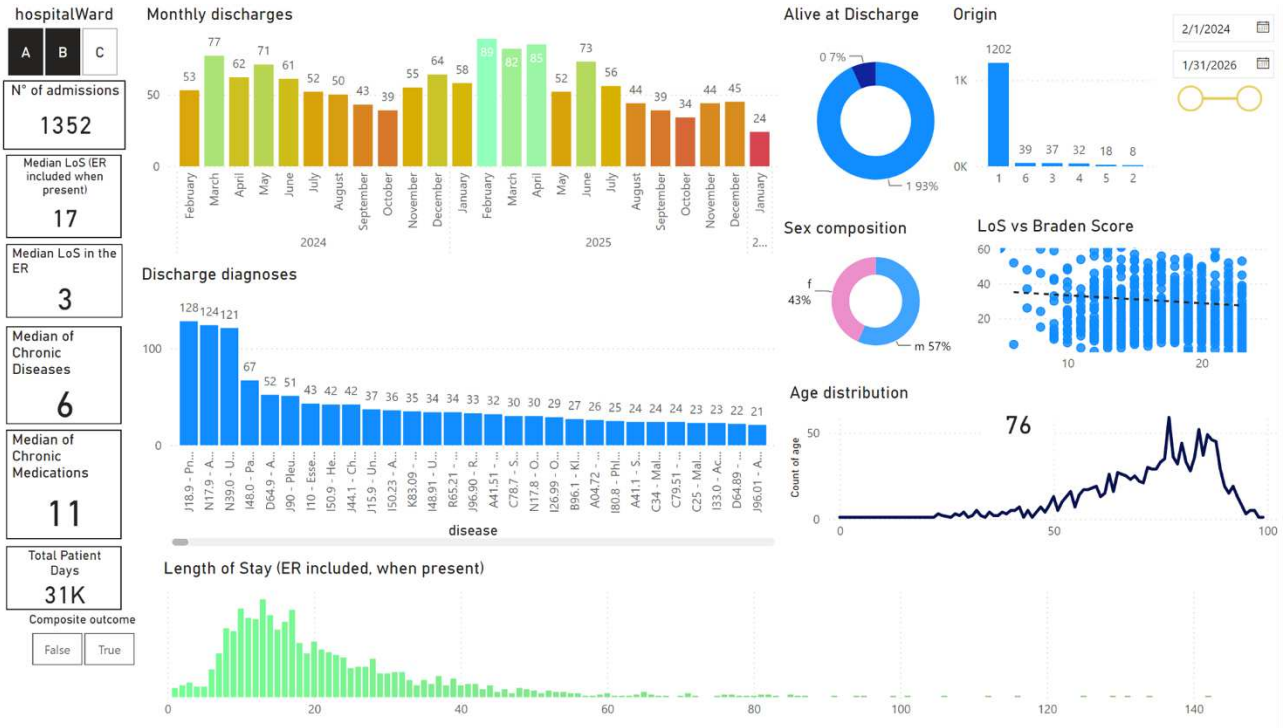

Figure 6 – ***Example of a PowerBI dashboard stemming from our Real-World Data collection pipeline.*** *We used Microsoft PowerBI to transform raw Real-World Data into a wide set of easy-to-navigate visualizations such as bar plots, pie/ring charts, scatter plots, density plots and cards. We added slicers to allow for live filtering of the data directly on the dashboard according to the hospital ward where the episode was recorded (upper left corner), the episode discharge date range (upper right corner) or the composite outcome condition being true or false (lower left corner).*

DISCUSSION

The present manuscript provides a comprehensive account of the methodology we implemented to continuously collect high-quality real-world data (RWD) from patients admitted to internal medicine wards at an Italian university hospital. This initiative is aimed at building a large, structured data lake to support future research and generate real-world evidence (RWE) that can inform healthcare professionals, hospital administrators, and policymakers alike. We here also share a data quality report on a database snapshot as of 31st January 2026 and display potential exploitation of the collected data to build visual dashboards for quick and live data analysis to foster data-driven medicine.

Importance of RWD Collection

This study is part of an ongoing ten-year observational project designed to systematically collect RWD on older inpatients. Its innovative design addresses a critical evidence gap, as older adults are often underrepresented in clinical trials, limiting the applicability of standard evidence-based interventions[14–16,30,31]. By collecting and analysing RWD in this population, the project aims to generate RWE that reflects the effectiveness of diagnostic and therapeutic strategies in routine clinical practice, beyond the selective context of randomized controlled trials. The infrastructure also enables the integration of routinely collected hospital data, including laboratory results, radiological reports and images, and administrative records, further enriching the clinical context and supporting longitudinal analyses. This is particularly relevant in internal medicine wards, where multimorbidity, frailty, and complex care trajectories are the norm[1,4].

Reducing Missing Values and Improving Data Collection

One of the key challenges in leveraging RWD for research and clinical decision-making lies in ensuring sufficient data quality, often affected by high proportions of missing values and outliers. Missing values not only reduce the statistical power of analyses but can also introduce bias and undermine the generalizability of predictive models. Although our interim analysis shows generally high completeness (Figure 4), certain variables remain suboptimally represented, particularly those related to functional status or lifestyle. These variables are gaining increasing relevance especially in internal medicine settings where a multidimensional evaluation of patients is of paramount importance[32–35].

Outliers alter the visualization of biomedical data distributions, the mean and standard deviation values and add noise during training of models possibly affecting their performance. Our interim analysis reveals a low percentage of outliers in our most relevant numerical variables (Table 2), a result obtained by adding validation rules to data fields in our eCRF (see materials and methods). Despite such countermeasure, we detected small numbers of outliers possibly stemming from unpredictable mechanisms such as unit of measurement changes or data errors in source databases, even affecting automatically retrieved information (Figure 5). Of note, the Tukey's fences method mathematically sets the lower and upper bounds outside of which datapoints are labelled as outliers. As previously discussed, we suggest for manual review of such bounds by clinicians, as biomedical data is often characterised by skewed distributions and values that, despite being far from the median of the distribution, are still plausible (see for example the fences set for the BMI)[29]. Additionally, Tukey's fences method fails when variables are characterised by values that are strongly centred around the median (see for example the fences of "Admissions in the last month" variable). Overall, our experience reflects a broader issue encountered in real-life clinical settings, where data acquisition often competes with limited time and high clinical workload, especially in internal

medicine wards managing complex, multimorbid patients. Achieving uniform data capture in such environments is inherently difficult. To mitigate these limitations, and according to our continuous data quality audits, we have adopted a set of measures, including the use of structured electronic case report forms with automated validation rules, ongoing training of healthcare personnel involved in data entry, and real-time monitoring systems to flag variables with high missingness rates. These strategies are already showing promising results, as shown by our overall completeness percentage of our most important variables rising from 82.5% to 88.0%, and represent an essential step toward the creation of a reliable, high-density data infrastructure.

CONCLUSION

This study presents a structured organizational model for the systematic collection and integration of real-world data (RWD) in internal medicine wards, with the aim of enabling future contributions to the understanding of clinical complexity and care trajectories in older inpatients. By building a scalable, high-quality data infrastructure, we seek to capture the interplay of multimorbidity, polypharmacy, frailty, and discharge barriers in a real-world context.

The growing literature on RWD underscores the need for advanced analytical tools and dynamic modelling to extract meaningful, actionable evidence[6,8,17,26,29,31,36]. Our infrastructure lays the groundwork for such developments by integrating clinical, functional, and administrative data to improve care delivery and inform evidence-based policy.

Ultimately, RWD can help shift the focus from treating conditions to caring for vulnerable individuals, ensuring that hospital care remains not only effective, but also ethically and humanely sustainable[31].

DECLARATIONS

Ethics approval and consent to participate. This study was approved by the institutional review board of IRCCS San Raffaele Scientific Institute with the code “MED-Cli”, and is registered at clinicaltrials.gov (NCT05780099; ethics approval 09 June 2021; study start 24 June 2022; first submitted to ClinicalTrials.gov on 21 February 2023 and first posted on 22 March 2023, i.e., registered retrospectively). All participants provide written informed consent prior to inclusion. This study was conducted in accordance with the Declaration of Helsinki.

Availability of data and materials. The data underlying this study were collected as part of a clinical trial. Participants provided informed consent that does not permit public sharing of individual-level data. Accordingly, the dataset is not publicly available. Researchers may request access to the data for specific scientific purposes; such requests will be considered on a case-by-case basis and, where appropriate, made available through a research agreement consistent with the participants' consent and applicable ethical and legal requirements.
The Python code used to query the study database, process the data, and perform the statistical analyses will be made publicly available upon publication of the final manuscript. In contrast, the source code of the Cohort Genomic Platform as well as the code underlying the electronic case report form itself, constitutes proprietary software developed and maintained by the study institution and is therefore not publicly available.

Competing interests. The authors declare no relevant conflict of interest to the aims of this study.

Funding. The authors received no financial support for the research, authorship, and/or publication of this article.

Authors' contributions. MM: Investigation, Data curation, Methodology, Writing - Original Draft; RDL, GAR, SD: Conceptualization, Investigation, Data curation, Methodology, Writing – Review & Editing; PRQ, MT: Conceptualization, Supervision, Writing – Review & Editing; EB, LL, GP, CP, ER, CS, GB: Investigation, Data curation, Writing – Review & Editing; CFS, VG, CM, GS, GT: software, Writing – Review & Editing.

Aknowledgements. The present work was performed by Marco Montagna in partial fulfilment of the requirements for obtaining the PhD degree at Vita-Salute San Raffaele University, Milano, Italy. The authors want to thank all the undergraduate students, residents and consultants that are taking part in the data gathering effort as members of the MED-Cli study group: Giulia Abatianni, Pietro Actis, Sara Angiulli, Maddalena Atzori, Costanza Bagnati, Laura Baldo, Giulia Barone, Alessandro Barruscotti, Silvia Belforte, Chiara Bellino, Francesca Biasco, Alberto Bongiovanni, Maristella Bonomo, Elisa Borgonovo, Giulia Buonsante, Chiara Calabrese, Valentina Canti, Filippo Canzi, Agostino Carpentiere, Francesco Cattelan, Andrea Cavazzana, Filippo Chiabrando, Francesca Conti, Federica De Nicola, Gaia Deonette, Carla Desapedroso, Giuseppe Di Lucca, Gaetano Di Terlizzi, Vittorio Dibenedetto, Alberto Domenighini, Mariagrazia Dorazio, Giusy Faggiano, Francesca Farina, Costanza Festorazzi, Serena Finocchiaro, Caterina Focacci, Maria Frazzetto, Agrimakarun Gera, Graziana Grimaldi, Thomas Hill, Elias Kheir, Martina Laffranchi, Giulia Lanzetta, Irene Limonta, Francescaanna Liso, Giulia Lucchini, Vanessa Maffi, Chiaramaria Maggiore, Emanuela Manzi, Giulia Marazzi, Alessandro Marinosci, Sara Menchi, Aurora Merolla, Anna Mirizzi, Emma Montano, Alessandro Montefusco, Lavinia Mulas, Debora Onorati, Giacomoleo Paolazzi, Francesca Paolelli, Silvia Pavani, Rita Peracino, Costanza Piccolo, Angelo Racioppi, Mariapia Ruggiero, Rabia Saeedbhatti, Chiara Salmaggi, Simona Santoro, Filippo Sciammetta, Martina Sculco, Marta Secci, Daniele Soddu, Antonella Spano, Federica Traversari, Despoina Tzinieri, Rita Venezia, Paolo Vergani,

Laura Vetrugno. We also would like to thank our present and past study coordinators/data managers: Marina Biganzoli, Federica Guerra, Alessandra Lucini, Gabriele Davide Mogliarisi, Ilaria Vergine.

Authors' information. MM is a consultant in internal medicine and holds a PhD in health data science at Vita-Salute San Raffaele University, Milan where he is currently a postdoctoral research associate in the AI team. He sits in the Working Groups in Digital Health of both the Italian Society of Internal Medicine and the European Federation of Internal Medicine.

REFERENCES

1. Naik, H. *et al.* Population-Based Trends in Complexity of Hospital Inpatients. *JAMA Intern. Med.* **184**, (2024).
2. RACGP aged care clinical guide (Silver Book) 5th edition Part A. Palliative and end-of-life care General principles.
3. Colacci, M. *et al.* Patient Complexity, Social Factors, and Hospitalization Outcomes at Academic and Community Hospitals. *JAMA Netw. Open* **8**, (2025).
4. Ceriani, E. *et al.* COmplexity of CARE and Discharge barriers: the 'modern internal medicine patient'. Results from the CO-CARED Study. *Intern. Emerg. Med.* **20**, 471–479 (2024).
5. Kennedy-Martin, T., Curtis, S., Faries, D., Robinson, S. & Johnston, J. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. *Trials* **16**, (2015).
6. Dreyer, N. A. Strengthening evidence-based medicine with real-world evidence. *Lancet Healthy Longev.* **3**, e641–e642 (2022).
7. Tenny, S. & Varacallo, M. A. Evidence-Based Medicine. *StatPearls* https://www.ncbi.nlm.nih.gov/books/NBK470182/ (2024).
8. Subbiah, V. The next generation of evidence-based medicine. *Nature Medicine 2023 29:1* **29**, 49–58 (2023).
9. Gregg, E. W. *et al.* Use of Real-World Data in Population Science to Improve the Prevention and Care of Diabetes-Related Outcomes. *Diabetes Care* **46**, 1316–1326 (2023).

10. Barlean, B. *et al.* 297 Identifying Opportunities and Challenges for Translational Informatics Approaches to Real-World Data: A Diabetes Case Study. *J. Clin. Transl. Sci.* **7**, 89–89 (2023).
11. Tinetti, M. E., Bogardus, S. T. Jr. & Agostini, J. V. Potential Pitfalls of Disease-Specific Guidelines for Patients with Multiple Conditions. *New England Journal of Medicine* **351**, 2870–2874 (2004).
12. Inouye, S. K., Studenski, S., Tinetti, M. E. & Kuchel, G. A. Geriatric syndromes: clinical, research, and policy implications of a core geriatric concept. *J. Am. Geriatr. Soc.* **55**, 780–791 (2007).
13. Tinetti, M. E. & Fried, T. The end of the disease era. *American Journal of Medicine* **116**, 179–185 (2004).
14. Yang, C. *et al.* Bridging the age gap: improving representation of older adults in clinical trials for type 2 diabetes and chronic kidney disease. *Lancet Healthy Longev.* **6**, (2025).
15. Tan, Y. Y. *et al.* Comparing clinical trial population representativeness to real-world populations: an external validity analysis encompassing 43 895 trials and 5 685 738 individuals across 989 unique drugs and 286 conditions in England. *Lancet Healthy Longev.* **3**, e674–e689 (2022).
16. Iannone, P. *et al.* Wrong guidelines: why and how often they occur. *BMJ Evid. Based. Med.* **22**, 1–3 (2017).
17. Concato, J. & Corrigan-Curay, J. Real-World Evidence - Where Are We Now? *N. Engl. J. Med.* **386**, 1680–1682 (2022).
18. Batko, K. & Ślęzak, A. The use of Big Data Analytics in healthcare. *J. Big Data* **9**, 1–24 (2022).
19. Liu, F. & Demosthenes, P. Real-world data: a brief review of the methods, applications, challenges and opportunities. *BMC Medical Research Methodology 2022 22:1* **22**, 1–10 (2022).
20. Dang, A. Real-World Evidence: A Primer. *Pharmaceut. Med.* **37**, 25 (2023).
21. Framework for FDA's Real-World Evidence Program. www.fda.gov (2018).
22. Hulsen, T. *et al.* From big data to precision medicine. *Frontiers in Medicine* vol. 6 Preprint at https://doi.org/10.3389/fmed.2019.00034 (2019).
23. Routinely Collected Health Data - MeSH - NCBI. https://www.ncbi.nlm.nih.gov/mesh/2052123.
24. Lo, J. J. M., Graves, N., Chee, J. H. & Hildon, Z. J. L. A systematic review defining non-beneficial and inappropriate end-of-life treatment in patients with non-cancer diagnoses: theoretical

development for multi-stakeholder intervention design in acute care settings. *BMC Palliat. Care* **21**, (2022).

25. Brucato, A. *et al.* Three-month mortality in permanently bedridden medical non-oncologic patients. The BECLAP study (permanently BEdridden, creatinine CLearance, albumin, previous hospital admissions study). *Eur. J. Intern. Med.* **72**, 60–66 (2020).
26. Traverso, A. *et al.* Powering responsible artificial intelligence with high-quality real-world data: the S-RACE platform for scalable, multi-specialty clinical research. *NPJ Digit. Med.* **9**, (2026).
27. Charlson, M. E., Pompei, P., Ales, K. L. & MacKenzie, C. R. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *J. Chronic Dis.* **40**, 373–383 (1987).
28. Tukey, J. W. *Exploratory Data Analysis by John W. Tukey*. *Biometrics* vol. 33 (1977).
29. Montagna, M. *et al.* From raw data to actionable insights: preprocessing real-world data for machine learning in diabetes care. *Front. Digit. Health* **8**, 1685842.
30. Savare', L. Enhancing the role of real-world data in healthcare research through advanced statistical methods. https://www.politesi.polimi.it/handle/10589/221954 (2024).
31. Mitchell, A. *et al.* Call to collaborate on data science for older people across Europe: an EuGMS Big Data Special Interest Group position paper. *European Geriatric Medicine* vol. 16 Preprint at https://doi.org/10.1007/s41999-025-01276-y (2025).
32. Fried, L. P., Ferrucci, L., Darer, J., Williamson, J. D. & Anderson, G. Untangling the Concepts of Disability, Frailty, and Comorbidity: Implications for Improved Targeting and Care. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences* vol. 59 Preprint at https://doi.org/10.1093/gerona/59.3.m255 (2004).
33. Kim, D. H. & Rockwood, K. Frailty in Older Adults. *New England Journal of Medicine* **391**, 538–548 (2024).
34. Damanti, S. *et al.* Frailty index, frailty phenotype and 6-year mortality trends in the FRASNET cohort. *Front. Med. (Lausanne).* **11**, 1465066 (2024).
35. Nicolaus, S., Crelier, B., Donzé, J. D. & Aubert, C. E. Definition of patient complexity in adults: A narrative review. *Journal of Multimorbidity and Comorbidity* **12**, (2022).
36. Hou, J. *et al.* Generate Analysis-Ready Data for Real-world Evidence: Tutorial for Harnessing Electronic Health Records With Advanced Informatic Technologies. *J. Med. Internet Res.* **25**, (2023).

Supplementary Files

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

- Supplementarymaterials.docx