Probabilistic Approaches for Data Integration in Biomedical Research

A thesis submitted to The University of Manchester for the degree of

Doctor of Philosophy

in the Faculty of Biology, Medicine and Health

2022

Alexia Sampri

School of Health Sciences

Division of Informatics, Imaging and Data Sciences

‘Blank page’

Table of Contents

Final word count (including footnotes, endnotes): 53,295

List of Figures

List of Tables

Abbreviations

The following table describes the significance of various abbreviations and acronyms used throughout the thesis.

Abbreviation	Meaning
ALMS	Aspreva Lupus Management Study
ALMS-I	ALMS - induction
ALMS-M	ALMS - maintenance
AZA	azathioprine
BILAG	British Isles Lupus Assessment Group
BMI	Body Mass Index
COMET	Core Outcome Measures in Effectiveness Trials
Complete Records	complete case analysis
COS	core outcome sets
Cov	coverage of confidence intervals
D10_diffsize	10 study study datasets, different number of individuals per dataset
D2_N1000	2 study datasets, 1000 individuals per dataset
D2_N200	2 study datasets, 200 individuals per dataset
D5_diffsize	5 study datasets, different number of individuals per dataset
D5_N1000	5 study datasets, 1000 individuals per dataset
D5_N200	5 study datasets, 200 individuals per dataset
EHR(s)	Electronic Health Records
EmpSE	mean/average empirical standard error
EXPLORER	Exploratory Phase II/III SLE Evaluation of Rituximab
FCS	fully conditional specification (multiple imputation)
FCS3group2	granularity's FCS imputation model excludes its categorical informative variable; mixed type's FCS imputation model includes its categorical informative variable
FCSgroup	FCS includes categorical informative 'group' variable in imputation model
FCSgroups	granularity's FCS imputation model includes both categorical informative variables; mixed type's FCS imputation model includes both categorical informative variables
Full Data	reference model with complete data before content heterogeneity applied
i.i.d	independent and identically distributed data
ICD	International Statistical Classification of Diseases
ICD-10	International Statistical Classification of Diseases 10th edition
LUNAR	LUpus Nephritis Assessment with Rituximab
MAR	Missing At Random

continues on next page

Abbreviation	Meaning
MASTERPLANS	The MAximizing Sle ThERapeutic PotentiaL by Application of Novel and Stratified approaches
MCAR	Missing Completely At Random
MI	multiple imputation
MICE	multiple imputation by chained equations
MMF	mycophenolate mofetil
MNAR	Missing Not At Random
mSE	mean/average model standard error
MTX	methotrexate
OAV	Object Attribute Value
RITUX	rituximab
SCD	sudden cardiac death
SD	standard deviation
SLE	Systemic Lupus Erythematosus
SNOMED CT	Systematised Nomenclature of Medicine – Clinical Terms
UMLS	Unified Medical Language System
WBC	white blood cells

Abstract

Data generated by the numerous medical studies conducted worldwide have the potential benefit the scientific and patient communities by generating new knowledge about health, disease, and treatments. This promise is well recognised by research communities, but it remains the case that many biomedical datasets are underutilised. To realise the potential of such datasets, these must be integrated with other existing data, to generate large-scale, research-ready data resources. However, datasets are often heterogeneous in content – i.e., they capture different information, or capture overlapping information at different levels of granularity. Traditional approaches for integrating heterogeneous datasets focus on harmonisation: they limit the combined dataset to information that was captured in all original datasets, which can be extremely wasteful. For instance, new biomarkers that were not measured in all original datasets may be left out of the combined dataset, and categorical data may be reduced to two or three levels, whilst some of the original datasets captured it in much more detail. We have developed new, probabilistic approaches for data integration, reducing content heterogeneity to a missing data problem, which is subsequently resolved with well-established multiple imputation methods. Subsequently we address three commonly occurring forms of content heterogeneity (i.e., variation in variables, varying granularity of categorical variables, and variation in variable types). For each form of content heterogeneity we first outline the theoretical solution using probabilistic approaches to data integration. Then, we evaluate the suggested solution through simulation studies. Finally, we illustrate the solution through application to real-world datasets from studies in Systematic Lupus Erythematosus. We also do this for combinations of different forms of content heterogeneity. The research in the thesis is methodological but with clear and direct application benefits.

Keywords: content heterogeneity; data integration; FCS; granularity; mixed type; multiple imputation; probabilistic methods; Systematic Lupus Erythematosus; systematically missing values

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

I. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the "Copyright") and they have given the University of Manchester certain rights to use such Copyright, including for administrative purposes.

II. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made **only** in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

III. The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the thesis, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

IV. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=24420), in any relevant Thesis restriction declarations deposited in the University Library, the University Library's regulations (see http://www.library.manchester.ac.uk/about/regulations/) and in the University's policy on Presentation of Theses.

Acknowledgements

There are many people that I would like to thank for supporting me throughout my PhD journey.

Firstly, I cannot begin to express my sincere gratitude and deepest thanks to my main supervisor Professor Niels Peek who took a leap of faith with me and has continually supported me with expertise, great insights, and unwavering support. I would never ask for a better supervisor than Professor Nophar Geifman who was always creative, to-the point, asking insightful questions and offering invaluable advice. This trip would not be real without Dr Philip Couch who was providing his expertise, and calming presence throughout this process. They taught me not only how to become a better researcher, but they also provided significant academic and emotional support throughout the process.

Secondly, a very warm thank you to all CHI members for all the PhD support and training. I would like to thank Professor Ian Bruce, Helen Le Sueur, and Dr Patrick Doherty. I would also like to thank the Maximizing SLE therapeutic Potential by Application of Novel and Systemic Approaches and the Engineering consortium and as well as the BILAG-br, EXPLORER, LUNAR and ALMS studies for making the data available for integration and subsequent research. I thank the Engineering and Physical Sciences Research Council for funding my PhD through the Centre for Health Informatics, University of Manchester.

I would like to thank all my friends in Manchester (and around UK of course) and in Greece who were there and supported me in my brightest and darkest moments. Thank you, Alex, Amy, Aspa, Camilla, Dimitra, Irini, Kostas, Maria, Marilena, Nasra, Rose, Valentina, Vasia, and Vicky. Your friendship has been very important to me.

My very special thanks to my Mancunian family – *Stefania* and *Boris* – for all the deep discussions about our future, crazy night outs and for helping take my mind off when I most needed it. Our funny moments and your positive vibes will be missed!

Στο τέλος, θα ήθελα να πω ένα τεράστιο ευχαριστώ από τα βάθη της καρδιάς μου στην οικογένειά μου.

Ευχαριστώ πολύ τις καλύτερες μου φίλες – τις αδερφές μου – *Μαριάννα* και *Φανή* – που δεν έχουν πάψει στιγμή να με νοιάζονται και να μου δείχνουν πόσο υπερήφανες είναι για μένα. Αυτό το ταξίδι δεν θα μπορούσε να γίνει πραγματικότητα χωρίς την εμπιστοσύνη, βοήθεια, και υποστήριξη που μου δείξατε και συνεχίζετε να μου δείχνετε. Ευχαριστώ πολύ τους – *Βάιο* και *Αποστόλη* – για την αμέτρητη υποστήριξη, τις καλοκαιρινές αναμνήσεις και τις όμορφες οικογενειακές εκδρομές.

Το πιο μεγάλο ευχαριστώ ανήκει στον προσωπικό μου ήρωα – την μαμά μου – *Κούλα*, που ήταν πάντα εκεί να μου δείχνει πως να κυνηγάω τα όνειρά μου, να μην φοβάμαι και να εμπιστεύομαι τις δυνάμεις μου. Καθ΄όλη τη διάρκεια των σπουδών μου, ήσουν εκεί ενθαρρύνοντας και εμψυχώνοντας κάθε μου προσπάθεια.

Αφιερώνω την παρούσα διδακτορική διατριβή στην οικογένειά μου και συγκεκριμένα στα ανιψάκια μου – *Βασιλική, Νεφέλη* και *Άγγελο*. Ήσασταν το στήριγμά μου σε όλο αυτό το ταξίδι και ιδιαίτερα στις δύσκολες στιγμές στην περίοδο της πανδημίας. Μου χαρίσατε αμέτρητες στιγμές γέλιου, συγκίνησης, αγάπης και συναισθηματικής υποστήριξης.

Many thanks to you all,

Alexia

Presentations arising from this thesis

Oral Presentations

1. **Sampri A**, Geifman N, Couch P, Peek N. Comparison of imputation methods that solve granularity problem resulting from healthcare structured data integration. 42nd Annual Conference of the International Society for Clinical Biostatistics, Bordeaux, France (virtual).
2. **Sampri A**, Geifman N, Couch P, Peek N. Using simulation studies to compare and evaluate traditional and probabilistic data integration approaches that solve missing variables problem in big biomedical and health datasets. Belgrade Bioinformatics Conference 2021, Bioinformatics and Data Mining of Biological Data (BiDMBD) (virtual).
3. **Sampri A**, Geifman N, Couch P, Peek N. Probabilistic data standardisation of big heterogeneous Datasets in biomedicine. International Conference on Big Data in Management (ICBDM 2020), Manchester, UK (virtual).
4. **Sampri A**, Geifman N, Couch P, Peek N. Probabilistic data integration methods in healthcare to solve content heterogeneity: Systemic Lupus Erythematosus exemplar. Data Science Perspectives 2020, Newcastle Upon Tyne, UK.
5. **Sampri A**, Geifman N, Couch P, Peek N. Probabilistic data aggregation in healthcare data in order to solve content differences. 41st Annual Conference of the International Society for Clinical Biostatistics 2020, Krakow, Poland (virtual).
6. **Sampri A**, Geifman N, Couch P, Peek N. Challenges in the aggregation of biomedical datasets and probabilistic approaches to overcome representational heterogeneity: Systemic Lupus Erythematosus exemplar. Doctoral Academy PhD Conference 2019, Manchester, UK.
7. **Sampri A**, Geifman N, Couch P, Peek N. Challenges in the aggregation of biomedical datasets and probabilistic approaches to overcome representational heterogeneity - Systemic Lupus Erythematosus exemplar. Jodrell Bank Centre for Astrophysics - Machine Learning Workshop 2018, Manchester, UK.
8. **Sampri A**, Geifman N, Couch P, Peek N. Challenges in the aggregation of biomedical datasets and probabilistic approaches to overcome representational heterogeneity, IEEE Symposium on Computer-Based Medical Systems (CBMS) 2018, Karlstad, Sweden.

9. **Sampri A**, Geifman N, Couch P, Peek N. Representational heterogeneity after healthcare data integration. What is it? How to solve it traditionally? How to solve it probabilistically?. The Farr Institute of Health Informatics Research meeting 2017, Manchester, UK.

Poster Presentations

1. **Sampri A,** Geifman N, Couch P, Peek N. Using simulation studies to compare and evaluate traditional and probabilistic data integration approaches that solve mixed type variables and granularity problems in big biomedical and health datasets, The AI and Decision Intelligence Summit 2021, Altitude X, Manchester, UK.
2. **Sampri A,** Geifman N, Couch P, Peek N. Probabilistic data standardisation of big heterogeneous datasets in biomedicine. Methods for Evaluation of medical prediction Models, Tests and Biomarkers (MEMTAB) 2020 Symposium (virtual).
3. **Sampri A**, Geifman N, Couch P, Peek N. Probabilistic data integration methods in healthcare to solve content heterogeneity: Systemic Lupus Erythematosus exemplar. Data Science Perspectives 2020, Newcastle Upon Tyne, UK.
4. **Sampri A,** Geifman N, Couch P, Peek N. Probabilistic data integration methods in healthcare to solve content heterogeneity: Systemic Lupus Erythematosus exemplar". Young Statisticians Meeting 2020, Manchester, UK (virtual).
5. **Sampri A,** Geifman N, Couch P, Peek N. Challenges in the aggregation of biomedical datasets and probabilistic approaches to overcome representational heterogeneity, Healthcare Text Analytics Conference 2019, Cardiff, UK.
6. **Sampri A,** Geifman N, Couch P, Peek N. Challenges in the aggregation of biomedical datasets and probabilistic approaches to overcome representational heterogeneity, JBCA Machine Learning Workshop 2018, Manchester, UK.

About the Author

Alexia is originally from Greece and grew up in a family with ecological consciousness and love for nature. She soon developed an interest in engineering, which resulted in her joining the Environmental Engineering Dep. at Democritus University of Thrace. During her undergraduate studies she was part of the BETECO Lab where she focused on Big Data analysis, statistical modelling, infodemiology and Internet behaviour. Getting a first insight in research, she realised she wanted to obtain in depth knowledge and become familiar with data science. That led her to move in Scotland and complete a MSc in Big Data with distinction, at the University of Stirling. For her dissertation, she applied unsupervised learning techniques to predict water quality parameters using remote sensing data. Whilst in Stirling, Alexia got offered different PhD studentships, but she decided to undertake the exciting opportunity to join the Centre for Health Informatics in the University of Manchester. She started her PhD on Health Informatics which was supervised by Prof Peek, Prof Geifman and Dr Couch, and funded by EPSRC. Her research focused on probabilistic methods to integrate structured biomedical data.

During her PhD, she co-authored research papers, was management committee member in different European Networks, participated in training schools, completed Short Term Scientific Missions and presented her research on multiple international and national meetings, symposia and conferences. She was also involved in teaching of many UG and PG courses which led her to obtain a Fellowship by Advance HE. Her PhD work helped her build a wide range portfolio of data engineering, mining, and analytics tools. She has learnt describing a theoretical problem, designing a method that solves it, creating simulated data and analysing real-world big datasets. Alexia's goal is to pursue a career in researching concepts by aggregating environmental and health data, developing AI technologies, applying machine learning algorithms and learning more about epidemiology, population health data, missing data, and prediction modelling.

After PhD submission, Alexia started her new role as a research associate in Health Data Science at the Cardiovascular Epidemiology Unit, University of Cambridge. Not only this was the fulfilment of a dream, but it gave Alexia the opportunity to join a multidisciplinary team, be an independent researcher and collaborate with partners that help her be the best version of a scientist. Her role is on applied research to further our understanding of the longer-term effects of Covid-19 using whole population electronic health records.

Chapter 1: General Introduction

Worldwide, the data revolution is unlocking many new opportunities for answering biomedical and public health research questions [1]–[4]. The broad adoption of electronic health records (EHRs) and different cohorts promises to make an unprecedented amount of data available for: generating new insights into population health and care; leveraging heterogeneity of populations and settings and understanding how clinical trials results generalise to the real-world. Furthermore, ubiquitous technologies promise to reveal the rhythms of health and disease hitherto invisible to infrequent clinical observation.

The realisation of these opportunities requires that data from multiple resources to be combined and made available to cross-cutting research. Combining different datasets is essential to achieving better generalisability of results (creating evidence outside highly controlled research environments) [2], [5], [6], ensuring the validity of comparative research [7], [8], larger denominators (for answering specific questions, and studying small subgroups and rare conditions) [9], encouraging more efficient secondary usage of existing data [10], providing opportunities for collaborative and multicentre research [11], [12], and to taking advantage of heterogeneity (of populations, geographical environments, and healthcare services). To realise the potential of these datasets, they must be integrated with each other to generate large-scale and therefore, useful data resources.

However, in most cases, the process of integration is cumbersome and requires significant investment of time and effort. Harmonising large amounts of data from different sources is a challenge such as balancing data integration with information governance. Specifically, as we continue to add more data and information, the risk of person identifiability increases. Ethical, legal and restrictions connected with sharing or pooling of individual level data is a continuous issue in international research projects and networks [13]–[15]. Federated analytics offers a way to measure and improve the performance of federated learning models and it might be useful for health data where the need for privacy and accuracy is heightened significantly. A remaining issue though is the heterogeneity of data generated as they are not identically distributed across the data systems which adds complexity in terms of modelling, analysis, and evaluation. Moreover, privacy-preserving methods for federated learning can be challenging to rigorously assess due to statistical heterogeneity in data [16].

The promise of personalised medicine cannot be delivered without addressing complex questions that require data from multiple sources and settings. This promise is well recognised by research communities, but it remains the case that many biomedical datasets

are underutilised. Different datasets will likely have non-uniform data content, apply different levels of abstraction [17], making integration of data from different sources difficult. This phenomenon, called content heterogeneity, is the main focus of this thesis. Traditional approaches for data integration resolve content heterogeneity either by relying on better coding of information at source, which is rarely implemented adequately, or on the prospective alignment of datasets which tends to be wasteful, labour-intensive and often impossible to achieve completely.

An important distinction is between pre-alignment and post-alignment of variables, databases, and vocabularies. Our goal is to create a database designed for keeping biomedical data from different studies. Pre-alignment is when the first data source uses certain standards and ontologies, and we use the same ones to all the following study designs. Therefore, all the data from different studies will be easily integrated and used together. Pre-alignment shows how the data are going to align with existing data before the start of collection and capture. On the contrary, in post-alignment, at least two biomedical databases use different terminologies, different vocabularies, and different use of similar data. Therefore, they have to change their terminologies, ontologies, and vocabularies, and choose common standards when these data sources/databases need to be integrated. Data collections are post-aligned in the preferred standards and ontologies.

1.1 Representational Heterogeneity

Health datasets typically grow through decentralised processes in which data collecting organisations meet local data needs and there is no requirement to standardise their data representations. This results in a patchwork of diverse, heterogeneous databases, making it very difficult to create a single, integrated database that uniformly captures all the relevant information. Integration of data sources is then problematic, as there is no single representation to which each source can be translated – i.e. the datasets are not interoperable.

We distinguish four different types of representational heterogeneity. *Structural heterogeneity* refers to differently structured data in different databases. It occurs when databases have different schemas and are not presented in a unified and global schema that provides transparency. For instance, structural heterogeneity is a common problem when harmonising normalised data in Object Attribute Value (OAV) format which is commonly used in EHR systems. In this context, the ‘Object’ is a patient, the ‘Attribute’ a clinical variable that was measured/recorded, and the ‘Value’ that actual value. Each record always has a timestamp (date + time). When this format is used, records are never deleted or overwritten, only new records are added at the bottom. We can use it to record anything i.e. symptoms, diagnoses, lab measurements, referrals, medication, prescriptions etc. Each

patient has many rows (often hundreds to thousands) because each attribute is kept per row. For instance, if we analyse someone's blood in the lab and measure 20 biomarkers, this will lead to 20 rows being added (all for the same patient).

Naming or syntactic heterogeneity is characterised by the use of distinct lexical terms for the same semantic objects. For example, in the cardiovascular domain 'coronary artery disease', 'myocardial infarction', 'acute coronary syndrome', and 'NSTEMI' may be used to capture overlapping conditions and states.

Semantic heterogeneity occurs when the meanings of table names, field names, and data values across local databases are similar but not identical [18]. An example is sudden cardiac death (SCD) which generally means 'death within 24h of onset of symptoms with a cardiac cause' but in reality it is hard to assess this (e.g. people are often found at home after they have died). Therefore, different studies often have different rules to classify a death as SCD. Reported incidences of SCD vary due to differences in definitions and methodology between cohorts [19]. With severe forms of semantic heterogeneity, one cannot construct a single, global value set to which all original data can be mapped. And apart from these semantic differences that appear from data and metadata themselves, there may be more subtle semantic differences arising from variation in coding habits of clinicians [20].

Finally, *content heterogeneity* - on which is the focus throughout the thesis - occurs when data represented in one data source is not represented in another. Content heterogeneity may occur, for instance, because a study recruited only men and therefore did not record sex in its database. Such implicit data, which are typically constant over a dataset, are often derivable from metadata, creating the possibility to resolve the content difference with other studies. But other forms of content heterogeneity, such as differences in measured biomarkers, are not easily resolvable as the underlying data are simply missing. Another example is the use of different variable types across datasets. Sometimes these differences amount only to variation in granularity, e.g. one dataset groups age by 10 years while another has age as integer - a mixed type variable problem. It is possible to resolve such differences by considering the 'least common denominator' (in this case, an age grouping by 10 years) but results in a loss of information that quickly increases as more datasets are integrated.

It should be noted that all types of representational heterogeneity can exist independently of each other, although there are often connections between heterogeneity in naming, semantics, and content. Furthermore, all types of representational heterogeneity reduce the possibilities for uniform cohort selection and for confounder adjustment in a consistent way across different datasets, and are therefore relevant for this thesis.

At this point there should be a distinction between data integration, data harmonisation and data enrichment. We will use the term ‘data integration’ to refer to the process of storing different datasets in a single location (a database), without any syntactic or semantic changes. In this case, information can often not be captured with a single query but will have to be captured via multiple queries – unless the data are also harmonised. By ‘data harmonisation’ we will refer to the process of creating a single cohesive dataset that contains all the information that was captured in the different datasets. So, this means that all the information is expressed at the same level of granularity and has the same meaning throughout. In addition, all the information can be captured with a single query, and data’s origin is known. By the term ‘data enrichment’ we refer to the process of appending collected data with relevant context obtained from additional sources. This may be achieved by using external data sources [21]. Data enrichment could be also accomplished by integrating taxonomies, ontologies, and third-party libraries as a part of the data processing architecture. An example is when enriched real-world data studies combine primary data collected directly from physicians and patients with existing (secondary) data such as EHRs. Throughout the thesis, the focus is on data integration.

1.2 The traditional approach: striving for data-level harmonisation

The health informatics literature has traditionally considered the use of universal data standards a core requirement to solving representational heterogeneity problems, especially those related to naming and semantics [22]. Broad utilisation of models and standards for coding such as ICD [23], HL7v3 [24] , Read [25], GALEN [26], SNOMED [27], CaBig [28] would essentially eliminate all forms of representational heterogeneity, thus enabling data-level harmonisation across different data sources, regardless of where, when and by whom the data were recorded. In other words, this approach advocates the use of a fine-grained, shared representation language in which all information is expressed at the time of recording, and therefore the same representation language can be used to express integrated datasets at any time point in the future. Such integrated datasets can subsequently be analysed for scientific purposes, for instance to answer the questions regarding the benefit of a drug and the prognostic value of the new biomarker. Because the data are pre-aligned to a granular representation, analytical issues with respect to cohort selection and confounder adjustment can be addressed in the analysis of the integrated dataset with existing statistical tools such as imputation analysis.

Data standards are often used successfully in clinical registries to pre-align data that are recorded at different sites [29], but other successful examples are rare. When datasets use different standards and are therefore not pre-aligned, it is sometimes possible to achieve

data-level harmonisation through post-alignment, for instance using the UMLS Metathesaurus [30]. However, such an approach is often laborious and rarely yields complete results [9]. In cases of severe representational heterogeneity, striving for data-level harmonisation often leaves no other solution than ignoring both specific data sources and data items (variables) from the shared representation. For instance, in case of severe content heterogeneity, integrated datasets often cover just a small set of data items that are present in most source datasets and leave out the other items. But this will obviously diminish the potential to accurately answer research questions as it will reduce possibilities to harmonise cohort selection and confounder adjustment. Therefore, we may not be able to answer our research question at all, because none of the available datasets cover all the relevant information. Provided these models/standards are properly and universally implemented they can enable consistent indexing, storage, retrieval and aggregation of clinical data across specialties and sites of care. They also facilitate structuring of the medical record, thereby reducing variability in the ways clinical information is encoded.

Unfortunately, these standardisation efforts have largely failed. For example, after a decade of development, HL7v3 was launched in 2006 but to date the vast majority of healthcare messaging is still done using older versions of HL7 that are semantically weak [31]. SNOMED CT is widely recommended as clinical terminology but few EHR vendors have actually implemented it [32]. This failure represents a considerable waste of resources in the time and costs of defining coding standards and information models. At the same time there has been vast growth in data sources, which remain underutilised as a source of knowledge for improving healthcare. Waiting until they are fully harmonised and interoperable is wasting time and opportunities for discovery − especially if that wait is likely to be indefinite if not infinite.

1.3 A probabilistic approach to integration for inference with biomedical data

In this thesis we question the assumption that perfect, data-level standardisation is needed for inference with data that stem from different sources. We believe that more progress can be achieved by adopting a top-down approach to integration that starts with research questions rather than data collections. We argue for an alternative focus on the development of analytic methods that embrace the data turmoil, rely less on standardised data items, and have the capacity to process content heterogeneous data. We suggest probabilistic approaches to integration in which content equivalence is a matter of uncertainty rather than a dichotomy. We address three commonly occurring forms of content heterogeneity (i.e.,

variation in variables, varying granularity of categorical variables, and variation in variable types), and essentially cast them as missing value problems, enabling a solution through well-established methods for resolving missingness (multiple imputation).

1.4 Case study – real-world data

Our work is about realising scale. If we bring different datasets together, we increase scale and therefore increase the statistical power to detect important relationships, even if the object of research is a rare disease. An extract of real-world (rare disease) data from a biomedical database was available from the start of the PhD project. This dataset formed the basis for the design of our methodology and has been used for subsequent analyses.

1.4.1 Systemic Lupus Erythematosus (SLE)

Systemic Lupus Erythematosus (SLE) is a complex and poorly understood condition that can affect many body systems with symptoms ranging from mild to severe. Some common symptoms include: achy joints, obscure fever, skin rash, chest pain, fatigue, weight loss, sensitivity to light and swelling. It is a chronic autoimmune disease (the body's natural defence against illness and infection) of unknown aetiology and with prevalence, in general population, of approximately 20 to 150 cases per 100,000 people [33]. Diagnosis of SLE can be difficult and is typically based on a combination of clinical symptoms and laboratory tests. High levels of two specific types of antibodies, anti-nuclear and anti-phospholipid, combined with a set of typical symptoms are usually indicative of lupus [33]. These antibodies significantly increase the risk of cardiovascular disease, which causes a reduced life expectancy in SLE patients. As with other more common autoimmune conditions, such as rheumatoid arthritis, it is thought a combination of genetic and environmental factors is responsible for triggering lupus in certain people. There is no cure for SLE; treatments may include non-steroidal anti-inflammatory drugs, corticosteroids, immunosuppressants, hydroxychloroquine, and methotrexate, but often none of these lead to full remission of symptoms and all have side effects. Two recent medicines (rituximab and belimumab) may be used to treat severe lupus. These biologic therapies act on immune system to decrease antibodies' number in the blood. In the last decade biologic therapies have become available to treat SLE with promising results including full and sustained remission of symptoms in some patients.

1.4.2 The MAximizing Sle ThERapeutic PotentiaL by Application of Novel and Stratified approaches (MASTERPLANS)

The MAximizing Sle ThERapeutic PotentiaL by Application of Novel and Stratified approaches (MASTERPLANS) programme aims to improve care for SLE patients by taking a precision medicine approach to identifying groups of patients that respond to particular biologic therapies [34]. The goal of stratified and precision medicine is to provide patients with the best treatments (i) by ensuring that existing therapies are targeted at those who will derive most benefit and by accelerating the development of new therapies, and (ii) by better understanding why some patients respond to specific treatments while others do not.

The data integration work-strand in MASTERPLANS focuses on the bringing together various existing datasets from cohort studies and clinical trials involving either the conventional treatment with the immunosuppressants such as mycophenolate mofetil (MMF) or treatment with new biologic therapies such as Rituximab, and Belimumab. The aim is to identify biomarkers that either singly or in combination predict remission of symptoms in patients receiving these treatments; enabled by integration and harmonisation of relevant data from multiple sources. All the data from the studies related to the programme are kept in the MASTERPLANS data warehouse, implemented on the tranSMART platform. All data pertaining to the project have been uploaded into tranSMART to provide a single integrated data repository. Loaded data are restructured to fit a consistent format, to suit analysis requirements; this requires some degree of harmonisation between datasets and removal of redundancy among the data sources.

As we mentioned before, an extract of data from the MASTERPLANS database was available almost from the start of the PhD project. Those data formed the basis for the design of our methodology and have been used for subsequent analyses. These data currently comprised of three cohort studies i) Aspreva Lupus Management Study (ALMS) [35], [36], ii) The lupus nephritis assessment with rituximab (LUNAR) study [37] and iii) The Exploratory Phase II/III SLE Evaluation of Rituximab (EXPLORER) [38]. As the MASTERPLANS projects progressed in parallel to this PhD, additional datasets became available. The three studies were conducted across different sites, with different coding frames and employing different entry and exclusion criteria. Furthermore, study protocols, the evaluated medication, recruitment numbers, and duration of follow-up varied across the studies with only partial overlap.

The ALMS study included two phases: induction (ALMS-I) and maintenance (ALMS-M). ALMS is a prospective and randomised trial aimed to assess the efficacy and safety of MMF with patients with lupus SLE. The objective of the study was to compare long term MMF efficacy to azathioprine (ALMS-M) and MMF efficacy to cyclophosphamide (ALMS-I). In ALMS-I patients were followed for 24 weeks and in ALMS-M for 36 weeks.

The entry into Phase II was determined by response during Phase I. In contract LUNAR and EXPLORER's objectives were to evaluate the efficacy and safety of rituximab in randomised, placebo-controlled trials. They were also randomised at a ratio of 2:1 to receive rituximab (1,000 mg) or placebo on days 1, 15, 168, and 182 in both studies. In LUNAR, 144 patients with SLE took part and they were previously treated with MMF and corticosteroids. The primary end point was at week 52. In EXPLORER 257 patients took part with background treatment distributed among azathioprine, mycophenolate mofetil and methotrexate. Patients entered with >=1 British Isles Lupus Assessment Group (BILAG), a standard disease severity index in lupus, ranging from A to E) A score or >=2 BILAG B scores despite background immunosuppressant therapy, which was continued during the trial.

1.4.3 Representational heterogeneity in lupus studies

We have identified three main categories of representational heterogeneity discussed above (naming, content and semantic heterogeneity) in the lupus datasets. They are shown in Table 1.1.

Table 1.1. Examples of representational heterogeneity among ALMS, LUNAR, AND EXPLORER datasets.

Variable's description	**Variables' representation in lupus datasets**			
	ALMS		***LUNAR***	***EXPLORER***
	ALMS-I	***ALMS-M***		
Drug dose	Drug Dose per time Total drug dose per visit		Non-existence of variable	
Visit date	day/month/year		As a number that uses 0 as a baseline for the first day that the study started	
Ethnicity	Caucasian Asian Black Other (many subcategories)		White Black or African American Other	
Smoking Status	Y/N/NA		never/current/ex-smoker	
Year of SLE diagnosis	Year of diagnosis		Non-existence of variable	

- *Naming (or syntactic) heterogeneity*

Naming (or syntactic) heterogeneity occurs when different lexical terms have identical meaning. In the lupus datasets, we have smoking status 'Y' in ALMS and 'current' in LUNAR/EXPLORER which amounts to the same thing.

- *Semantic heterogeneity*

Semantic heterogeneity occurs when variables have identical or similar names but their meanings are different [18]. An example is the variable 'visit date' which occurs across different lupus datasets but its meanings are different. All the studies have different baseline data and in particular ALMS compared to LUNAR and EXPLORER have different relative visit time points for their patients (different time points relative to baseline). Therefore, when it comes to integrate patients' visit days (from baseline) we need to fully understand the study's design in order to apply a common starting point, following visits and end point. Another example is smoking status, in LUNAR and EXPLORER this variable includes the categories of ex-smoker and never smoker which do not directly and accurately map to 'N' as it captures in ALMS as they have different meanings.

- *Content heterogeneity*

Content heterogeneity occurs when data are not equally represented across all the datasets. For example, the year of SLE diagnosis is captured in ALMS-I and ALMS-M but not in LUNAR and EXPLORER. Another important difference relates to drug dose of medication used. As we see in Table 1.1, in ALMS there are variables to capture drug's dose per time and the total per visit. However, in LUNAR and EXPLORER there is no capture of drug dosage as this is likely uniform within each of these studies and may have been described in study protocols but not in the data themselves. Another example would be of a common variable, *Ethnicity*, which occurs in all three datasets (Table 1.1), however the subcategories which fall under this are different in each study. Furthermore, even the common subcategories such as 'Other', are likely to have different content (will include different ethnicities).

As discussed previously, all the aforementioned kinds of representational heterogeneity may exist with or without the existence of the others and it may be the case that several forms of representational heterogeneity exist alongside each other. In the lupus datasets, 'smoking status' includes all the forms of the representational heterogeneity.

1.5 Aim and objectives

Our over-arching aim is to explore the potential of probabilistic methods based on multiple imputation to address content heterogeneity in biomedical dataset integration. Specifically, we aim to assess the accuracy of statistical inference based on probabilistic integration of heterogeneous datasets via simulation studies, focusing on (i) the problem of variation in variables; (ii) the problem of varying granularity of categorical variables; (iii) the problem of variation in variable types; and (iv) combinations thereof. Furthermore, we aim to illustrate the application of these methods in real-world data on SLE.

1.6 Thesis structure

The overall structure of the thesis takes the form of seven chapters. Chapter 1 is the general introduction and Chapter 2 starts with a review of existing literature on data integration, representational heterogeneity, content heterogeneity, existing probabilistic machine learning/data mining methods, types of missing data and methods to address missing data. We also reviewed probabilistic methodologies to evaluate potential routes for efficient data integration. Then, four chapters follow where we investigate three commonly occurring forms of content heterogeneity. In particular, in Chapter 3 we present the problem of variation in variables, in Chapter 4 the problem of varying granularity of categorical variables and in Chapter 5 the problem of variation in variable types. For each of form of content heterogeneity we first outline the theoretical solution using probabilistic approaches to data integration. Then, we evaluate the suggested solution through simulation studies. At the end, we demonstrate the utility of the developed probabilistic approaches by applying them to real-world biomedical datasets - studies in Systematic Lupus Erythematosus. Chapter 6 addresses combinations of different forms of content heterogeneity and it follows similar format as the previous three chapters. The thesis finishes with a general discussion (Chapter 7) on the utility of probabilistic methods for data integration in biomedical research, limitations, future steps, and conclusions.

Chapter 2: Literature Review

2.1 Objectives

In order to be successful with the general scope of this thesis a robust review of current achievements in the field had to be carried out. Therefore, in this chapter we review the scientific literature on representational heterogeneity and methods for integration of heterogeneous datasets, with a focus on health data. Further, we explore which probabilistic methods have been used for integration of heterogeneous datasets. At the end we focus on the issue of content heterogeneity, types of missing data and techniques to deal with data missingness.

Therefore, our review has four objectives:

1. To review the issue of representational heterogeneity in health data;
2. To review methods for integration of heterogeneous health datasets;
3. To explore probabilistic methods for integration of heterogeneous datasets;
4. To identify content heterogeneity issues after data integration and explore potential methods to overcome them.

The chapter continuous with the search strategy to direct literature review (2.2). Results start with a discussion of different types of representational heterogeneity and a clarification of thesis' main aim (2.3.1). We continue with different forms of data integration (2.3.2), provide an overview of traditional data integration methods in biomedicine (2.3.3 and suggestion to alternatives i.e. probabilistic data integration methods (2.3.4). Subsequently, we describe in detail content heterogeneity (missing data (2.3.5), statistical methods to handle them (2.3.6.) and the problem of systematically missing values (2.3.7). The chapter ends with a discussion about literature review' findings and limitations (2.4).

2.2 Search strategy

To conduct the literature review, we have selected appropriate sources for literature that would be analysed and included in the final report. These included Google Scholar, Scopus, and the largest open biomedical database PubMed [39]. The main sources of information included here are published papers, specifically, peer reviewed articles and the scientific articles (that either were cited by others or not). There was not a specific minimum number of citations for a paper in order to be included because for some parts the provided literature

was limited. In some basic terminology that needed to be covered, only papers with a minimum number or 50 citations were included. In order to provide good coverage, contrast and a meaningful discussion, the selected papers covered different methodologies and opinions on the subjects covered by this review. Searching was not limited to specific editorials and publishing companies in order to avoid bias. A wide range of keywords and key phrases and synonyms were used to conduct the searches; these included: 'data integration', 'health data integration', 'data standardisation', 'probabilistic data integration', 'data standardisation in health', 'representational heterogeneity', 'health data harmonisation', 'integrating the healthcare', 'data integration in health care', 'medical data integration', 'uncertainty', 'uncertainty and data integration', 'missing data, 'systematic missing variables', 'and 'probabilistic approach in data integration'. Additionally, searches focused also on manuscripts published by authors that are prominent in the field of heterogeneous databases, data integration, probabilistic models for data integration in different fields and the field of biomedical data integration [17], [40]–[46].

Applying the criteria listed above, a snowballing approach was taken to identify a set of papers to be included; this process started with significant publications such as systematic reviews and had two phases for selecting relevant publications: backward and forward snowballing [47]. In backward snowballing, leading scientific papers' reference list was used to identify other published papers. In forward snowballing, articles that have cited the leading ones are identified. With both approaches, the reference lists were reviewed and used to decide which publications should be included. The preferred articles were chosen based on the following criteria (in addition to those outlined above): year of publication, number of citations (played an important role only to cover basic concepts and terminology), relativity, contribution, clarity, impact and correlation with other work. Once the papers were found, the abstracts (and in some cases, other parts of the paper) were used to decide on their inclusion or exclusion from this review. Figure 2.1. denotes the process of the literature search. During the electronic database search, 110 papers were identified using specific keywords (see above). From their reference lists another 70 papers were also included in the first collection of papers. A total of 18 papers were identified in both collections (mainly reviews with a large number of citations. Some of the initial papers were duplicated and were removed leaving a total of 155 papers. Of these, 30 papers were deemed irrelevant from reading their abstract. An additional 63 papers were removed after being found not useful for our review. This final step included reading papers in more detail and especially parts like introduction, discussion and conclusion

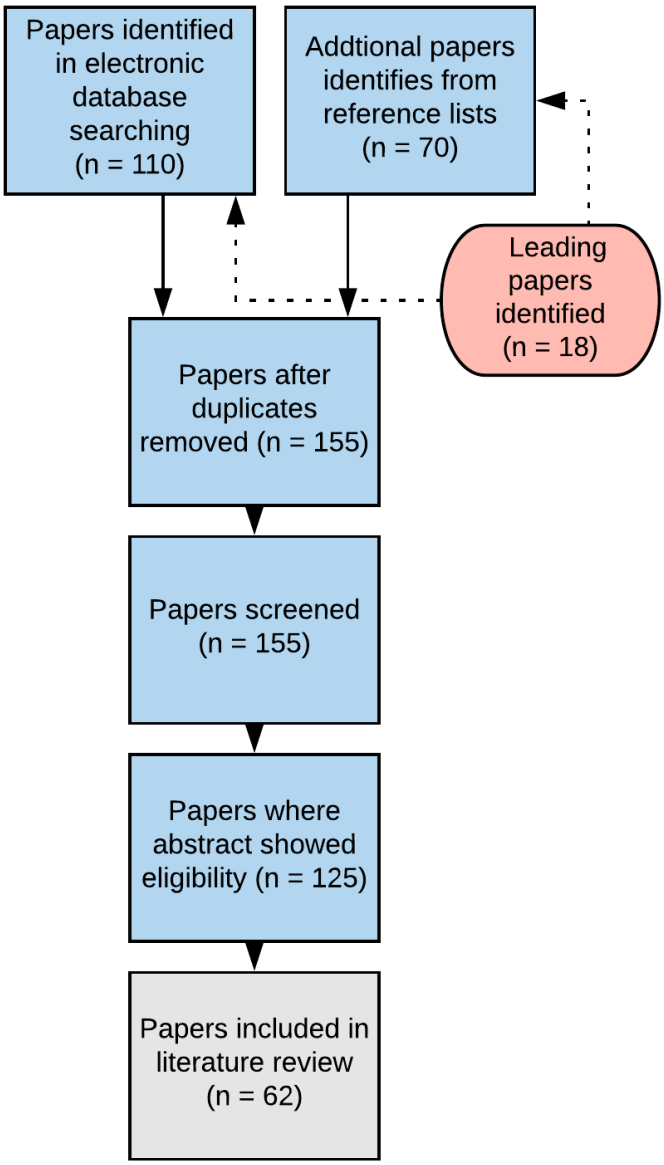


Figure 2.1. Methodology flowchart of the literature review.

2.3 Results

2.3.1 Representational Heterogeneity

In this section we give a detailed overview of subcategories of representational heterogeneity. We have identified four categories of representational heterogeneity: structural heterogeneity, naming heterogeneity, content heterogeneity and semantic heterogeneity. In Chapter 1, we briefly discussed and presented a few examples for the four heterogeneity types. In this section we describe them in more detail. As mentioned in section 1.4 in page 31 the overall goal of the thesis is overcoming content heterogeneity in the form of (i) variation in variables; (ii) varying granularity of categorical variables; (iii) variation in variable types; and (iv) combinations thereof resulting from data integration.

Firstly, we introduce a type of representational heterogeneity that was mentioned in the previous chapter i.e *structural heterogeneity*. It occurs when there are differences between different kinds of table decompositions such as horizontal and vertical, data - metadata representation, and structured or free text variables. Different 'horizontal table decomposition' requires different normalisation degree and leads to split the data across a changing number of tables [17]. Concerning normalisation degree, we mean different processes of data organisation in tables, to achieve elimination of redundancy, ensurance that data are in the correct table, and elimination of restructuring the database. Figure 2.2 (A) shows that doctor and patient relationship might be presented in at least 2 ways. Different 'vertical table decomposition' tables could happen by splitting rows in one or more tables. For example, Figure 2.2 (B) shows how inpatients and outpatients' records, of a hospital database, could be presented in at least two ways. With regards to data and metadata representation, relational models do not have to follow a specific hierarchy. Therefore, there is a variety of encoding data which leads to different usage of metadata. For a better understanding of this concept, see Figure 2.2 (C) which illustrates how blood tests results could be represented in at least 3 ways. Concerning free and structured records, there are differences in the allocation of the original data across many fields against a unique field. For example, the division and connection in how laboratory results and addresses are recorded. Some databases represent the address in a single data field: < 'Amy Smith, 123 Hulme Street, M18 6GB, Manchester' > while in others are represented as separated data fields: < 'Amy Smith', '123 Hulme Street', 'M18 6GB', 'Manchester'>.

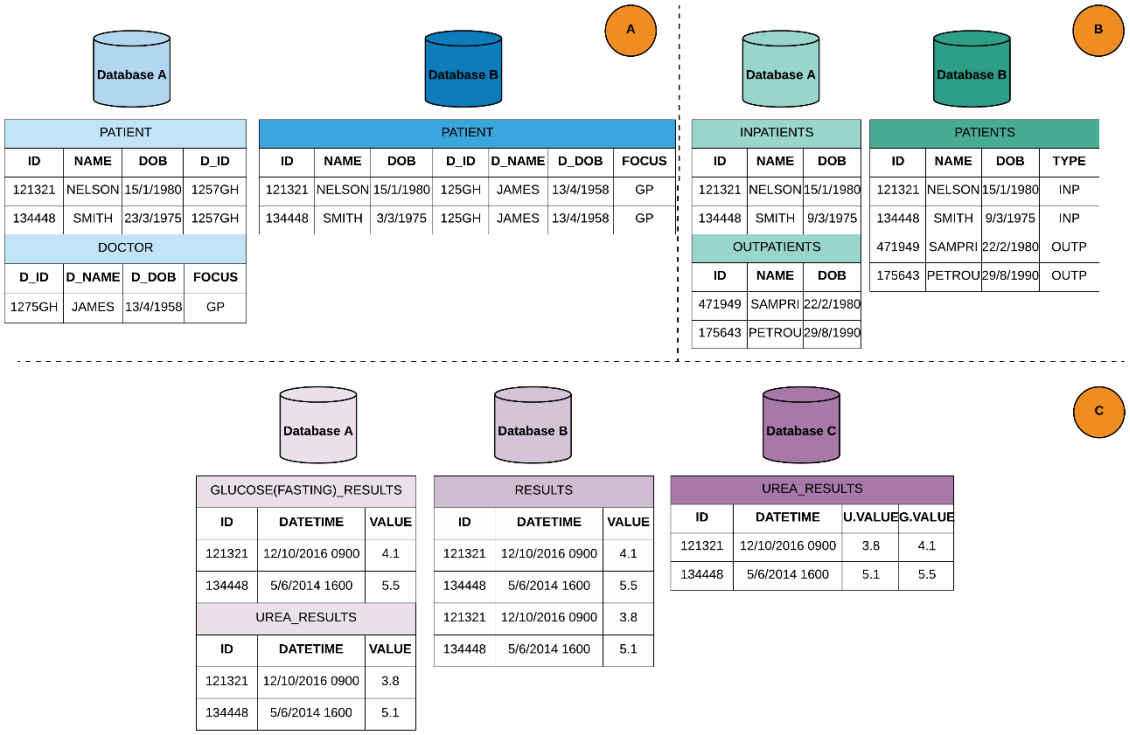


Figure 2.2. Different horizontal (A), vertical (B) table decompositions, and different encodings of a simple type hierarchy [17].

Secondly, as briefly discussed in Chapter 1, *naming or syntactic heterogeneity* occurs when same meanings are presented using different lexical terms. These differences in naming conventions could occur at the data level (e.g. in dataset A has 'Body Mass Index' and dataset B uses the abbreviation 'BMI') or at the metadata level (e.g. data are collected for dataset A in term 'sex' and in dataset B in a term called 'gender'). These examples are simple, and sometimes metadata naming differences could also indicate different meaning which may not be so easy to understand. This specific issue is not only syntactic, but it is also semantic in the metadata level. Therefore, it is often unclear and makes it difficult for different data sources to be integrated. For example, records for a patient's id may designate the social security number rather than medical record number. This example catches semantic differences (which will be explained later), and it is a time-consuming data integration process.

Thirdly, as described in section 1.1, *semantic heterogeneity* occurs when variables in different data sources have same terminology; sometimes it is described as variation in granularity. This occurs when variables of the data fields are kept under identical and similar variables but their meanings are different [18]. A simple example for semantic heterogeneity would be where one database uses the term 'cold' to mean flu like disease and where a second database uses the term 'cold' to mean temperature. Despite the semantic differences

among the datasets, in terms of data and metadata, another consideration comes from the coding preferences of clinicians [20]. Semantic heterogeneity can be divided into *one-to-many* and *many-to-many* relationship between similar variables. 'One-to-many' means that values of a variable of dataset A may be linked to many values of the same variable of dataset B but not the opposite. This means that, in 'one-to-many' a value of this variable of dataset B is linked *only* to one value of dataset A. For example, there are two databases that have the common variable 'Blood_growth'. Database A differs the growth level as: 'no growth', 'moderate_growth', 'significant_growth' while database B differs it as '0', '1', '2', '3', '4', '5'. Based on the example, we see that there is a semantic difference and 'one-to-many' relationship between the values of database A and database B Otherwise, in 'many-to-many' many values of a variable of dataset A may be linked to many values of the same variable of dataset B and vice versa. For example, database A differs the growth level as 'no_or_low_growth', 'moderate_or_significant_growth', while database B differs it as 'no_growth', 'moderate_growth', 'significant_growth'.

Lastly, as we discussed in Chapter 1, *content heterogeneity* characterises data that are not represented across all the data sources. Some data are recorded in one or in some databases, but they do not exist in others. Data might be inherent, deducible, or missing. 'Inherent' data mean that are often part of a local database and cannot be taken easily in global databases' environment. 'Deducible' data mean that one could be derived from the other (some information loss possibly occurs). An example for deducible data is the representation of postal code vs. town name, or use of birth date vs. use of age. A general example of content heterogeneity is when, a given study may focus on only males or only children while a second study may include both males and females or all age groups. 'Missing' data occur when no values are stored for some or all variables. The problem of missing data takes place when the global databases include data that are not represented in local databases. A missing data example occurs when medical centres' local databases do not record any variables for HIV status (for confidentiality issues) but in global and integrated databases the 'HIV' status is recorded. Therefore, the integrated database has null values for some patients which means that either the status is negative, or unknown, or unavailable. Another typical example of content heterogeneity is where patient's age is recorded as '<20', '20-40', '40-60', '>60' in dataset A, and recorded as an integer number in dataset B; here there is uncertainty about the ages of patients from dataset A when we try to express them as integer numbers (for integration with dataset B). Another similar example is when the age, in dataset A, is kept in range by 5 years and in dataset B is grouped by 10 years.

All the aforementioned kinds of representational heterogeneity may exist with or without the existence any of the others and it may be the case that several forms of representational heterogeneity are combined.

2.3.2 Data Integration

According to Sujansky [17], the process of data integration could be defined as *'the creation of a single uniform query interface to data that are collected and stored in multiple, heterogeneous databases'*. Data integration is a way of combining heterogeneous (in our case structured) data at the same time from multiple and different data sources [43]. To realise the potential of these datasets we need to build a consistent interface that gives the ability of querying to users.

There are several possible approaches for heterogeneous data integration such as vertical, horizontal, historical, longitudinal, integration for application portability, and cross indexing [46].

- *Vertical data integration* is the combination of data from different sources that do not have the same type or format, into one common database. This type of integration demands a methodical planning process of how to correlate and combine all the findings from the data sources. Each database may follow a specific hierarchy which needs to be maintained after the aggregation [43], [48].
- *Horizontal data integration* is the combination of data from similar data resources and applications with common format. Through this process, data follow a non-hierarchical way of integration and there is no particular order or preference [48].
- *Historical Integration* is used when data from different databases and potentially in different formats are combined but they also need further handling in order to be together [46]. Patient health records that come from many systems and in dissimilar formats could need much more processing time and effort to give a useful summary of information.
- *Longitudinal Integration* is the aggregation of data in a continuous manner. Health care data are captured all the time and there is a need to keep the records updated and to offer flexibility in terms of new data entries [46]. Also, our knowledge of treatments increases all the time, therefore it is of highest importance to find new ways to keep records and data constantly merged and without inaccurate information.
- *Cross Indexing Integration* is used when there are new records for a specific patient and we would like to add them to their existing record (e.g., medical history), and then aggregate records from the patient's family members. This case offers a great

opportunity of investigating similarities, differences and any correlations among generations and to understand in detail multiple exposures based on family history records.

In this thesis, instead of focusing on the data, we focus on the research question, and try to use all the information that is available in the individual study datasets to address that question, even if those individual datasets express that information in mutually incompatible ways. In the process of doing so, we do create a very large 'stacked' dataset that encompasses all the information from the individual study datasets. But this stacked dataset, by necessity, also has a lot of 'gaps' that are emerge because those datasets express information in mutually incompatible ways. We then estimate the distribution of data that can plausibly fill the gaps using established methods from the field of missing data, i.e. multiple imputation.

2.3.3 Traditional data integration methods in Biomedicine

In bioinformatics and in health informatics, attempts have been made to solve the problem of non-uniform data; researchers have considered that the complex problem of localised data collections could be solved by adopting more evolutionary and advanced techniques [22], [41], [49]. In the past, the problem of representation heterogeneity has been addressed through formal knowledge representation by developing semantic standards such as information methods and strict terminologies/vocabularies, taxonomies, ontologies, thesaurus and coding systems [17], [22], [44], [45], [50], [51].

First of all, we should understand the distinction between the different types of knowledge representation. An 'ontology' is a structural taxonomy with formalised relationships among concepts. Ontologies are used more to express complex and difficult relationships such as complex information systems (Semantic Web). According to Guarino, Oberle and Staab [52] ontology means *'to formally model the structure of a system, i.e., the relevant entities and relations that emerge from its observation, and which are useful to our purposes'.* The other distinction of a controlled vocabulary is a taxonomy. A 'taxonomy' is often presented by a hierarchical tree structure and the used variables are represented by nodes. 'Taxonomy' could be found as tree - vocabulary that is less complex than an ontology or a thesaurus. It could also be found as any kind of controlled vocabulary that is used more in web sites and enterprises than scientific libraries. A 'controlled vocabulary' is a systematised grouping of words and phrases with explanations and is used to recapture the meaning through searching. It does not have any particular structure, they are used by companies and persons. 'Controlled vocabulary' is a general category that includes thesauri, and taxonomies. A

‘thesaurus’ has a standard structure and its variables have connection among them. Thesauri also have notes and further explanations. They are created for certain eras and reasons.

The success of the Gene Ontology (GO) in standardising gene characteristics and functions has led to the spread of the usage of the term ontologies and especially in the biomedical sciences became very popular [53]. There are more than 200 biomedical ontologies and that number keeps increasing. Ontologies are now routinely used in biomedicine and GO is a classic source [54], [55]. Table 2.1, lists some of the existing universal data standards, vocabularies and ontologies such as ICD [23], SNOMED [27], HL7v3, Read [56], GALEN [26], and caBIG [57]. The adoption of these standard vocabularies offers a (partial) solution to representational heterogeneity problems especially with those relating to semantics and naming [22].

The Systematised Nomenclature of Medicine – Clinical Terms (SNOMED CT) is a very broad and multilingual clinical healthcare terminology [58]. It defines terminology that can be used to capture information about a patient’s medical history, diseases, treatments, and laboratory results [59]. It is widely used in many countries and provides a central way to represent medical concepts. A drawback is that SNOMED’s terminology and coding results are affected by the browser in which a user searches. SNOMED’s browsers tend to show slightly different results [60]. While SNOMET CT is widely proposed for clinical terminology, only a small number of EHR promoters have put it into practice [32]. SNOMED CT is supposed to be a reliable ontology that is easily comprehended by users. Although, this does not always happen, and it tends to be misunderstood. Also, SNOMED CT struggles with linking of significant medical concepts [61] such as myocardial infarction, diabetes, and hypertension. Systematic errors in SNOMED’s schemas have been detected and major changes are needed. These changes could make SNOMED a more reliable technique in data integration [62].

Another well-known data standard, the International Statistical Classification of Diseases and Related Health Problems (ICD) 10th edition (ICD-10), was developed by the World Health Organisation [63], [64]. The main objective of ICD-10 is standardisation of terminology that is used for clinical diagnoses and procedures. Also, ICD-10 offers multiple uses in health statistics, clinical work, health management, medical billing, epidemiology, and health reporting [64], [65]. This medical coding application is used mostly by physicians [65].

Table 2.1. Key comparison of biomedical terminologies and ontologies.

Name	Scope	Number of elements	Applications	Source
ICD-10 (International version)	Diseases	Around 16,000 codes	Health Statistics,Epidemiology,Health,Reporting,Billing	[64]
GO	Cellular components, molecular functions, biological processes	28,912 terms- (Biological process) 10,900 terms (Molecular Function) 4,016 terms (Cellular component Research)	Research on genes, proteins	[55]
GALEN	Anatomy, surgical deeds, diseases, health care	Over 10,000 terms	Electronic healthcare records, clinical user interfaces, decision sup- port systems, knowledge access systems, natural language processing	[66]
FMA	Anatomy content	120,000 terms	Education, biomedical research	[67]
Read	Patient records, Health Care systems, primary and secondary use of health data, health care, health/socia	298,102 discrete concept codes	Certain vocabulary	[68]

	l care IT systems			
SNOMED CT	- Everything encoded in the electronic health record	326,734 active concepts	Information about a patient's medical history, illnesses, treatments, and laboratory results	[69]
UMLS	Biomedical and health related concepts	Over 1 million concepts	Scientific literature, guidelines, and public health data, natural language processing	[70]

The existing ontologies, vocabularies and data standards offer some advantages and have been used traditionally as a partial solution to problems with data integration. However, there are some universal drawbacks in the usability and the adoption of terminologies and ontologies. One of the main concerns is the plethora of available ontologies and standard terminologies; this has led to different standards being adopted by different data resources. The adoption of different standards, which may not easily align with each other, defeats the purpose, making integration of the different data systems and resources laborious and even impossible. For example, if one clinical study used Read to formalise relevant concepts while a second clinical study used SNOMED CT, data are inevitably captured differently by the two studies. A solution to this would be to impose the same standards on all data resources but would be difficult to implement.

The Unified Medical Language System (UMLS) provides a feasible solution to the problem of the variety of ontologies. The UMLS was developed by the US National Library of Medicine and could be briefly described as an integration tool of more than 150 biomedical vocabularies [71]. UMLS assists the progress of the development of computer systems that function as if they comprehend the biomedical and health related used language. National Library of Medicine's scope is to forward the UMLS databases and programs in order to provide to informaticians an electronic system that develops, recaptures, and integrates not only biomedical/health information but also research information. The UMLS is not designed to satisfy particular scopes, but it is a flexible tool. The UMLS includes integration of vocabularies such as GO, MeSH, Online Mendelian Inheritance in Man and external

sources such as GenBank [72]. The UMLS knowledge sources include: the Metathesaurus, the Semantic Network, and the SPECIALIST Lexicon.

The Metathesaurus is a multi-scope tool that recognises and connects names and same ideas from multiple and different vocabularies. The metathesaurus has been created by the linkage of a lot of thesauri, code sets, controlled terminologies for: patient treatment, health care information, biomedical literature, clinical and health related research, biomedical statistics etc.

The UMLS's semantic network's goal is to support semantic classification in a wide range of terms in numerous eras and is made up of two subcategories. The first one is called 'semantic type' and it supports a homogeneous classification of concepts and terms presented in the UMLS Metathesaurus. The second subcategory is called 'semantic relations' and expresses valuable relationships that can be found among semantic types.

The SPECIALIST Lexicon is an English dictionary that has lots of biomedical and health related terms. It provides numerous words that are necessary for the SPECIALIST Natural Processing System. It is created to help users take over biomedical text differences.

In conclusion, UMLS is designed to achieve post-alignment data harmonisation to the biomedical datasets that are not pre-aligned. However, this approach is not trivial and has yet to attain successful and trusted results [9].

In terms of another traditional approach, there is The Core Outcome Measures in Effectiveness Trials (COMET). It connects people interested in the development and application of agreed standardised sets of outcomes, known as 'core outcome sets' (COS) [73]. These sets represent the minimum that should be measured, captured and reported in all clinical trials of a specific health, and are also suitable for use in clinical audit or research other than randomised trials. There is also no restriction on what it needs to be captured and what the COS will include. People using the COMET initiative expect, that the COS that will be captured, will offer an integrated way to compare, combine and analyse different sources of data due to the common initial format. The COMET's scope is to create and promote applied and methodological with which people could exchange ideas, knowledge and information to improve methodologies of biomedical and health sciences.

This pre-alignment data integration may not always work. If every trial/study would collect these core outcome data, then integration of the associated datasets would be very easy. However, the reality is different: each trial/study collects their own outcome data – as well

as using their own choice of baseline data. but it needs to be mentioned that the COMET initiative becomes even more popular in our days. In this thesis, we therefore focus on a different approach that does not require pre-alignment and will be useful in scenarios where pre-alignment is not possible or has not happened.

2.3.4 Probabilistic data integration methods in health data

Heterogeneous biomedical data integration remains a challenge especially due to noisy biomedical resources, coverage diversity, and functional biases [74]. There is a growing effort to integrate these data using probabilistic knowledge representation.

Huttenhower and Troyanskaya [74] used Bayesian networks to predict useful protein relationships under multiple and different conditions. The study took into consideration the structure of the network and compared different probabilistic methods. It compared the golden value of the conditional probabilities with those from a method called 'expectation maximisation' and another one called 'extended logistic regression'. The behaviour of Bayesian data integration varied according to the parameters chosen to accomplish computational and biological scope. They concluded that Bayesian networks are indeed a steady and functional way of integrating data, but their performance should be carefully considered: the performance is different for each individual functional category and could produce a different accuracy to the general model. The proposed technique is not only for Bayesian networks and is a general method for biological data integration.

Gevaert's PhD thesis [75] focused on Bayesian networks as a biomedical decision support model and in particular how they could be used to formulate a model that integrates heterogeneous and high dimensional data. His research scope was to formulate a structure for biomedical data integration and for this he took the following steps: a. demonstration of Bayesian networks for primary data i.e. clinical and genomic; b. development of algorithms for primary data integration; and c. extension of the framework to allow the incorporation of secondary data sources. This study focused on primary data integration and a Bayesian network integration technique was applied to integrate clinical and microarray data from breast cancer patients. Bayesian networks were applied in full, partial, and decision integration and the difference among those was in which step the integration took place. Partial Bayesian integration showed the best results and the final model had 3 clinical variables and 13 genes to predict prognosis of breast cancer patients.

The thesis also examined the usage of Bayesian networks in secondary data integration, broadening the Bayesian network framework to include secondary data as priors. Secondary

data sources present high dimensionality in general and with this study Gevaert [75] used his methods including 4 datasets. In this case, secondary Bayesian data integration showed that using a text prior in every case, improved the prognosis prediction of breast cancer. Next, Bayesian networks were applied successfully for the prediction of malignancy in ovarian masses in clinical data. Lastly, a specific Bayesian class, hidden Markov modelling, was applied to understand the molecular mechanisms that lead to carcinogenesis in some particular tumours. In conclusion, Bayesian networks were used for the development of a framework that integrates heterogeneous and high dimensional data to support a decision-making system. This provides an example for the successful use of probabilistic methods in biomedicine to model a single primary data source, to integrate primary sources and finally to integrate secondary data sources.

Other examples include the use of Bayesian networks by Savage et al. [76] to infer transcriptional modules by integrating gene expression and transcription factor binding data. They concluded that their model, could extract clusters with better functional consistency than other existing methods.

2.3.5 Content heterogeneity and missing data

Missing data refer to values that are not stored or recorded for a variable in the observation of interest. It is quite common in different studies and in research in general, even if they are well controlled, the existence of missing data. It is a critical consideration as it can lead to biased results, decrease of statistical power (potential invalid results and conclusions, decrease of sampling representation, difficult study analysis [77]. All these issues that missing data bring might affect trials and conclusions' accuracy. Until recently, researchers have focused on discussing their conclusions on the assumption of a complete dataset.

Types of Missing Data

There are three different categories of missingness that applies to data: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [78].

MCAR

It means that the probability of data missingness of a variable Y is not connected with either the specific record that was supposed to be obtained or any other variable X in the data set. However, it does allow for the possibility that missingness on that variable Y is related to the missingness on another variable X [79], [80]. Data missing due to technical reasons or lost in transit are characterised as MCAR. MCAR in most cases is an assumption which may

not reflect reality. Having MCAR data, when applying complete case analysis, gives unbiased results with less power but would still lead to valid conclusions.

MAR

This category of missingness is more realistic than MCAR. It means that the probability of data missingness is connected with the set of observed records but is not connected with the specific value that is supposed to have been recorded. In other words, the probability of missing data on Y is unrelated to the value of Y after controlling for other variables in the analysis (X). In a formal way, P(Ymissing|Y,X) = P(Ymissing|X) [79]. It means that any systematic differences between the missing and observed records, can be explained by other observed variables. So, if we can control this dependent variable, we can control a random subset. Therefore, the solution to that is the usage of techniques to absorb variables connected to missingness.

MNAR

This is the most problematic category of missing data. MNAR means that data are neither MCAR nor MAR. It occurs when records are missing because the values are related to the reason of missingness. MNAR occurs when the missingness is dependent on the missing variable itself, or in fact dependent on any other variable that is unobserved (i.e. unmeasured records). In MNAR, the probability of data missingness depends on unobserved records. The value of the unobserved responses depends on information not available for the analysis (i.e. not the values observed previously on the analysis variable or the covariates being used), and thus, future observations cannot be predicted without bias by the model.

Nonignorability (MNAR) means that we need to model the missing data mechanism to get accurate estimates of the variables of interest, and this requires advanced methods. On the contrary, ignorability means that we do not need to model the missing data mechanism as part of the estimation process. MCAR is ignorable and MAR can be made ignorable under the appropriate analysis. If we include the mechanism variables, then we can ignore the problems with MAR data, but we apply specific techniques to use the data in an efficient way.

2.3.6 Statistical methods for handling missing data

Most analytical methods cannot handle missing values. For instance, if we feed a dataset with missing values to a logistic regression analysis algorithm, it will result in an error. This is independent of the missingness mechanism: any missing value will cause such an error.

Different methods to handle missing values have been proposed. One of the most commonly used methods is complete case analysis where observations with missing data are omitted from the analysis [81]. This is a simple method that is easy to carry out and may give unbiased estimates when data is MCAR. However, if data are not MCAR, complete case analysis can give biased estimates. Furthermore, the reduced number of participants results in a loss of statistical power [82]. If the number of missing values is low, researchers often still choose this solution because then anticipated bias and effects on statistical power will be low. The other deletion method is called pairwise. It eliminates information when specific data are needed to test a specific hypothesis. Pairwise deletion keeps and uses all the information needed, therefore it has a larger number of data than complete case analysis. One common problem with this method is that the model's parameters may differ due to different input datasets, and therefore different statistics will exist.

An alternative approach is to use methods that fill in or impute the missing data. Instead of deleting the records that have missing values, imputation replaces the missing data with estimated/substituted values. All the missing data are replaced with a probable value that was calculated by other available information [77]. Mean imputation is one such method in which the mean of the observed values for each variable is computed and the missing values for that variable are imputed by this mean and the resulting completed data set is used for inference. Single mean imputation does not take into account the uncertainty in the imputations. It is a simple and unconditional method but it introduces biased estimations when data are not MCAR. However, it gives more accurate results than complete case analysis method. Also, there are other single imputation methods like median, mode, regression (the imputed value is predicted from a regression equation), hot-deck (replaces the missing data by realistic scores that preserve the variable distribution), last observation caried forward (imputes the missing value with the last observation of the individual) etc [83]. Another single imputation technique is single conditional imputation. It is a potentially more accurate method of single imputation, where the missing value is replaced with a value that is conditional on another. In case that missing data are not MCAR, the results are less biased than the previous methods mentioned. The issue with this technique is that the standard errors are underestimated when dataset is considered as a complete dataset (no missing values) [84]. One of the problems with all single imputation methods is that the filled-in observations are treated as actual observations in subsequent analysis. However, the filled-in values are estimates, which have standard errors.

Multiple imputation (MI) can be used to take the uncertainty about the estimates into account That's why MI is recommended as an appropriate way of handling missingness in data [85]–

[87]. It is more advantageous than the single imputation because it uses several complete data sets and provides both the within-imputation and between-imputation variability. There are three steps of multiple imputation process. MI is the process of replacing each missing data point with a set of m > 1 plausible values to generate m complete data sets. These complete data sets are then analysed by standard statistical tools, and the results combined, to give parameter estimates and standard errors that take into account the uncertainty due to the missing data values [88].

According to van Buuren and Groothuis-Oudshoorn [89] there are two general approaches to multiple imputation: joint modelling presented by Schafer and Olsen [90] and fully conditional specification developed by van Buuren [91]. Joint modelling results to a specified multivariate distribution for the missing data and imputation from their conditional distributions by Markov Chain Monte Carlo techniques. Fully conditional specification (FCS) is based on the loop that shapes a conditional distribution for each incomplete variable. It does not explicitly assume a particular multivariate distribution, but assumes that one exists and can be generated by it, using Gibbs sampling [92].

Finally, a popular choice for imputing binary variables is logistic regression which is also commonly used in multiple imputation, much like a propensity score. It is a parametric method that assumes an underlying logistic model for the imputed variable (given other predictors) [91], [93], [94]. In FCS, the default imputation techniques are predictive mean matching for a numeric variable, logistic regression for a binary variable, multinomial logit model for a categorical variable with more than 2 levels, and ordered logit model for an ordered categorical variable with more than 2 levels [95].

It needs to be mentioned that there is no method to handle missing data that is perfect and that does not come with several limitations. The best approach is different for each case and depends on a number of critical choices, the degree of data missingness and the relationships/correlations between covariates and outcome data [96].

2.3.7 The Problem of Systematically Missing Values

Missingness can be due to sporadically missing data (i.e., values are absent for a proportion of participants within a study) and systematically missing data (i.e., one or more variables were collected in one study but not the other) [97]. In scenarios of data integration and individual patient data meta-analyses, datasets are analysed post hoc. When problems like sporadically missing data are encountered complete case/records analysis. This is a when only participants for which we have no missing data on the variables of interest are included in subsequent statistical analysis. Participants with any missing data are excluded. An

alternative solution is imputation. In those cases of sporadically missing data, much research has been done and it is suggested as a gold standard approach to apply FCS with the MICE algorithm [95], [98]. It is widely considered the best approach to resolve sporadic missingness.

The issue of systematically missing data is one that has, to date, not been fully addressed. Schafer and Yucel [99], [100], a Gibbs sampler to generate multiple imputations of continuous missing variables from a joint multivariate linear mixed model [99], [100] implemented in the PAN package [101]. Most recently, Quartagno and Carpenter [102] suggested that MI is flexible to allow for between-study heterogeneity when studies have missing covariates [97]. Their research led to the extension of the PAN package by the REALCOM software [103] and specifically the R package jomo [104] that allows missing data in any level and handles latent categorical data through latent normal variables. In their published paper they proposed and evaluated a joint modelling approach to multiple imputation of individual patient data in meta-analysis, with an across-study probability distribution for the study specific covariance matrices. In addition, Van Buuren [105], [106] suggested a MICE extension to allowance of multilevel imputation by a Bayesian approach with recent extension for 2-level variables' imputation [98]. The aforementioned approaches are able to impute multilevel data either systematically or sporadically. Resche-Rigon and White [107] proposed a method that handles, at the same time, systematically and sporadically missing data. Based on their simulation study the proposed methods can be successfully combined in a multilevel MICE procedure, when cluster means are not included in the imputation models. Unfortunately, the multiple imputation methods and packages presented above are currently not able to handle systematically missing data except for jomo [107]. While no gold standard approach exists, and in many cases, variables that are systematically missing from one dataset would just be omitted from the analysis – complete case analysis.

2.4 Summary

In summary, we highlighted the need for biomedical data integration of structured data. Bringing literature together, studies provided important insights into the importance of the issue of representational heterogeneity resulting from data integration. We then discussed traditional and probabilistic approaches to tackle heterogeneous biomedical data. The studies presented thus far provided evidence on unresolved issues concerning content

heterogeneity, types of missing data and probabilistic methods to overcome data missingness.

In conclusion, the scope of this literature review was to argue that probabilistic approaches to resolving content heterogeneity are under-researched. Our proposal is characterised by a probabilistic rather than a functional view on content harmonisation; by post-alignment rather than pre-alignment of data sources; and by a pragmatic, top-down approach to answering research questions rather than a laborious, bottom-up approach to data harmonisation.

Limitations

Searching was broad in terms of used phrases and synonyms and limited to specific databases such as PubMed, Scopus, and Google Scholar. The majority of papers exist in those, and the basic reason for not choosing others is that we tried to avoid specific publishing companies (avoid bias). The review also was limited to papers written in the English language. Potential gaps in the literature may exist because the concepts were very broad and there was a systematic effort to be included most of the information.

Chapter 3: Systematically missing values

3.1 Introduction

In Chapter 2, we reviewed the scientific literature on concepts such as representational heterogeneity, data integration, and methods for integration of heterogeneous datasets, with a focus on health data. We further examined content heterogeneity and explored which probabilistic methods have been used for integration of heterogeneous datasets. We also discussed in detail the different types of missing data, and methods that resolve data missingness. We concluded with a description of the chronic autoimmune disease SLE and the MASTERPLANS project – data from which will be used throughout this thesis.

In this first chapter of the results, we describe a probabilistic approach to resolving content heterogeneity that is caused by systematically missing values because a variable exists in one dataset but not in the other. First, in Section 3.2, we identify the problem at hand and provide an illustrative example. We next (in Section 3.3) describe a probabilistic approach to data harmonisation in generic terms followed by an evaluation of the approach through a series of simulation studies (in Section 3.4). Subsequently, in Section 3.5, we apply our method to real-world data provided by MASTERPLANS project, and present the results obtained. We conclude with a discussion in Section 3.6 on the utility of the approach and directions for future research. In Appendix A we can find detailed simulations' results presented in tables.

3.2 Problem identification of systematically missing values

The issue of systematically missing data has still not been fully researched. As mentioned extensively in Chapter 2 (sections 2.3.5-2.3.7), there has been some research that exploring situations of systematic missing variables. Here, we briefly mention the main points and terminology around missing data.

Missingness can be due to sporadically missing data (i.e., values are absent for a proportion of participants within a study) and systematically missing data (i.e., one or more variables were collected in one study but not the other) [97]. In scenarios of data integration and IPD meta-analyses, datasets are analysed post hoc. When problems like sporadically missing data are encountered one (traditional) approach is complete case analysis. Participants with any missing data are excluded. However, in cases of sporadically missing data, much research has been done and it is suggested as a gold standard approach to apply FCS with the MICE algorithm [95], [98]. MI is widely considered the best approach to resolve sporadic missingness.

As mentioned in more detail in chapter 2, in this study, we question the assumption that perfect, data-level standardisation is needed for inference with data that are from different sources. We believe that more progress can be achieved by adopting a top-down approach to harmonisation that starts with research questions rather than data collections. We suggest a probabilistic approach to data harmonisation in which content equivalence is a matter of uncertainty rather than a dichotomy.

3.3 Theoretical solution

First, we assumed that we had a number (more than one) of study datasets. These study datasets were assumed to be non-overlapping in terms of patients included. In order to focus on the problem of content heterogeneity due to systematically missing values, we also assumed that the datasets did not contain sporadically missing values, that each study dataset was described by a flat table where each row corresponds to one study participant, and that naming differences between study datasets had already been resolved.

From a probabilistic perspective, imperfect alignment of different data sources is not problematic as long as we can derive which information each of the sources provides in answering our research question and properly quantify the uncertainty that is caused by the imperfect alignment. In our case we would like to use all the available information across the datasets being integrated. The general idea behind our integration method is that the problem of content heterogeneity, presented as a missing variable problem, can be translated into a missing value problem and then solved using established methods for addressing missing values (imputation).

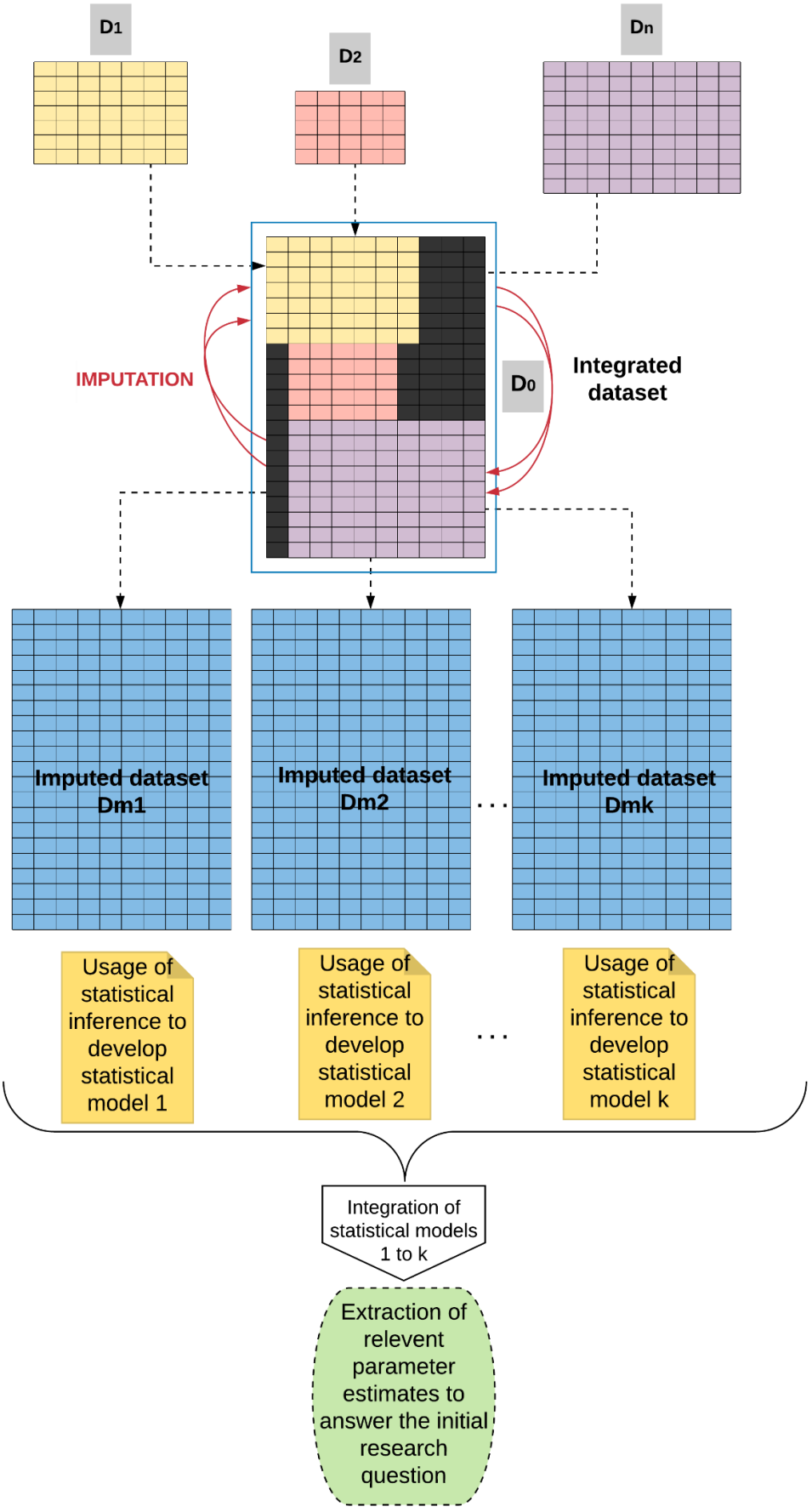


Figure 3.1. Main tasks of our probabilistic data integration process to solve systematically missing values problem. The black squares denote missing data.

As shown in figure 3.1, we propose an approach that comprises a number of tasks to handle systematically missing values problem during the analysis, assuming that any naming differences have been resolved beforehand. We aim to create an integrated dataset that includes all the information from the constituent datasets, but the reality is that that integrated dataset would have many 'gaps' due to content heterogeneity. The next step is to 'fill' these gaps using multiple imputation, producing k complete datasets, after which statistical inference can be applied to each of the k imputed datasets, producing k statistical models. After integration of these models, yielding the final model, we can extract the parameter estimates that we were looking for to answer the specific research question.

Here, we present the aforementioned goal for handling the content heterogeneity problem in more detailed tasks:

1. We select and stack datasets D_1 to D_n in one large integrated dataset (D_o). Selection is based on datasets potentially relevant for answering a research question of interest, based on the available metadata. More specifically, we select all participants for which there is a nonzero probability that they are relevant to answering the research question at hand. These would be all patients across selected datasets that have a nonzero probability of meeting the inclusion criteria for the research question. For some patients that probability will be 1 because we have complete information. For others it might be smaller than 1 because we do not have complete information on them with respect to the inclusion criteria.
2. In order to answer the research question, we want to fit a statistical model in which one or more variable(s) present a content heterogeneity problem. We approach the systematically missing values problem as a missing values problem; we solve it by applying *imputation* – a method well established to solve data missingness – to the integrated dataset D_0.
3. We create multiple copies (imputed datasets/imputations) of the dataset D_0, and the missing values replaced by imputed values. This means imputation process uses information from other variables and has a random component. Each variable with missing data is modelled conditional upon the other variables in the data [108]. The imputed values are sampled from the known observed data's predictive distribution. Multiple imputation is therefore a Bayesian approach. This imputation procedure must fully consider all uncertainty when predicting the missing values. It does it by inserting suitable variability into the multiple imputed values. This variability is needed as we do not the truth because of data missingness [109]. Once the data have

been imputed in D_0, each imputed dataset (Dm_1 to Dm_k) is 'complete' in the sense that it has no missing values.

4. We perform the analysis (modelling) designed to answer our research question (e.g., a regression analysis) on each of the imputed and complete dataset Dm_1 to Dm_k. In other words, we use statistical inference to develop 1 statistical model in each imputed dataset (Dm_1 to Dm_k).
5. In each of the imputed datasets, estimates will differ due to variation introduced in missing values' imputation. So, we combine/integrate the estimates of the k statistical models to obtain the final result — average of the overall estimates. Standard errors are calculated using Rubin's rules [94], [110]. The variance estimates, calculated in this step, involve both the 'within' variance calculated for each dataset individually, as well as the 'between' variance that reflects the uncertainty in the imputations (equation 3.1) — how variable the results are across the imputed datasets Dm_1 to Dm_k.

$$Var_{total}(\theta) = \sum Var_{within}(\theta) + (1 + \frac{1}{k})Var_{between}(\theta) \quad (3.1)$$

(where θ, a vector of unknown parameters)

While it is possible to write a short computer program to do the combining, many standard statistical software packages include procedures to combine results across datasets automatically. Thus, from the user's perspective doing these two steps and obtaining the final estimates are often no more complicated than running a single regression in a single dataset.

6. We extract relevant parameter estimates that answer the initial research question.

MICE's steps can be found in more detail in other published research [91], [98], [108], [111]. Missing values are replaced by imputed values based on the chosen method. All available FCS's imputation techniques can be found in detail in table 1, page 16 in MICE published paper [98]. Each variable with missing data is modelled conditional upon other variables in the dataset, which we refer to later as *imputation model* [108]. Imputations are generated according to the default method, which is, for numerical data, predictive mean matching (pmm) and for categorical data with >2 levels, multinomial logit model factor (polyreg).

Our goal ideally focuses on the construction of a single, large table that captures all the information that is described by the underlying datasets (tables) that we have integrated. Furthermore, that integrated table should not have 'gaps' or (systematically) missing

values, and the interpretation of all the values in the table should be consistent and ambiguous.

3.4 Simulation studies

In this section, we present the results of a series of simulation studies designed to evaluate our approach to handle the content heterogeneity problem described in the preceding sections. We first define synthetic data that contain enough variables to illustrate this particular content heterogeneity problem. Without loss of generality, we specify the bare minimum required to capture this problem. We take this approach for evaluation as in synthetic data the true associations between the predictors and the outcome are known and can be used reliable to quantify our integration method's performance.

3.4.1 Simulation study design

We performed a series of simulation studies designed to investigate our probabilistic methods in a simplified and generalised setting. Common sample sizes for simulation studies are 500 and 1000 iterations. As shown in figure 3.2 and table 3.1, the process of each simulation described below is repeated 1000 times [112] as an agreement between estimate accuracy and computational time, to obtain different datasets under the specified parameters, that are then used for analysis.

Table 3.1. Description of data used for the simulations to unerstand the distributions of the variables X_1, X_2, X_3 in datasets D_n.

	D_n
X_1, continuous	~ N(0,1)
X_2, continuous	~ N(0,2)
X_3, categorical (%)	
'A'	p(X_3=A) =0.10
'B'	p(X_3=B) =0.20
'C'	p(X_3=C) =0.65
'D'	p(X_3=D) =0.05

Note: X_1, X_2 are correlated by 20% in each study D_n

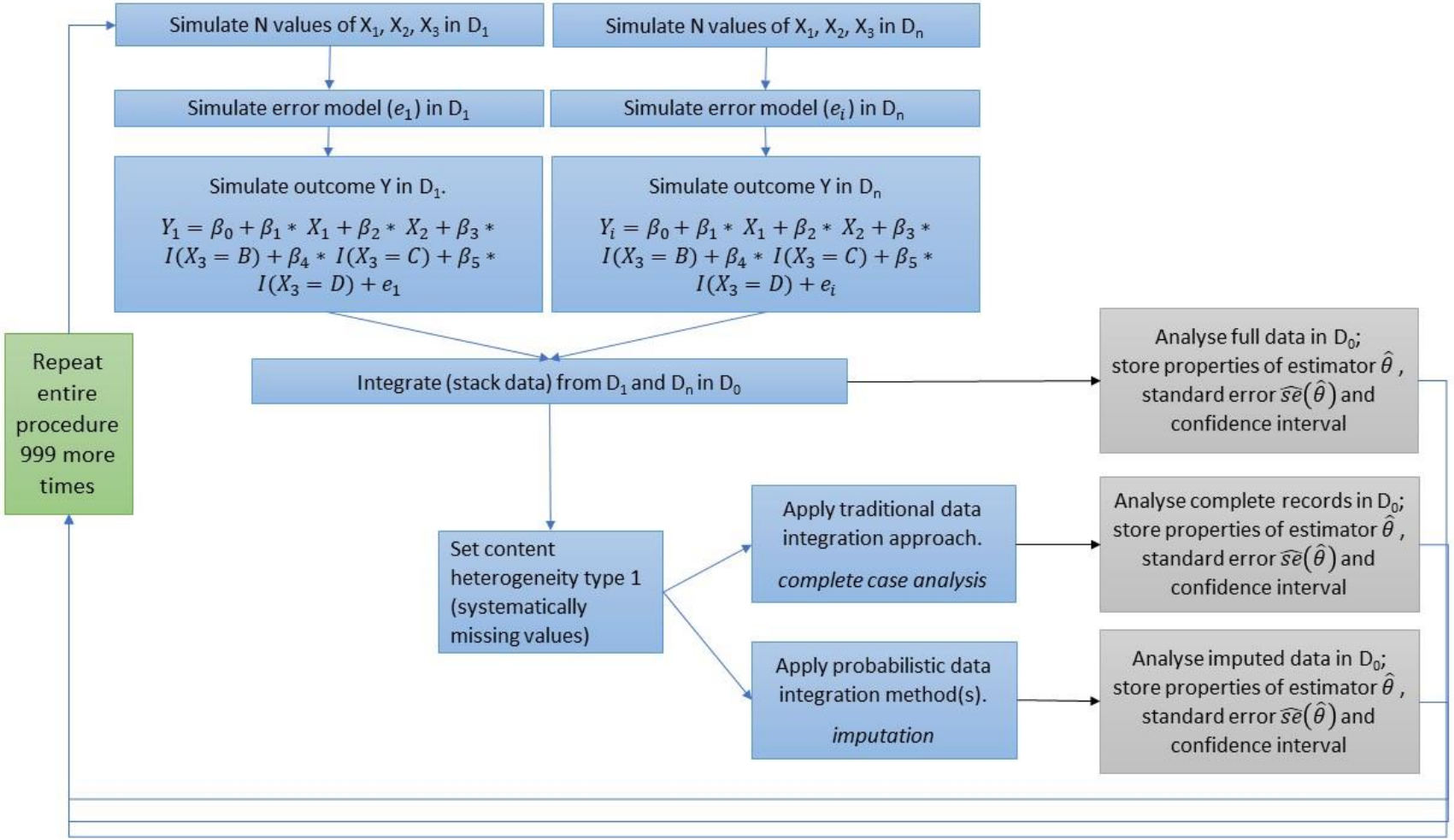


Figure 3.2. A pictorial representation of the simulation procedure for systematically missing values.

The simulation procedure (Figure 3.2) goes as follows:

1. We assume a synthetic dataset D_1 size of N individuals each for which we originally have complete information on a continuous covariate X_1, a continuous covariate X_2, a categorical variable X_3 (see table 3.1). The variables X_1, and X_2, are continuous and follow a normal distribution. X_3 is a categorical variable with levels 'A', 'B', 'C', and 'D'. The variables X_1, X_2 are correlated by r =0.2, X_3 is assumed to be independent of X_1 and X_2. They follow the following distributions: X_1, X_2 ~ MVN((0,0),(1,2),r), X_3: p(X_3=A) =0.10, p(X_3=B) =0.20, p(X_3=C) =0.65, p(X_3=D) =0.05.

2. Simulate model error (e_i) with a normal distribution.
3. Simulate a continuous outcome, Y, under a given data-generating model, in D_1. Y is complete, has no missing data and has known dependency on X variables. The data generating model from model in equation 3.2 with β_0 = 2.5079, β_1 = 0.6874, β_2 = -2.0591, β_3 = -0.7876, β_4 = 0.6238, β_5 = -0.9491.

$$Y_i = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * I(X_3 = B) + \beta_4 * I(X_3 = C) + \\ + \beta_5 * I(X_3 = D) + e_i \quad (3.2)$$

4. Repeat steps 1-3 for synthetic dataset(s) D_n.
5. Integrate/stack X_1, X_2, X_3, Y from D_1 and D_n into one integrated dataset D_0.
6. Fit a linear regression model: Y ~ X_1 + X_2 +X_3 to the true Full Data in D_0, analyse them and store properties of estimator $\hat{\theta}$, standard error $\widehat{se}(\hat{\theta})$ and confidence interval.
7. Apply content heterogeneity type 1. For example, for individuals in D_0, set X_1 (i.e., individuals from D_1) values to missing. So, X_1 is a variable with systematically missing values for individuals from D_1.
8. Solve content heterogeneity problem 1 by applying the traditional approach – complete case analysis (Complete Records). Fit a linear regression model: Y ~ X'_1 + X_2 + X_3 in D_0, analyse Complete Records and store properties of estimator $\hat{\theta}$, standard error $\widehat{se}(\hat{\theta})$ and confidence interval.
9. Solve content heterogeneity problem 1 by applying the probabilistic approach FCS as presented in Figure 1. Fit a linear regression model: Y ~ X''_1 + X_2 + X_3 in D_0, where X''_1 is imputed and complete. Then, analyse data and store properties of estimator $\hat{\theta}$, standard error $\widehat{se}(\hat{\theta})$ and confidence interval.

All imputation analyses carried out using the R package *mice* freely available on CRAN [98]. We used the seed function and set seed to 156524. The starting seed number is used to generate a sequence of random number and it ensures that we get the same result if we start with that same seed each time, we run the same process. The chosen values of betas used in simulations are selected empirically, and at random. We chose specific values as we want to focus on checking how the probabilistic methodologies work on solving content heterogeneity without the effect of a range of betas. We expect to see similar results with other betas. We decided randomly which variable and which dataset contained the content heterogeneity problem in each scenario, so we achieve generality. In the next section, we present the simulations' results.

3.4.2 Simulations' scenarios

Tables 3.2 and 3.3 show the different scenarios to generate data following Figure 3.2 and its aforementioned tasks. We explore different simulation scenarios with different number of individuals (N) per study, number of studies (D_n), and model errors (e_i). The number of m is set to five (default). The number of iterations is set to five (default).

Table 3.2. Scenarios 1 - 5 used to generate data from Figure 3.2.

	Scenario 1	**Scenario 2**	**Scenario 3**	**Scenario 4**	**Scenario 5**
Number of individuals per study (N)	200	1000	200	1000	D_1: 200, D_2:150, D_3:50, D_4:75, D_5:100
Model error e_i: (same for each study)	N~(0,0.2) N~(0,2) N~(0,20)	N~(0,0.2) N~(0,2) N~(0,20)	N~(0,0.2) N~(0,2) N~(0,20)	N~(0,0.2) N~(0,2) N~(0,20)	N~(0,0.2) N~(0,2) N~(0,20)
Number of studies (D)	2	2	5	5	5
Imputations (m)	5	5	5	5	5
Missingness applied to:	X_1 from D_1	X_1 from D_1	X_1 from D_4,D_5	X_1 from D_2, D_5	X_1 from D_4, D_5

Table 3.3. Scenarios 6 - 10 used to generate data from Figure 3.2.

	Scenario 6	**Scenario 7**	**Scenario 8**	**Scenario 9**	**Scenario 10**
Number of individuals per study (N)	D_1:800, D_2:150, D_3:50, D_4:75, D_5:350, D_6:200, D_7:150, D_8:500, D_9:750, D_{10}:100	100	200	500	1000
Model error e_i: (same for each study)	N~(0,0.2) N~(0,2) N~(0,20)	e_1: N~(0,1.2), e_2: N~(0,1.3)	e_1: N~(0,1.2), e_2: N~(0,1.3)	e_1: N~(0,1.2), e_2: N~(0,1.3)	e_1: N~(0,1.2), e_2: N~(0,1.3)
Number of studies (D)	10	2	2	2	2
Imputations (m)	5	5	5	5	5
Missingness applied to:	X_1 from D_3, D_6, D_9, D_{10}	X_1 from D_1	X_1 from D_1	X_1 from D_1	X_1 from D_1

3.4.3 Performance measures

The performance measures describe a numerical quantity used to assess the performance of a method [112]. To examine the simulations' results we choose as performance measures the following:

Properties of estimator of θ

- *Bias*

  Bias is frequently of central interest and quantifies whether a method targets estimand θ on average. We check whether our estimate coefficients are biased or not. In the simulation setting, bias is how far from the average estimate $\hat{\theta}$ exceeds estimand θ (equation 3.3). There is no bias when the average is similar or close to

the real value. If it is not, then the method produces biased results. The absence of bias is one property of an estimator; while it is often of central interest, we may sometimes accept small biases because of other good properties.

$$\text{Bias} = E\left[\hat{\theta}\right] - \theta \tag{3.3}$$

In this setting that we had 1,000 simulations, mean bias was estimated as shown in equation 3.4.

$$\text{Bias}_{estimate} = \frac{1}{1000}\sum_{i=1}^{1000}\left|\hat{\theta}_i - \theta\right| \tag{3.4}$$

- *Empirical standard error (EmpSE)/Precision*

The empirical SE is a measure of the precision or efficiency of the estimator of θ. We want to obtain precise estimates such as low variance. The EmpSE should be as small as possible for good results. The EmpSE estimates the long-run standard deviation of $\hat{\theta}_i$ over the number of simulations repetitions. EmpSE is defined in equation 3.5 and in this setting was calculated as in equation 3.6.

$$\text{EmpSE} = \sqrt{Var(\hat{\theta})} \tag{3.5}$$

$$\text{EmpSE}_{estimate} = \sqrt{\frac{1}{1000-1}\sum_{i=1}^{1000}(\hat{\theta}_i - \bar{\theta})^2} \tag{3.6}$$

Properties of Standard Error (SE) ($\widehat{se}(\hat{\theta})$)

- *Mean/Average model Standard Error (mSE)*

The average model SE should be unbiased for the variance of the estimator [112]. We want the mSE to be almost equal to the true SE, EmpSE. The mSE is defined in the equation 3.7 and in this setting was calculated as in equation 3.8.

$$\text{mSE} = \sqrt{E[Var(\hat{\theta})]} \tag{3.7}$$

$$\text{mSE}_{estimate} = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}\widehat{Var}(\hat{\theta}_i)} \tag{3.8}$$

Properties of Confidence Interval ($\hat{\theta}_{low}, \hat{\theta}_{upp}$)

- *Coverage*

Coverage of confidence intervals is a key property for the long-run frequentist behaviour of an estimator. It is defined as the probability that a confidence interval contains θ [112]. If we estimate a regression model for each coefficient, we get a point estimate and we get a standard deviation for variance. So, that is a measure of uncertainty. We check whether the 95% confidence interval of our regression coefficients is accurate, and we compare it with true model and complete case analysis' coverage level. Coverage is defined in the equation 3.9 and in this setting was calculated as in equation 3.10.

$$\text{Coverage} = \Pr\left(\hat{\theta}_{low} \leq \theta \leq \hat{\theta}_{upp}\right) \tag{3.9}$$

$$\text{Coverage}_{estimate} = \frac{1}{1000}\sum_{i=1}^{1000} 1\left(\hat{\theta}_{low,i} \leq \theta \leq \hat{\theta}_{upp,i}\right) \tag{3.10}$$

For each iteration within a given simulation scenario, all performance measures were calculated between the estimates coefficients of each (probabilistic/traditional) integration model and the generating coefficients. Then, all performance measures were averaged across iterations for all the coefficients. Across all the simulation scenarios in all chapters, we present the results of the analyses with different methods in terms of mean estimates, mean standard error (mSE) estimated from the models, standard deviation of the simulation estimates (EmpSE) and coverage level (Cov). A valid method should yield unbiased results, similar model, and empirical standard errors and coverage levels close to 95%. In this chapter, estimand θ was the estimate $\hat{\theta}_i$ of X_1 coefficient in each model fit.

3.4.4 Results

In this section, we present the results of a series of simulation studies. It is often good to include a familiar 'benchmark' method to check the simulation result. In our case – the systematically missing values problem – we also include complete case analysis as a check. The aim is to compare the complete case analysis with imputation in the concept of systematically missing values and evaluate our probabilistic data integration approach.

We are aware that the amount of information of simulations could be huge. Therefore, in order to provide the reader with a gentler introduction to the main findings of the simulation studies, we decided to include graphical representations of the main results in figures for all

scenarios separately and together per category, so we make necessary comparisons. All complete simulation results can be found in Appendix A.

Simulations with studies of same sizes (Scenarios 1 - 4)

Scenario 1

Here, we simulated data from two studies, **each with 200 patients**. For each simulated dataset, we completely removed X_1 from D_1. We present the simulation results in figure 3.3 with e_i: (N ~ (0, 0.2)), (N ~ (0, 2)), (N ~ (0, 20)) respectively. Analysis of datasets imputed gave very good results both in terms of bias (figure 3.7), precision (figures 3.9 and 3.10) and confidence interval coverage (figure 3.8). We see clear gain in precision compared with the Complete Records analysis. When imputing with very large model error e_i: N ~ (0, 20), mSE and EmpSE's difference was slightly larger in FCS than in Full Data (figures 3.9 and 13.0 respectively). However, they were still equal in FCS.

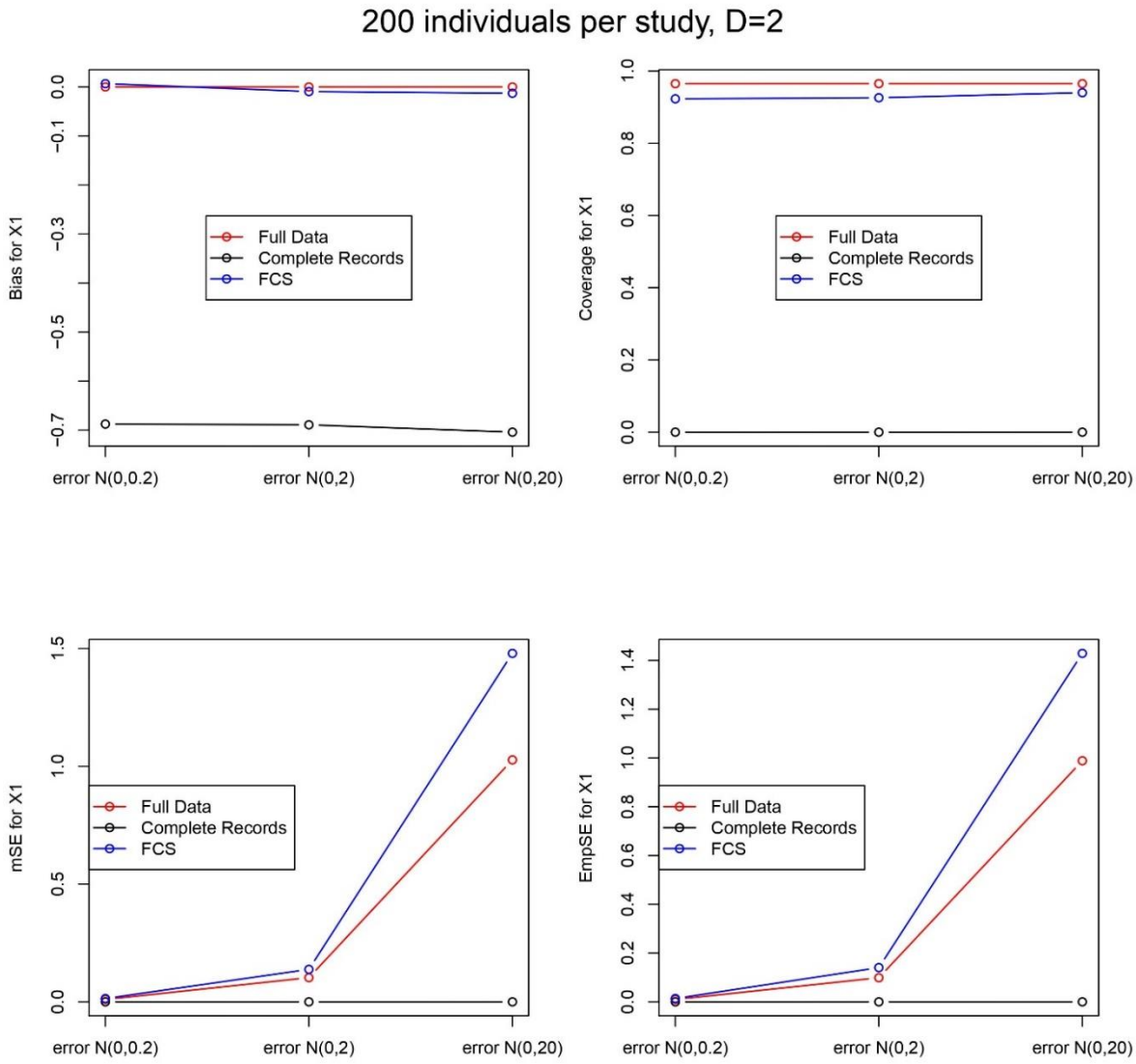


Figure 3.3. Main results from scenario 1's simulation study: Comparison of Bias, Coverage level, mSE and EmpSE for X_1 after 1000 simulations with Full Data (red), handling systematically missing values with Complete Records (black) and FCS (blue) for three model errors.

FCS: Multiple imputation by fully conditional specification; *D:* number of studies, *mSE:* mean model standard error; *EmpSE:* mean empirical standard error.

Scenario 2

In this scenario, we simulated data from two studies, **each with 1000 patients.** For each simulated dataset, we completely removed X_1 from D_1. We present the simulation results in figure 3.4 with e_i: (N ~ (0, 0.2)), (N ~ (0, 2)), (N ~ (0, 20)) respectively. FCS model was still compatible with Full Data and the findings were very good. Bias, mSE, EmpSE were slighlty lower here than in scenario 1 but still very good. Bias was higher and coverage was slighlty lower when e_i: N ~ (0, 20) in this scenario than in scenario 1.

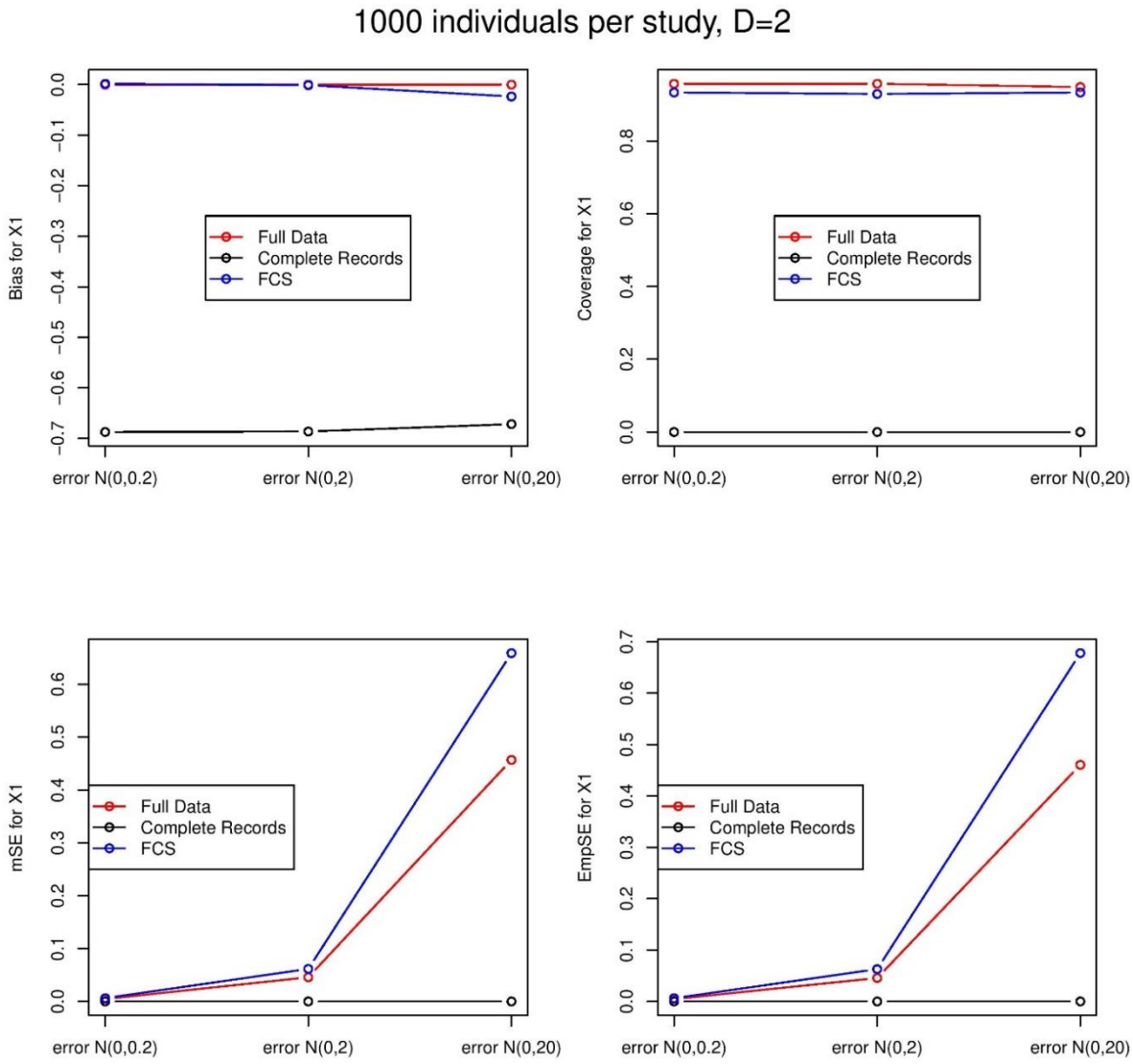


Figure 3.4. Main results from scenario 2's simulation study: Comparison of Bias, Coverage level, mSE and EmpSE for X_1 after 1000 simulations with Full Data (red), handling systematically missing values with Complete Records (black) and FCS (blue) for three model errors.

FCS: Multiple imputation by fully conditional specification; *D:* number of studies, *mSE:* mean model standard error; *EmpSE:* mean empirical standard error.

Scenario 3

In this scenario, we simulated data from **five studies**, each with **200 patients.** For each simulated dataset, we completely removed X_1 from two studies (D_4 and D_5) in D_0. We present the simulation results in figure 3.5. FCS provided very good results. This scenario had better coverage in the three errors than scenario 1 and lower bias, mSE and EmpSE than in scenario 1.

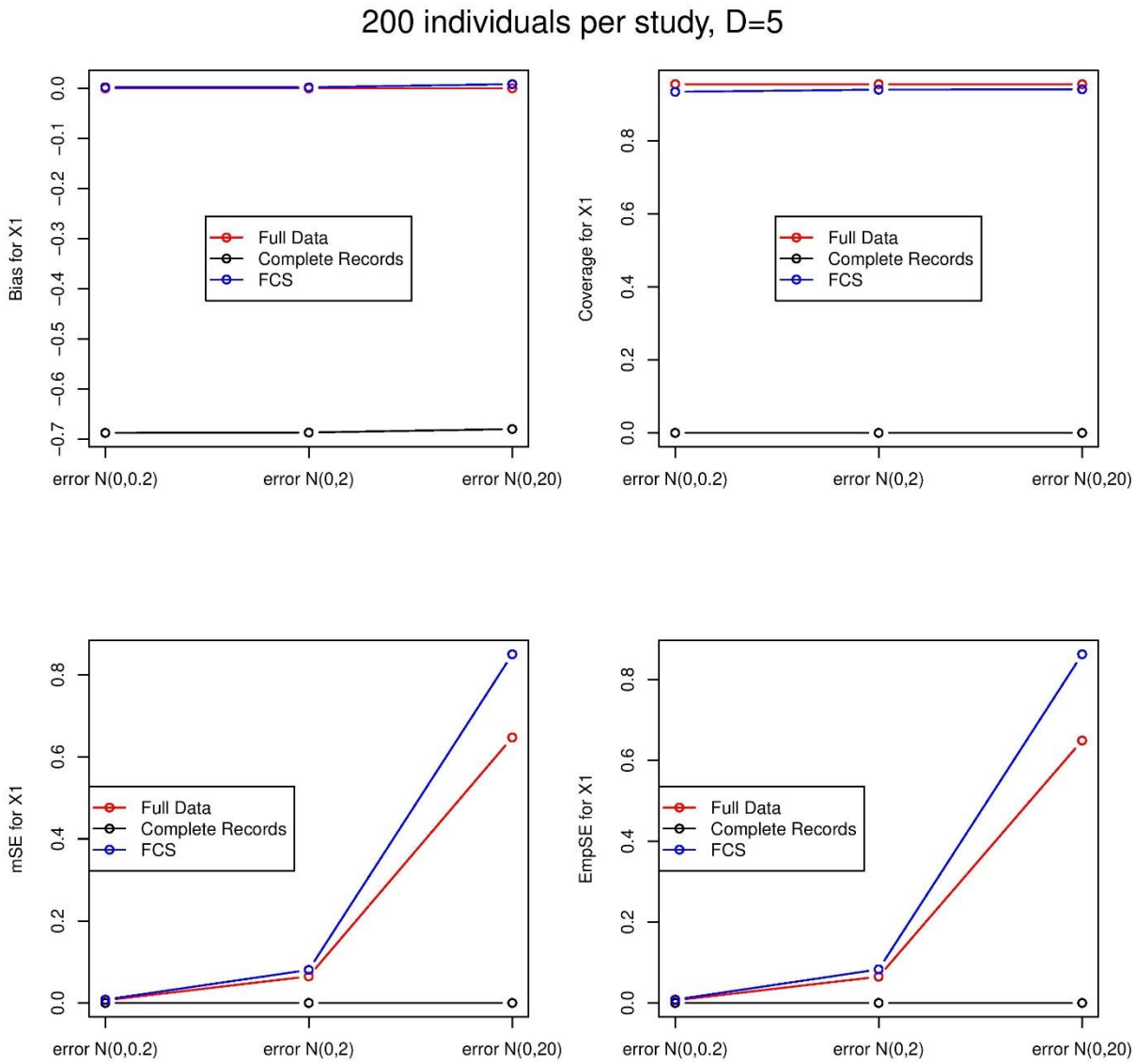


Figure 3.5. Main results from scenario 3's simulation study: Comparison of Bias, Coverage level, mSE and EmpSE for X_1 after 1000 simulations with Full Data (red), handling systematically missing values with Complete Records (black) and FCS (blue) for three model errors.

FCS: Multiple imputation by fully conditional specification; *D:* number of studies, *mSE:* mean model standard error; *EmpSE:* mean empirical standard error.

Scenario 4

Here, we simulated data from **five studies**, each with **1000 patients.** For each simulated dataset, we completely removed X_1 from two studies (D_2 and D_5) in D_0. We present the simulation results in figure 3.6. FCS had almost identical results in bias, precision and coverage with Full Data and the best scenario comparing with scenarios 1, 2, 3.

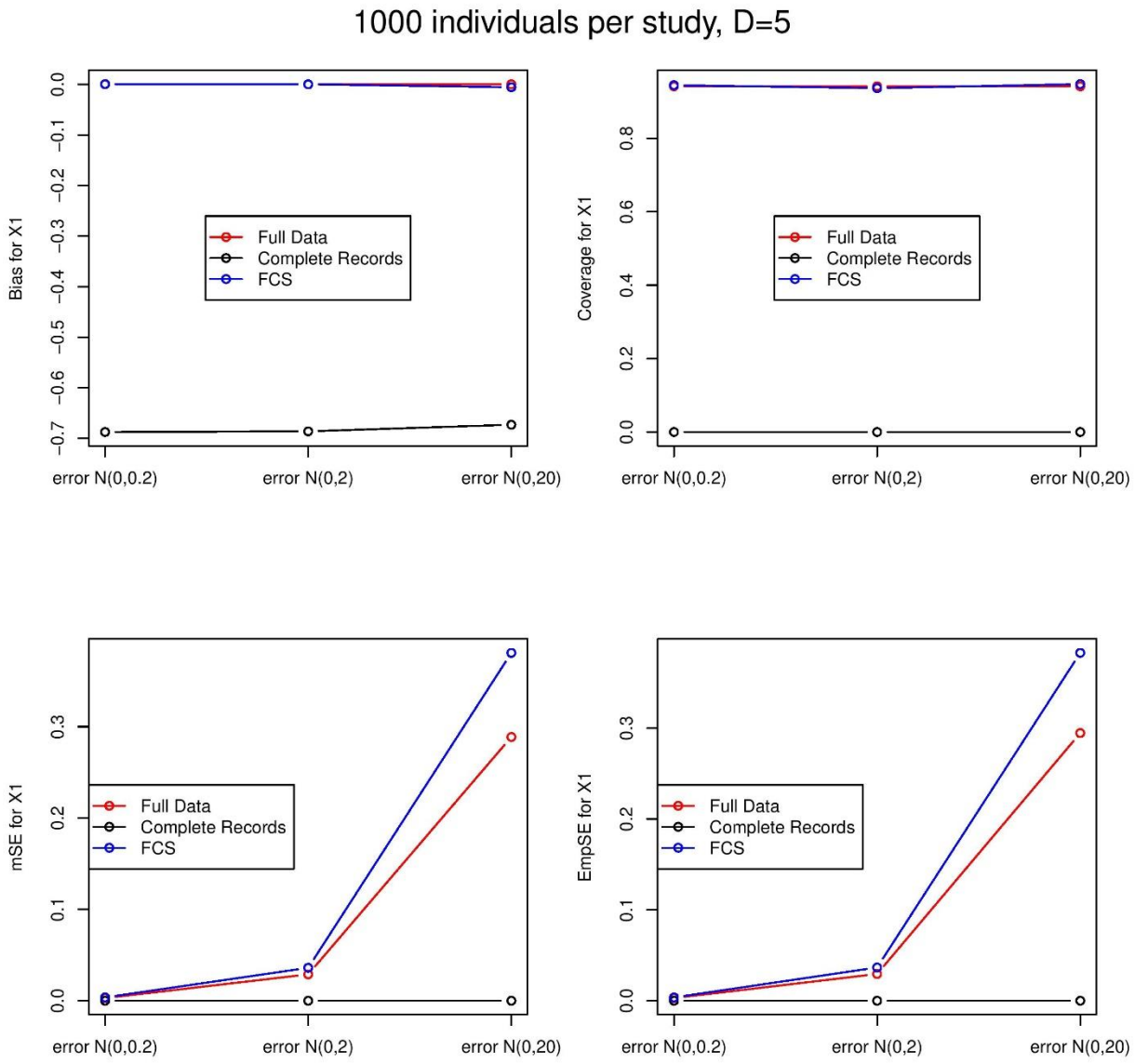


Figure 3.6. Main results from scenario 4's simulation study: Comparison of Bias, Coverage level, mSE and EmpSE for X_1 after 1000 simulations with Full Data (red), handling systematically missing values with Complete Records (black) and FCS (blue) for three model errors.

FCS: Multiple imputation by fully conditional specification; *D:* number of studies, *mSE:* mean model standard error; *EmpSE:* mean empirical standard error.

In figures 3.7 to 3.10 we compare Bias, Coverage, mSE and EmpSE for X_1 for scenario 1(D2_N200), scenario 3 (D5_N200), scenario 2 (D2_N1000) and scenario 4 (D5_N1000) for the three model errors.

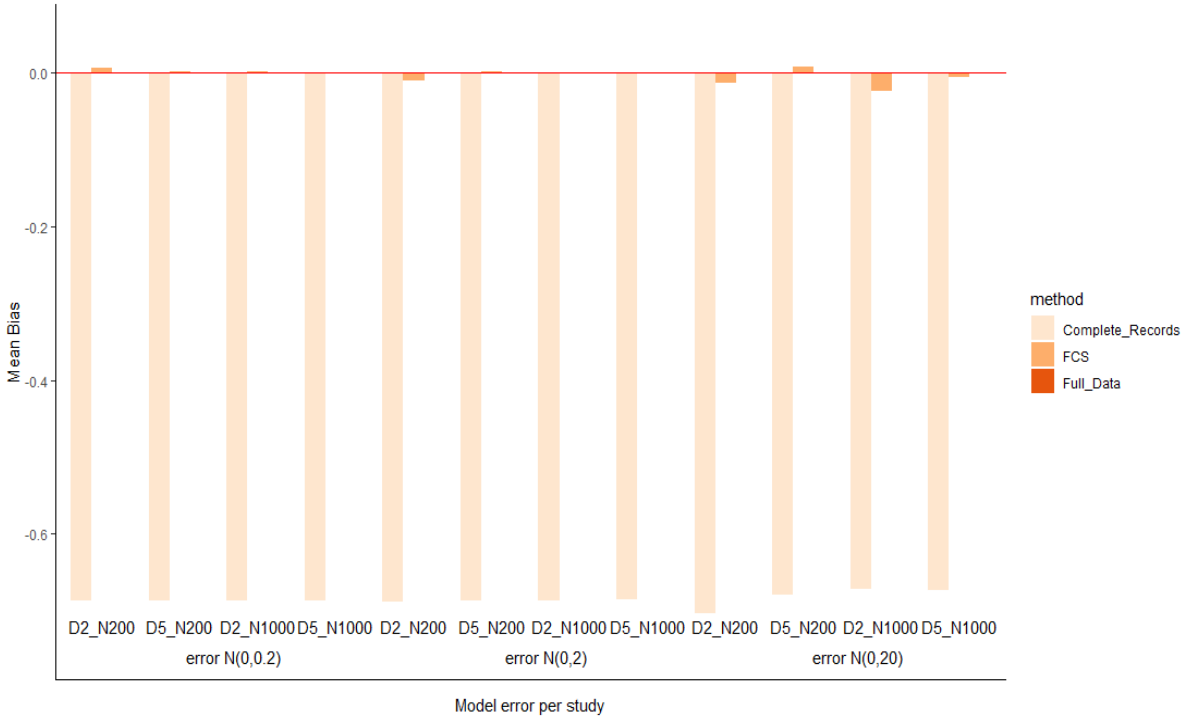


Figure 3.7. Bias for X_1 for '2 datasets, N=200 per dataset' (D2_N200), '5 datasets, N=200 per dataset' (D5_N200), '2 datasets, N=1000 per dataset' (D2_N1000), and (D5_N1000) '5 datasets, N=1000 per dataset' for three model errors.

FCS: Multiple imputation by fully conditional specification; *D:* number of studies, *N:* individuals per study.

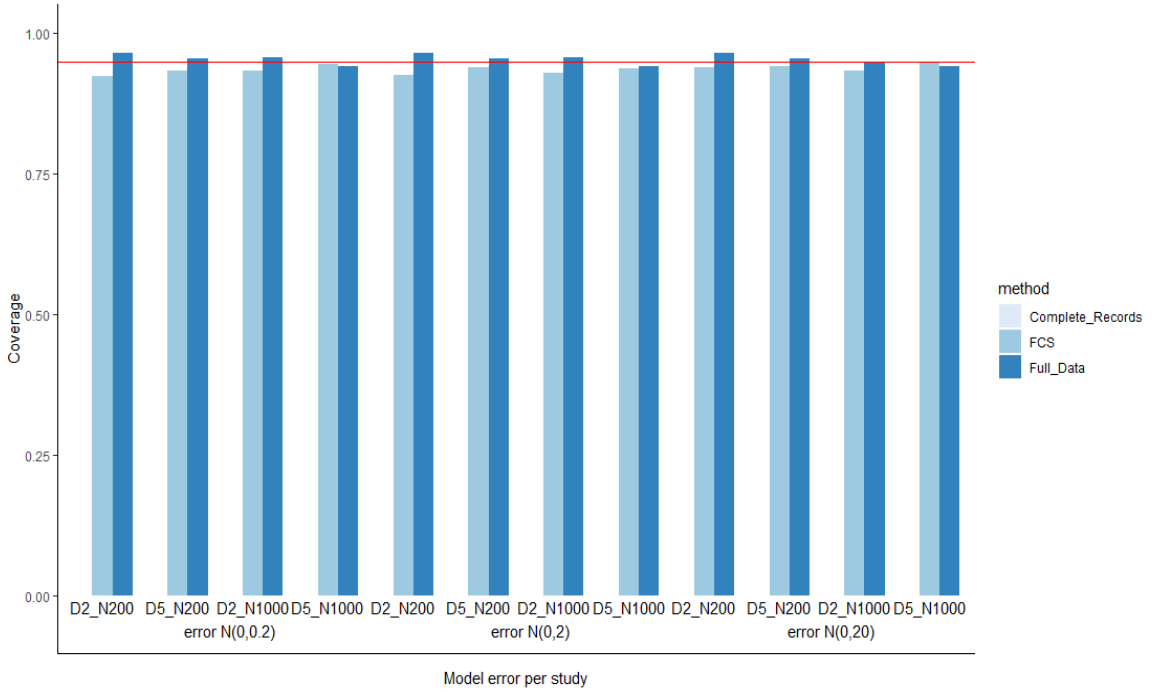


Figure 3.8. Coverage for X_1 for '2 datasets, N=200 per dataset' (D2_N200), '5 datasets, N=200 per dataset' (D5_N200), '2 datasets, N=1000 per dataset' (D2_N1000), and (D5_N1000) '5 datasets, N=1000 per dataset' for three model errors.

FCS: Multiple imputation by fully conditional specification; *D:* number of studies, *N:* individuals per study.

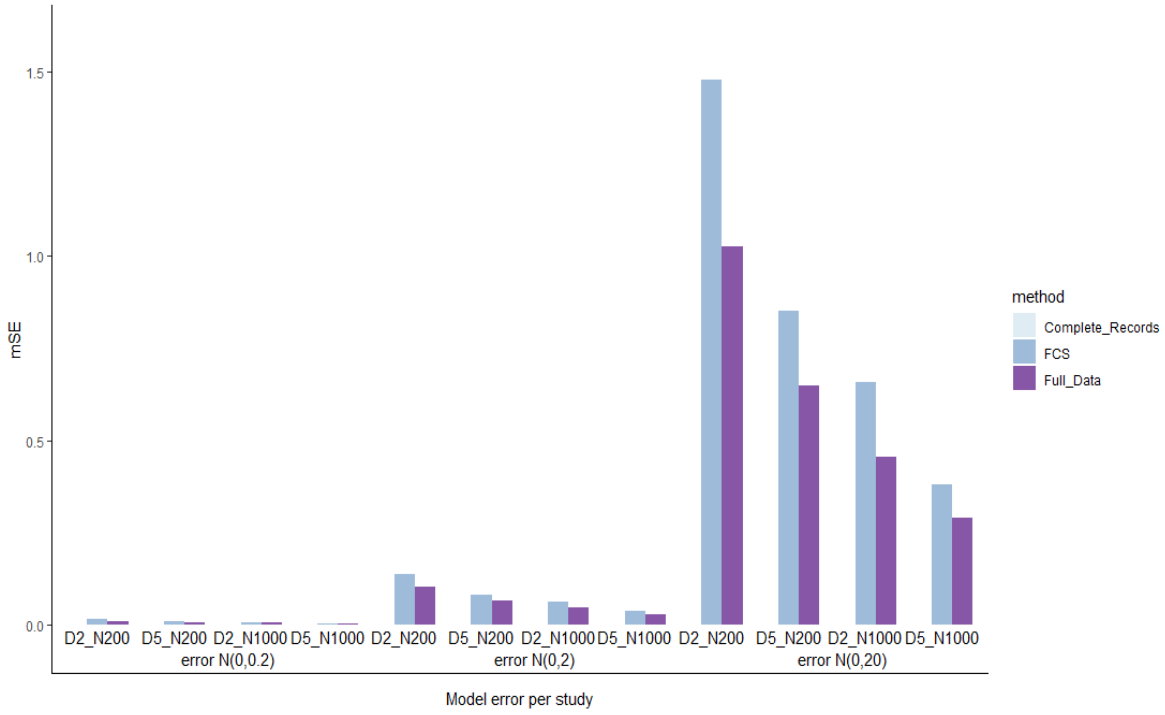


Figure 3.9. mSE for X_1 for '2 datasets, N=200 per dataset' (D2_N200), '5 datasets, N=200 per dataset' (D5_N200), '2 datasets, N=1000 per dataset' (D2_N1000), and (D5_N1000) '5 datasets, N=1000 per dataset' for three model errors.

FCS: Multiple imputation by fully conditional specification; *D:* number of studies, *N:* individuals per study; *mSE:* mean model standard error.

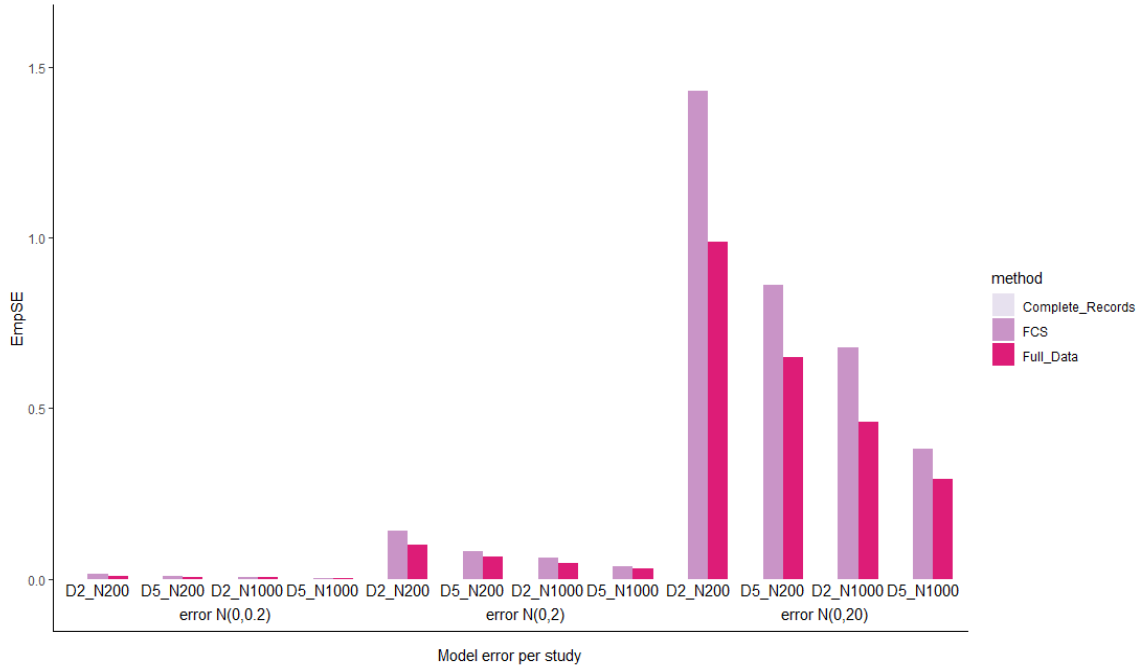


Figure 3.10. EmpSE for X_1 for '2 datasets, N=200 per dataset' (D2_N200), '5 datasets, N=200 per dataset' (D5_N200), '2 datasets, N=1000 per dataset' (D2_N1000), and (D5_N1000) '5 datasets, N=1000 per dataset' for three model errors.

FCS: Multiple imputation by fully conditional specification; *D:* number of studies, *N:* individuals per study; *EmpSE:* mean empirical standard error.

Simulations with studies of different sizes (Scenarios 5 - 6)

Scenario 5

We simulated data from five studies, each with different number of individuals (D_1:200, D_2:150, D_3:50, D_4:75, D_5:100). We completely removed X_1 from two studies (D_4 and D_5). The results are presented in figure 3.11.

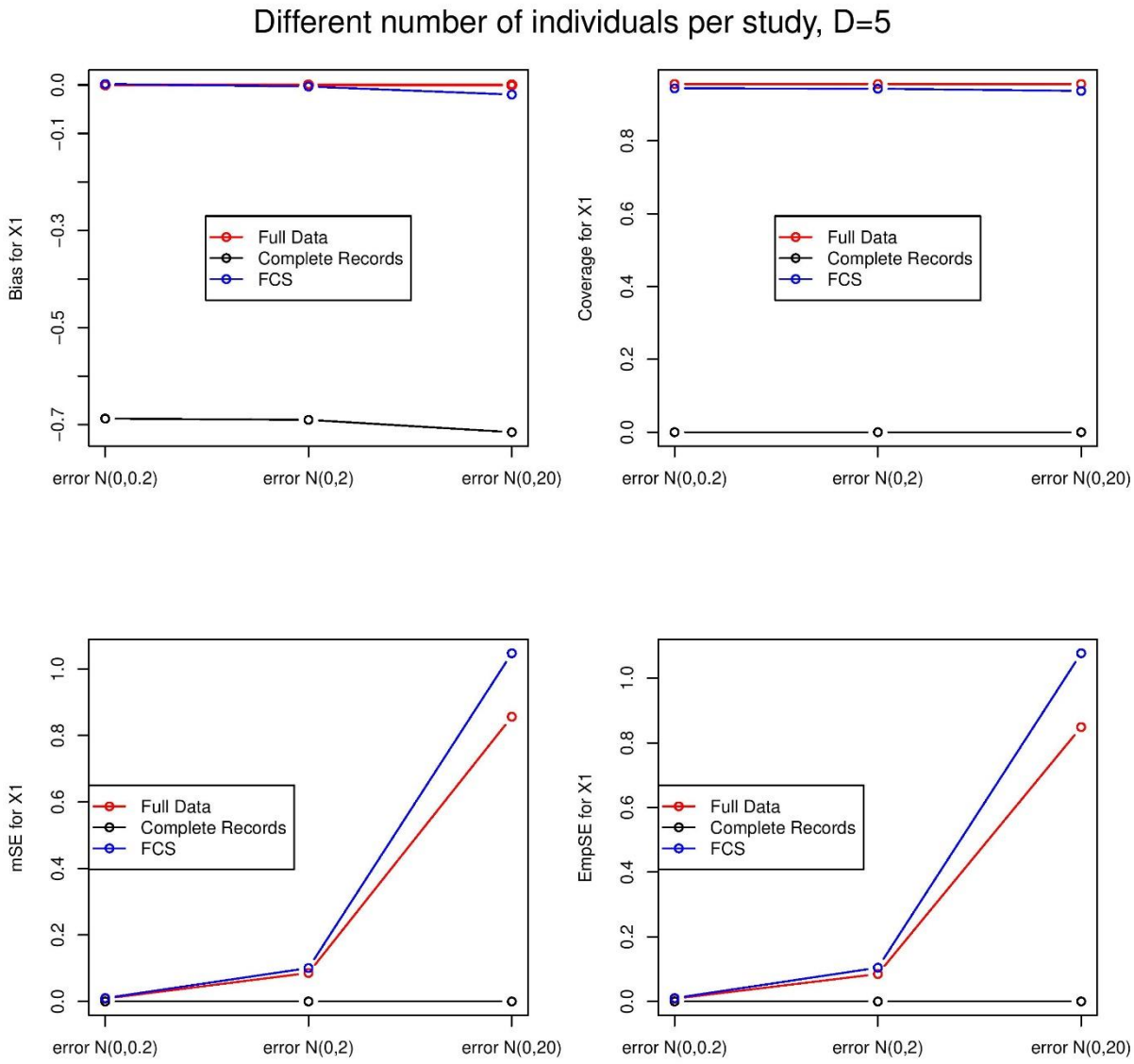


Figure 3.11. Main results from scenario 5's simulation study: Comparison of Bias, Coverage level, mSE and EmpSE for X_1 after 1000 simulations with Full Data (red), handling systematically missing values with Complete Records (black) and FCS (blue) for three model errors.

FCS: Multiple imputation by fully conditional specification; *D:* number of studies, *mSE:* mean model standard error; *EmpSE:* mean empirical standard error.

Scenario 6

We simulated data from ten studies, each with different number of individuals (D_1:800, D_2:150, D_3:50, D_4:75, D_5:350, D_6:200, D_7:150, D_8:500, D_9:750, D_{10}:100). We completely removed X_1 randomly from four studies (D_3, D_6, D_9, and D_{10}) with small, medium, and large size. Simulation's results are shown in figure 3.12.

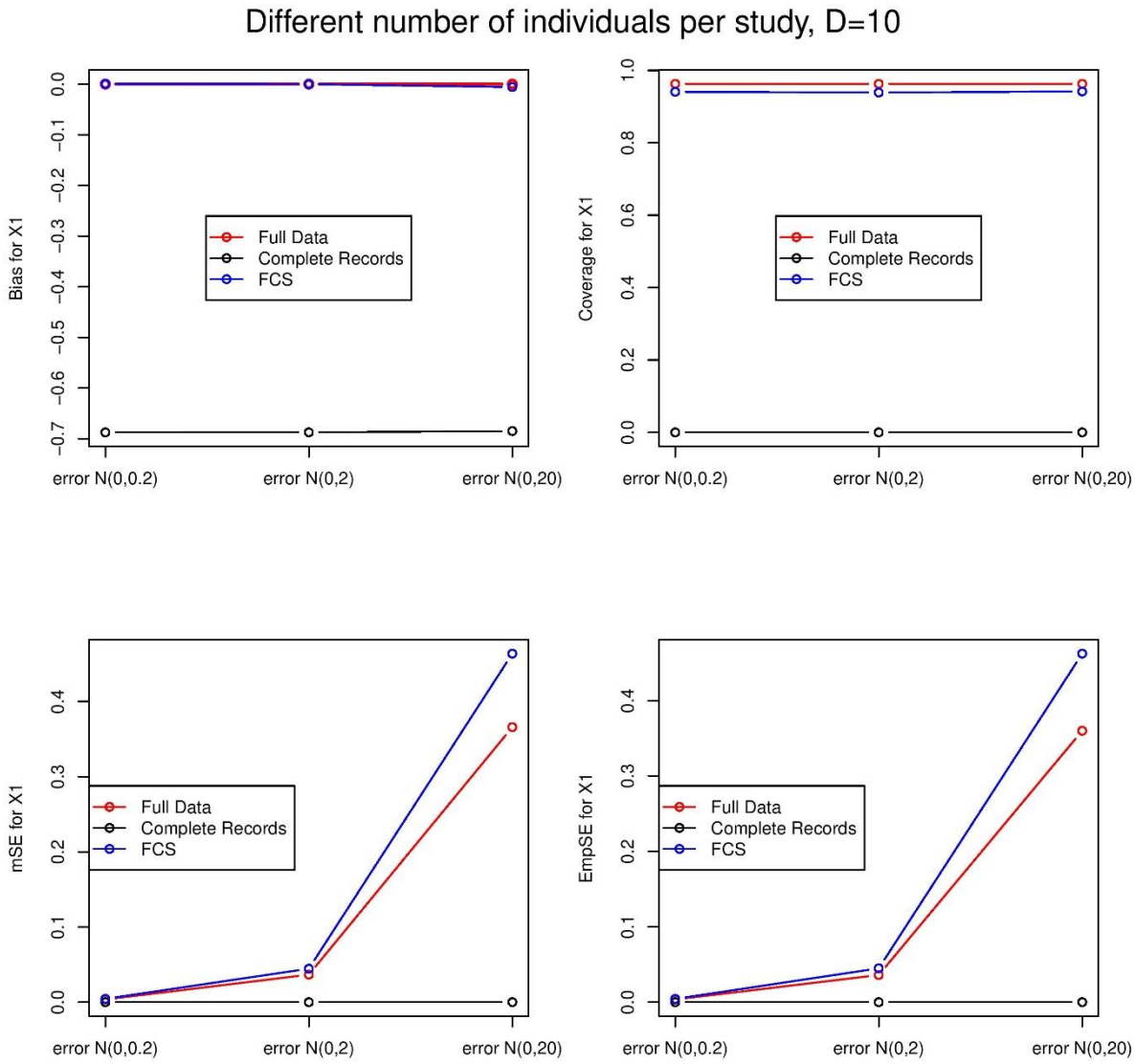


Figure 3.12. Main results from Scenario 6's simulation study: Comparison of Bias, Coverage level, mSE and EmpSE for X_1 after 1000 simulations with Full Data (red), Complete Records (black) - handling systematically missing values with complete case analysis and FCS (blue) - handling systematically missing values with FCS for three model errors.

FCS: Multiple imputation by fully conditional specification; *D:* number of studies, *mSE:* mean model standard error; *EmpSE:* mean empirical standard error.

In figures 3.13 to 3.16 we compare Bias, Coverage, mSE and EmpSE for X_1 for scenario 5 (D10_diffsize) and scenario 6 (D10_diffsize) for the three model errors. For scenarios 5 and 6, data simulated from correctly specified data generation model, we see no bias, correct coverage and minimal loss of information compared with the Complete Records analysis.

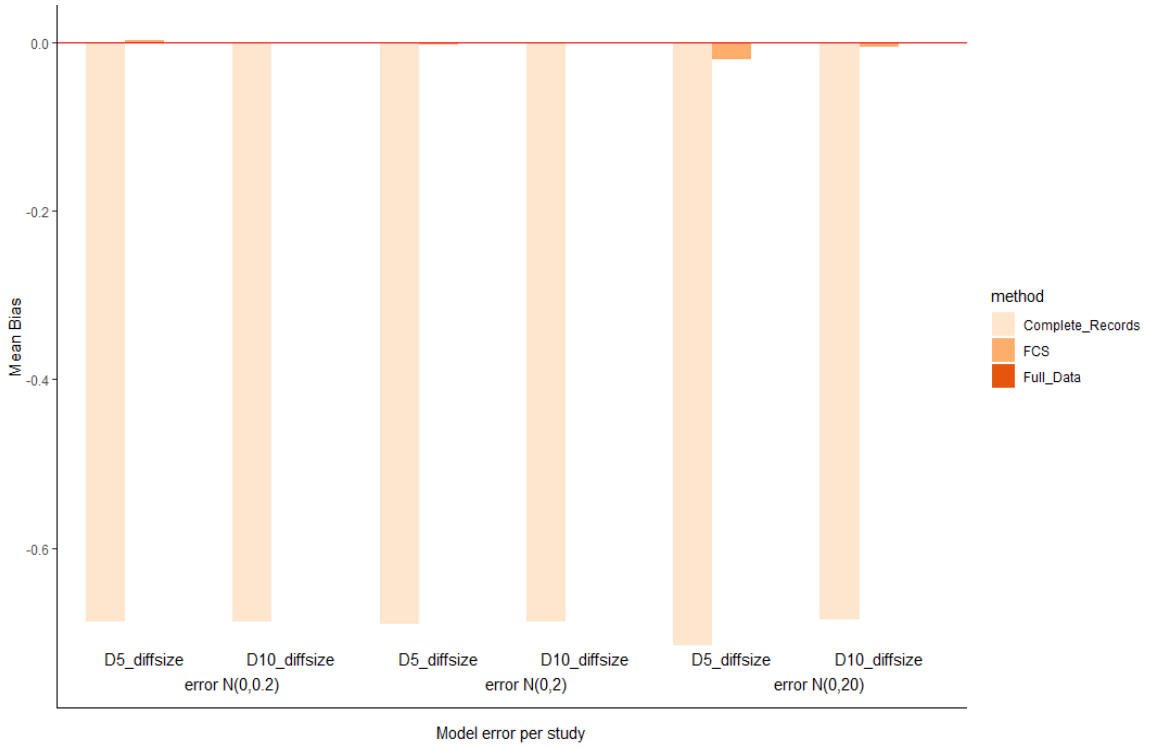


Figure 3.13. Bias for X_1 for '5 datasets, N=different per dataset' (D5_diffsize), '10 datasets, N=different per dataset' (D10_diffsize) for three model errors.

FCS: Multiple imputation by fully conditional specification; *D:* number of studies.

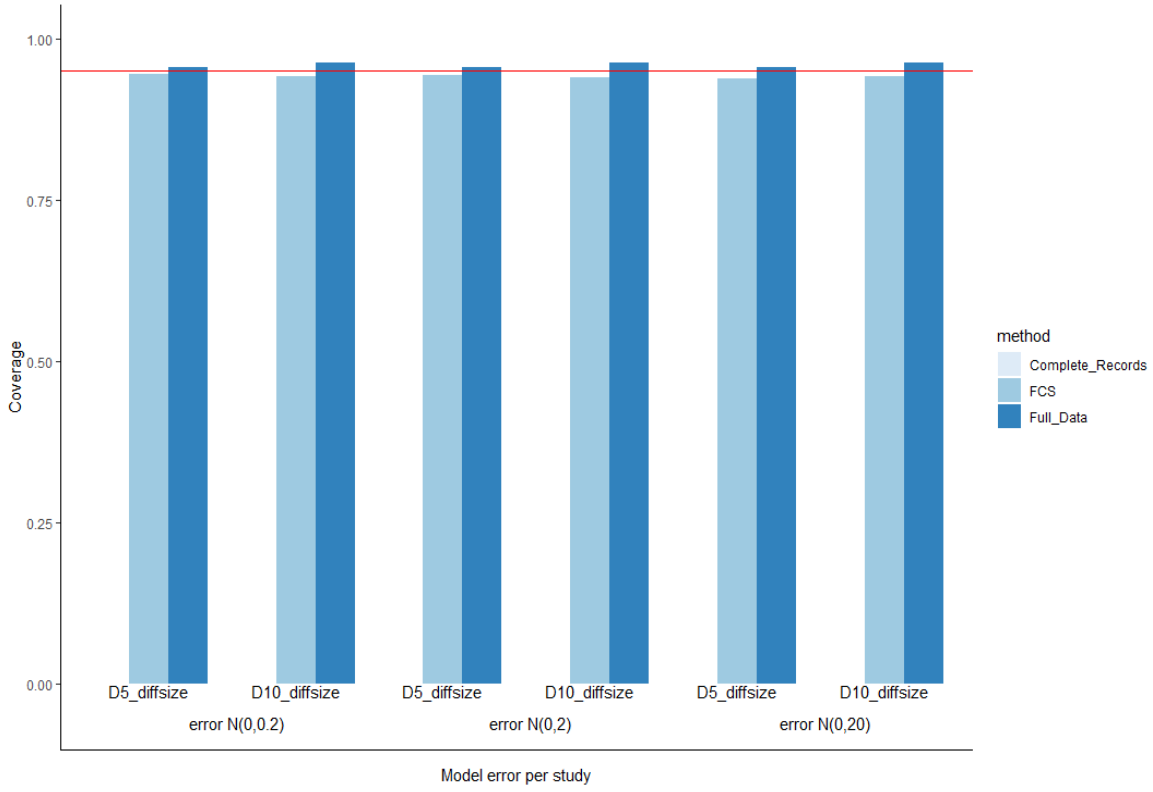


Figure 3.14. Coverage for X_1 for '5 datasets, N= different per dataset' (D5_diffsize), '10 datasets, N= different per dataset' (D10_diffsize) for three model errors.

FCS: Multiple imputation by fully conditional specification; *D:* number of studies.

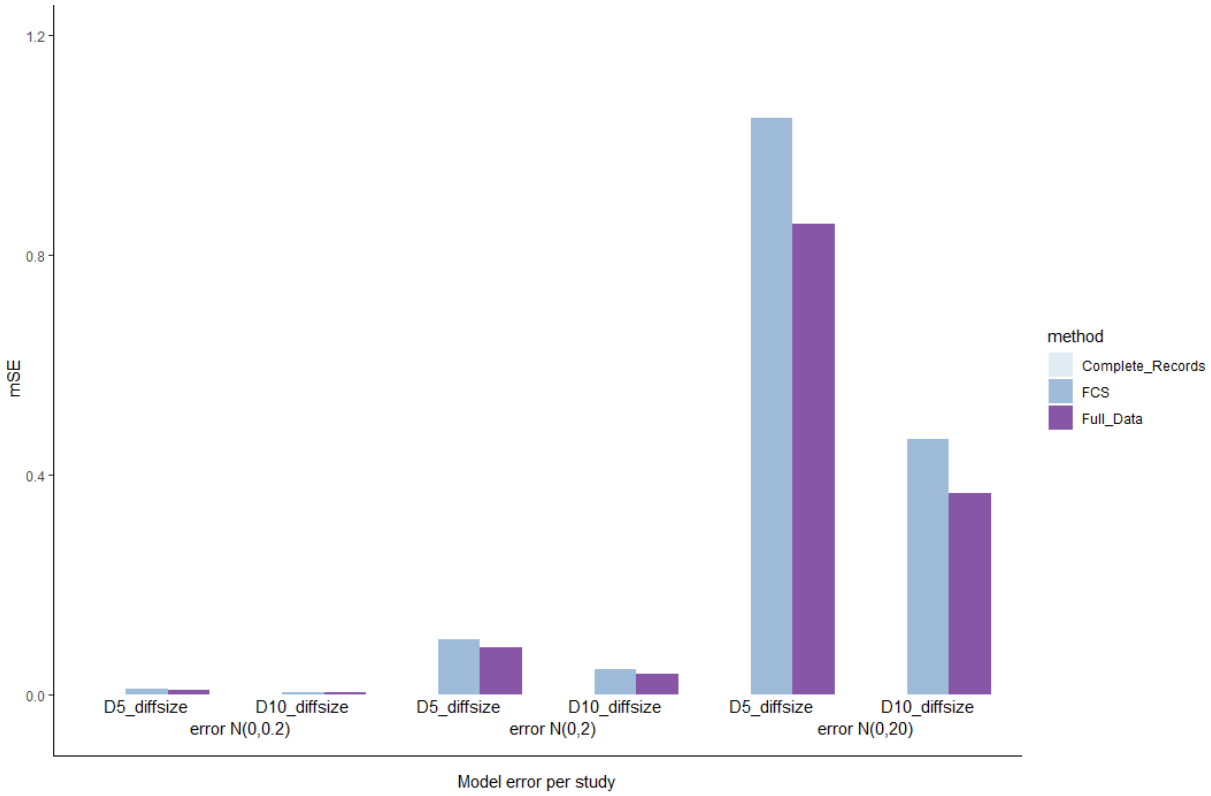


Figure 3.15. mSE for X_1 for '5 datasets, N= different per dataset' (D5_diffsize), '10 datasets, N= different per dataset' (D10_diffsize) for the three model errors.

FCS: Multiple imputation by fully conditional specification; *D:* number of studies, *mSE:* mean model standard error.

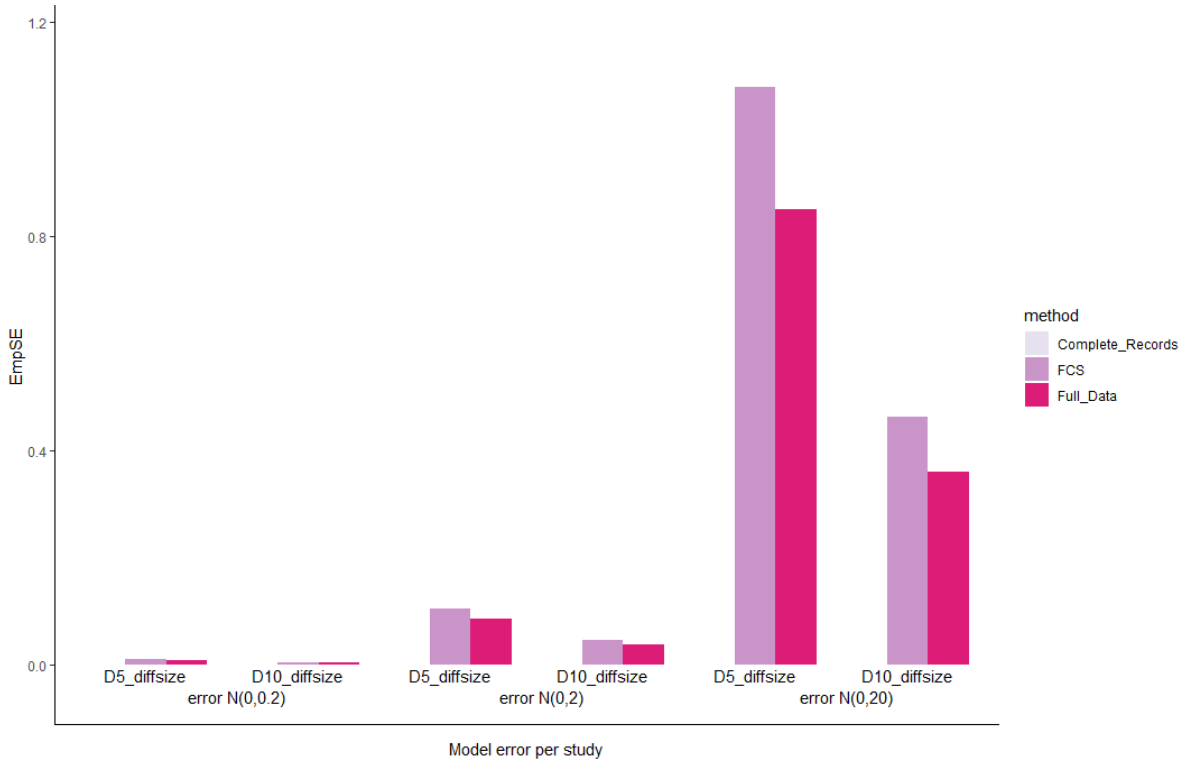


Figure 3.16. EmpSE for X_1 for '5 datasets, N= different per dataset' (D5_diffsize), '10 datasets, N= different per dataset' (D10_diffsize) for the three model errors.

FCS: Multiple imputation by fully conditional specification; *D:* number of studies, *EmpSE:* mean empirical standard error.

Simulations with same size studies, different model error per study (7 - 10)

Scenarios 7 – 10

Here, we again simulated data from 2 studies, each study with 100, 200, 500, 1000 individuals for each scenario respectively. We completely removed X_1 from a dataset (D_1). Results are shown in figure 3.17.

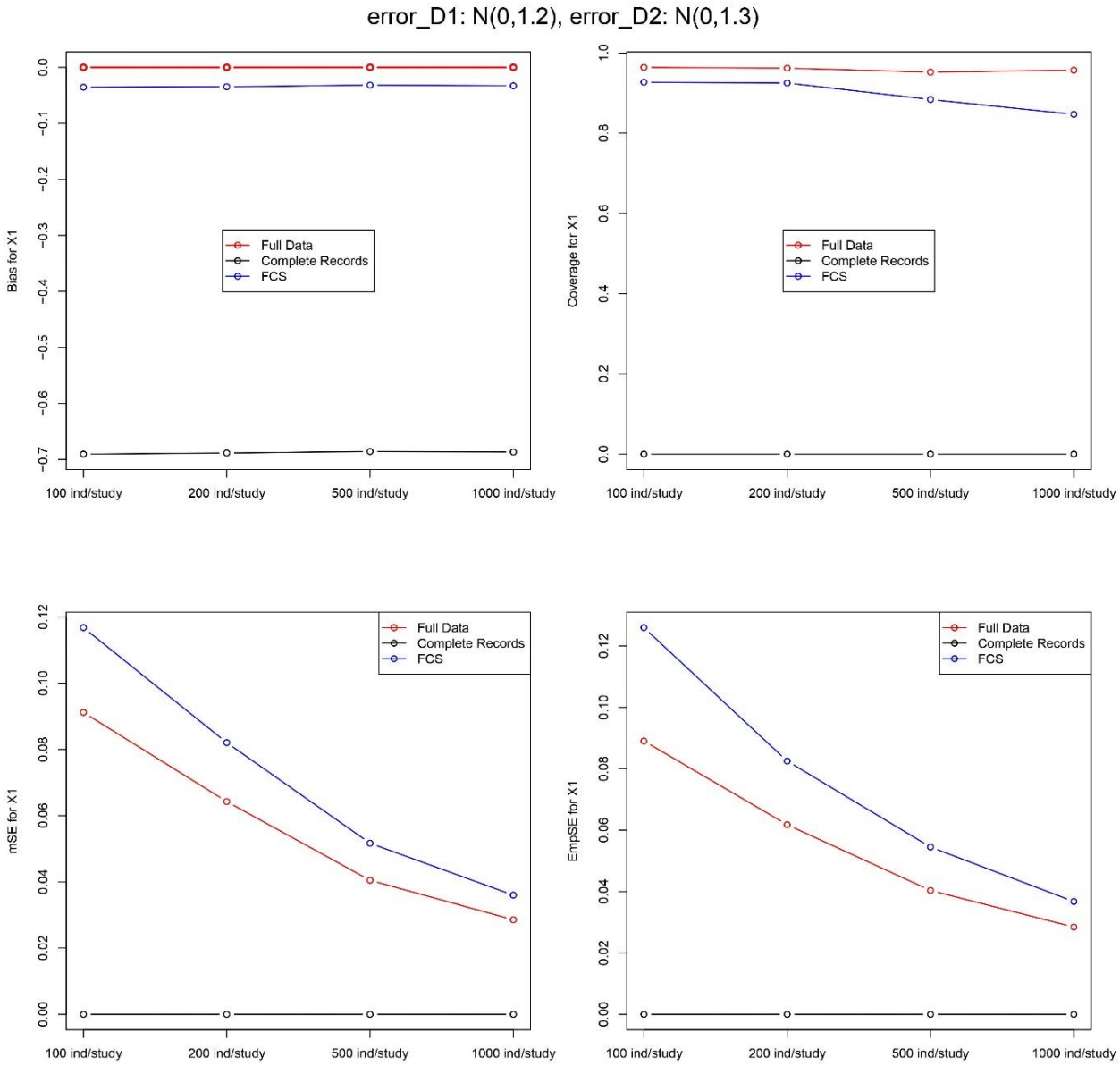


Figure 3.17. Main results from Scenario 7-10's simulation study: Comparison of Bias, Coverage level, mSE and EmpSE for X_1 after 1000 simulations with Full Data (red), Complete Records (black) - handling systematically missing values with complete case analysis and FCS (blue) - handling systematically missing values with FCS for three model errors.

FCS: Multiple imputation by fully conditional specification; *D:* number of studies, *mSE:* mean model standard error; *EmpSE:* mean empirical standard error.

3.4.5 Summary of findings from simulation studies

Drawing together the results from the simulation studies, we conclude the following:

Studies of same sizes: in our analyses we observed that when more studies are integrated, the smaller the bias in estimate. We also see that the larger the number of individuals per

study the smaller the bias. The fewer studies with systematically missing values were being integrated, the smaller the data missingness, therefore the estimates were more precise, and the coverage was better. The larger the model error per study, the larger the difference between FCS and Full Data in mSE and EmpSE. Another observation here is that EmpSE and mSE became smaller when individuals and number of studies were increased. We see the lowest EmpSE and mSE in Scenario 4 for all the three model errors.

Studies of different sizes

Similarly, here, the smaller the number of studies with systematically missing values, the smaller the data missingness, therefore more precise the estimates and better the coverage. Precision and good coverage is affected by the number of individuals with systematically missing data. The larger the study size the more accurate and closer to reality the estimates. The larger the model error per study, the larger the difference between FCS and Full Data in mSE and EmpSE.

Studies with same model errors

The data generation mechanisms are similar to the imputation models and there is no important bias introduced in any scenarios. Therefore, the closer the data generation mechanism is to the imputation model, the greater the gain in information through multiple imputation, with excellent results when they were the same.

Studies with different model errors

The results indicate that even when the FCS was producing slightly biased results (Scenarios 9-10) the results were much better and close to reality than complete case analysis. We see that when imputing with 500 and 1000 individuals per study, additional variability resulting in smaller standard errors and mild under-coverage.

Size of the model error

As expected, the smaller the model error per study the smaller the EmpSE and mSE. We also see that as the model error increased, the difference between EmpSE and mSE slightly increased. When the model error is e_i: N ~ (0, 20), we understand that outcome y was affected a lot by the model error. Therefore, model's coefficients estimate may have not affected the final results and may have not let us see how data missingness affected final analyses. With the same logic, when e_i: N ~ (0, 0.2), estimates played a very important role to the final outcome.

Overall

First, although realistic, our simulated scenarios cannot be exhaustive, and results may vary in alternative scenarios with different hypothesised associations between exposure, covariate and outcome and different distributions. Second, a sample of 1,000 per study might seem too large especially in scenario 4 (where we have 5000 individuals in total) if compared against trial data, but it was a necessity if we were to investigate very low and high model errors. The computational time led us to select our largest datasets to include 5,000 and 3,125 individuals. Unfortunately, this is not necessarily representative of a contemporary EHRs dataset which can hold hundreds of thousand or even millions of records. However, our method is based on FCS method have been thoroughly and routinely evaluated in the past in different datasets sizes. Therefore, we argue that we may provide an incomplete evaluation of the suggested method to solve systematically missing values problem in larger datasets.

We argue, as the simulations suggest that probabilistic approach is an accurate way to integrate structured healthcare data. When the data integration will be complex, a higher number of imputations may be needed. In general, the closer the data generation mechanism to the imputation model, the greater the gain in information through multiple imputation, with excellent results when they are the same. The probabilistic approaches gave valid results across a range of scenarios and in all cases outperformed the complete case analysis results and gave results close to real true data. It did not introduce practically important bias in any of the scenarios considered. Therefore, in applications the worst that may happen when taking this approach, is that inferences may be slightly conservative.

3.5 Application – MASTERPLANS exemplar

3.5.1 Data characteristics and systematically missing values problem description

Using three datasets in SLE, we describe and illustrate our approach to handling missing data problems in data integration.

To demonstrate the utility of the developed approach (figure 3.1) we applied this to real-world health datasets in SLE. For the datasets D_1, D_2, D_3, we have datasets that contain data from ALMS, LUNAR, and EXPLORER respectively. We applied our method to the three SLE datasets, by considering the research question ‘Finding the set of variables that best predict drug response’. To answer the research question, we included only ALMS-M (maintenance) information from the ALMS study data since we required patients to have a 12-month follow-up visit (with disease severity evaluated) which was not available in the

ALMS-I set (which followed patients for only 6 months). We rescaled the first visit as zero days and calculated days for each following visit relative to that. For this research question we used datasets that included patients from all the three studies together, with multiple visits per subject. For the response measure only, in order to have only one visit per patient, we kept the visit that had the least absolute difference from 365 days (12 month). The used dataset (table 3.4) consists of the 545 patients and 12 variables i.e., gender, age, ethnicity, height, weight, drug response (BILAGScore total), creatine, body mass index (BMI), current treatment, white blood cells (WBC), platelets, lymphocytes, and smoking status

Table 3.4. MASTERPLANS' data characteristics after integrating lupus studies ALMS, LUNAR, EXPLORER: systematically missing values.

Data characteristics	Integrated dataset N = 530 (%)	ALMS N = 204 (38.50)	LUNAR N = 127 (37.50)	EXPLORER N = 199 (24.00)
Age, years				
Mean ± SD	34.88 ± 11.71	31.64 ± 10.87	30.69 ± 9.27	40.88 ± 11.51
Mix – Max	12.00 – 71.00	12.00 – 64.00	17.00 – 56.00	18.00 – 71.00
Median (IQR)	34.00 (25.00 – 43.00)	32.00 (24.00 – 39.00)	30.00 (23.00 – 36.50)	41.00 (32.50 – 50.00)
BILAG score (total) (%)				
Mean ± SD	7.06 ± 6.26	4.76 ± 4.75	6.17 ± 6.35	10.00 ± 6.44
Min – Max	0.00 – 53.00	0.00 – 23.00	0.00 – 53.00	0.00 – 36.00
Median (IQR)	6.00 (2.00 – 10.00)	5.00 (1.00 – 6.00)	5.00 (2.00 – 8.00)	9.00 (5.00 – 13.00)
BMI, kg/m^2				
Mean ± SD	26.31 ± 6.48	24.04 ± 4.70	26.44 ± 5.55	28.56 ± 7.72
Min – Max	13.85 – 56.75	13.85 – 40.12	16.69 – 42.65	16.55 – 56.75

Median (IQR)	24.82 (21.64 – 29.55)	23.23 (20.34 – 26.86)	25.06 (22.45 – 29.55)	26.81 (23.07 – 32.84)
Creatinine, mg				
Mean ± SD	0.84 ± 0.31	0.85 ± 4.70	0.91 ± 0.42	0.80 ± 0.20
Min – Max	0.30 – 2.80	0.37 – 2.10	0.30 – 2.80	0.40 – 1.70
Median (IQR)	0.80 (0.67 – 0.96)	0.80 (0.68 – 0.95)	0.78 (0.65 – 1.03)	0.80 (0.70 – 0.90)
Ethnicity (%)				
Black or African American	100 (18.85)	22 (10.80)	33 (26.00)	45 (22.60)
Caucasian	251 (47.35)	90 (44.10)	41 (32.30)	120 (60.30)
Other	179 (33.80)	92 (45.10)	53 (41.70)	34 (17.10)
Gender (%)				
Female	466 (87.90)	175 (85.80)	112 (88.20)	179 (89.90)
Male	64 (12.10)	29 (14.20)	15 (11.80)	20 (10.10)
Height, cm				
Mean ± SD	163.09 ± 9.08	161.49 ± 9.38	163.24 ± 8.85	164.65 ± 8.67
Min – Max	132.00 – 198.10	132.00 – 191.00	143.50 – 195.60	141.00 – 198.10
Median (IQR)	162.50 (157.00 – 168.90)	160.00 (156.00 – 166.00)	162.50 (157.50 – 168.00)	163.80 (159.20 – 170.20)
Lymphocytes, (K/cmm)				
Mean ± SD	1.16 ± 0.59		1.23 ± 0.62	1.11 ± 0.57
Min – Max	0.15 – 3.52		0.18 – 3.52	0.15 – 3.30
Median (IQR)	1.01 (0.15 – 1.48)		1.14 (0.79 – 1.52)	0.98 (0.66 – 1.44)

Missing data	215	204	4	7
Platelets, (K/cmm)				
Mean ± SD	292.30 ± 92.21		315.46 ± 90.60	276.90 ± 90.03
Min – Max	48.00 – 703.00		151.00 – 703.00	48.00 – 679.00
Median (IQR)	283.50 (233.20 – 240.50)		298.50 (255.00 – 350.80)	262.50 (219.00 – 327.00)
Missing data	224	204	5	15
Smoking status (%)				
Current	51 (9.60)		10 (7.87)	41 (20.6)
Never	209 (39.40)		101 (79.50)	108 (54.3)
Previous	66 (12.50)		16 (12.63)	50 (25.1)
Missing data	204 (38.50)	204 (100.00)		
Treatment (%)				
Placebo + AZA OR AZA	124 (23.40)	98 (48.00)		26 (13.10)
Placebo + MMF OR MMF	189 (35.70)	106 (52.00)	59 (46.50)	24 (12.10)
Placebo + MTX OR MTX	19 (3.58)			19 (9.50)
RITUX + AZA	42 (7.92)			42 (21.10)
RITUX + MMF	123 (23.20)		68 (53.50)	55 (27.60)
RITUX + MTX	33 (6.20)			33 (16.60)
WBC, (K/cmm)				
Mean ± SD	6.32 ± 6.87		6.60 ± 2.76	6.15 ± 2.93

Min – Max	0.96 – 17.72		0.96 – 16.60	1.57 – 17.72
Median (IQR)	5.72 (4.17 – 7.79)		6.06 (4.75 – 8.19)	5.61 (3.98 – 7.69)
Missing data	214	204	4	6
Weight, kg				
Mean ± SD	70.33 ± 19.61	63.05 ± 15.17	70.58 ± 16.38	77.63 ± 22.65
Min – Max	34.20 – 156.63	34.20 – 114.30	41.77 – 120.31	42.00 – 156.63
Median (IQR)	65.62 (55.74 – 80.81)	61.10 (51.80 – 71.33)	67.50 (57.66 – 81.27)	74.20 (60.20 – 90.71)

Content heterogeneity occurs because white blood cells (WBC), platelets, lymphocytes, and smoking status were not recorded in ALMS. We removed records for 7 patients with missing value for height and 9 with missing creatinine (figure 3.18) leading to 530 patients to keep things clear for the illustration of the data integration problem (figure 3.19). In figure 3.20. we can see systematically missing values' visualisation across the three integrated lupus dataset.

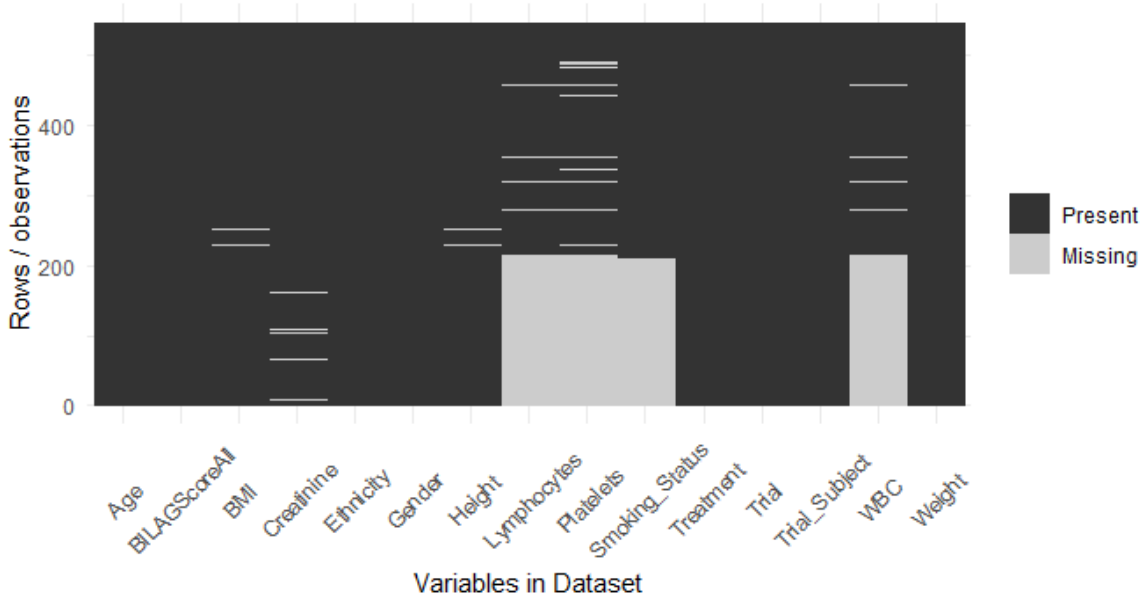


Figure 3.18. Visualisation of missing data in the integrated lupus data before we remove 15 patients' records to keep things feasible and make it easier to illustrate the problem (total patients: 545).

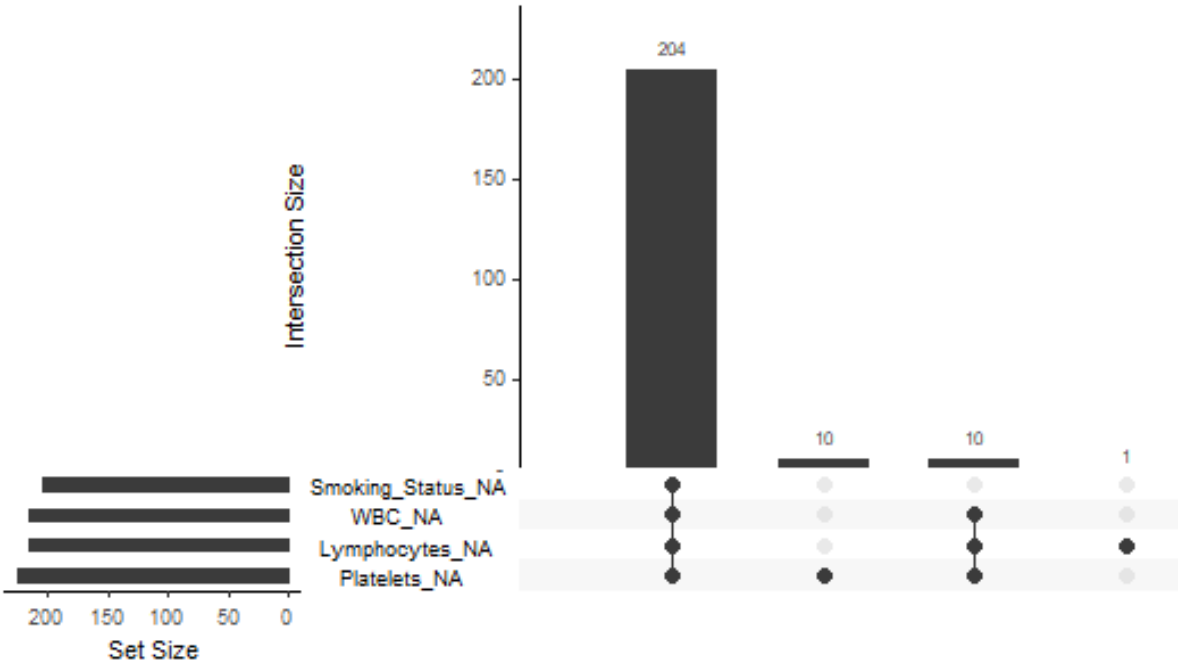


Figure 3.19. Visualisation of missing data in the integrated lupus data (total patients: 530).

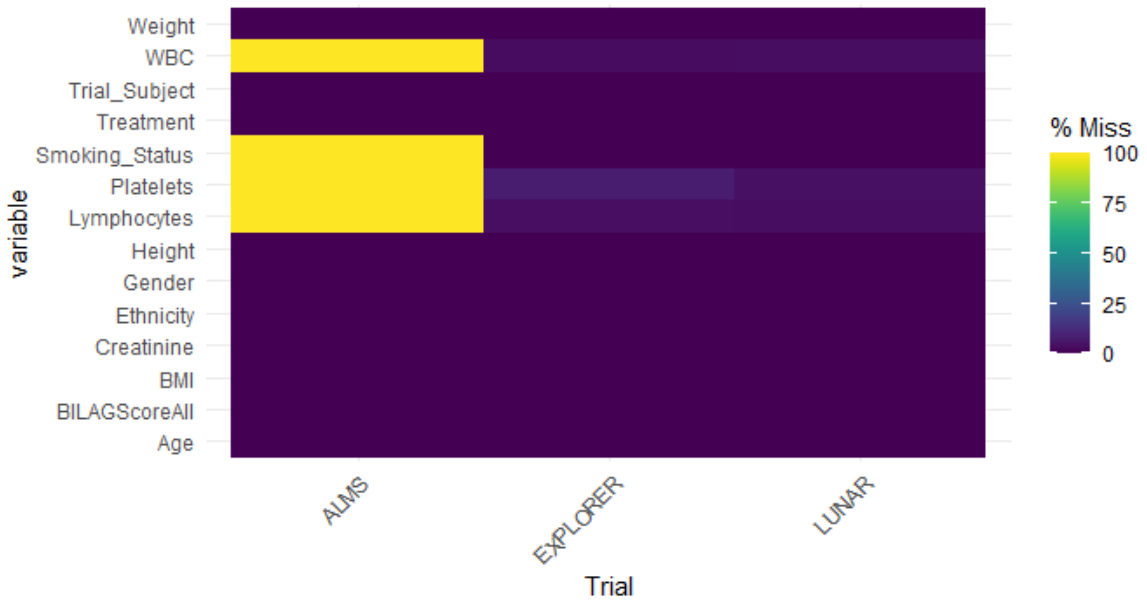


Figure 3.20. Systematically missing values' visualisation across the integrated lupus dataset (ALMS, LUNAR, EXPLORER). Yellow colour shows the systematically missing values.

3.5.2 Results

Data analysis of real-world lupus data helped us find which variables best predicted the outcome - BILAG response. We performed both traditional and probabilistic data integration in lupus datasets to make things comparable. We included enough variables that would help us compare the different methods.

3.5.2.1 Traditional Data Integration – complete case analysis

Taking the 'traditional' approach to data integration, the complete case analysis, we applied a prediction model of the disease measure BILAG, using the predictor variables gender, ethnicity, BMI, treatment, and creatinine (equation 3.10). Table 3.5 shows the coefficients for the linear regression model that estimate the drug response based on equation 3.10.

$$BILAG\ Score \sim Gender + Ethnicity + BMI + Treatment + Creatinine \quad (3.10)$$

Table 3.5. Coefficients (estimate, standard error, t statistic and p-value) for linear regression model from equation 3.10 after applying complete case analysis in SLE data.

	estimate	**standard error**	**t statistic**	**p-value**
(Intercept)	4.5579	1.5228	2.9930	0.0029 **
Ethnicity				
Caucasian	1.2472	0.7345	1.6980	0.0901
Other	-1.1860	0.7817	-1.5170	0.1299
Gender				
Male	-1.9100	0.8353	-2.2870	0.0226 *
BMI	0.0302	0.0427	0.7080	0.4791
Treatment				
Placebo + MMF OR MMF	0.4858	0.7062	0.6880	0.4919
Placebo + MTX OR MTX	3.6544	1.5218	2.4010	0.0167 *
RITUX + AZA	2.5395	1.1127	2.2820	0.0229 *
RITUX + MMF	1.8638	0.7838	2.3780	0.0178 *
RITUX + MTX	2.5681	1.2055	2.1300	0.0336 *
Creatinine	0.7724	0.8900	0.8680	0.3859

*Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

3.5.2.2 Probabilistic Data Integration – multiple imputation

Here, we took the alternative approach to data integration, the probabilistic approach, and we applied a prediction model of the disease measure BILAG, using the predictor variables gender, ethnicity, smoking status, BMI, treatment, creatinine, lymphocytes, and platelets (equation 3.11).

$$BILAG\ Score \sim Gender + Ethnicity + Smoking\ Status + BMI + Treatment + Creatinine + Lymphocytes + Platelets \quad (3.11)$$

Figure 3.21 shows the density distributions of WBC, lymphocytes, and platelets. Results show that density of the imputed datasets (showed in magenta colour) matched the density of the observed data (showed in blue colour). The density distributions appeared to be very similar to the observed data for the three variables. We argue that extreme values affected the shape of the plots. However, the central tendencies of the density plots of imputed data appeared relatively similar to the observed ones. Figure 3.22 presents a barplot for the categorical variable 'Smoking Status'. We see that the observed data in the top figure match completely the imputed data with FCS that are in the bottom figure.

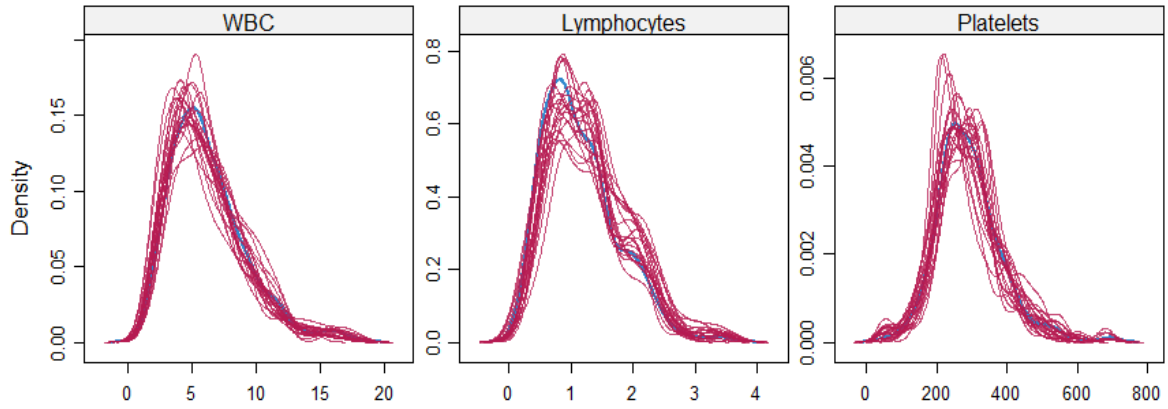


Figure 3.21. Density plots for the variables: 'WBC', 'Lymphocytes', 'Platelets', in content heterogeneity problem 1. Blue line shows the observed data and the magenta lines the imputed data from each of the imputations in FCS.

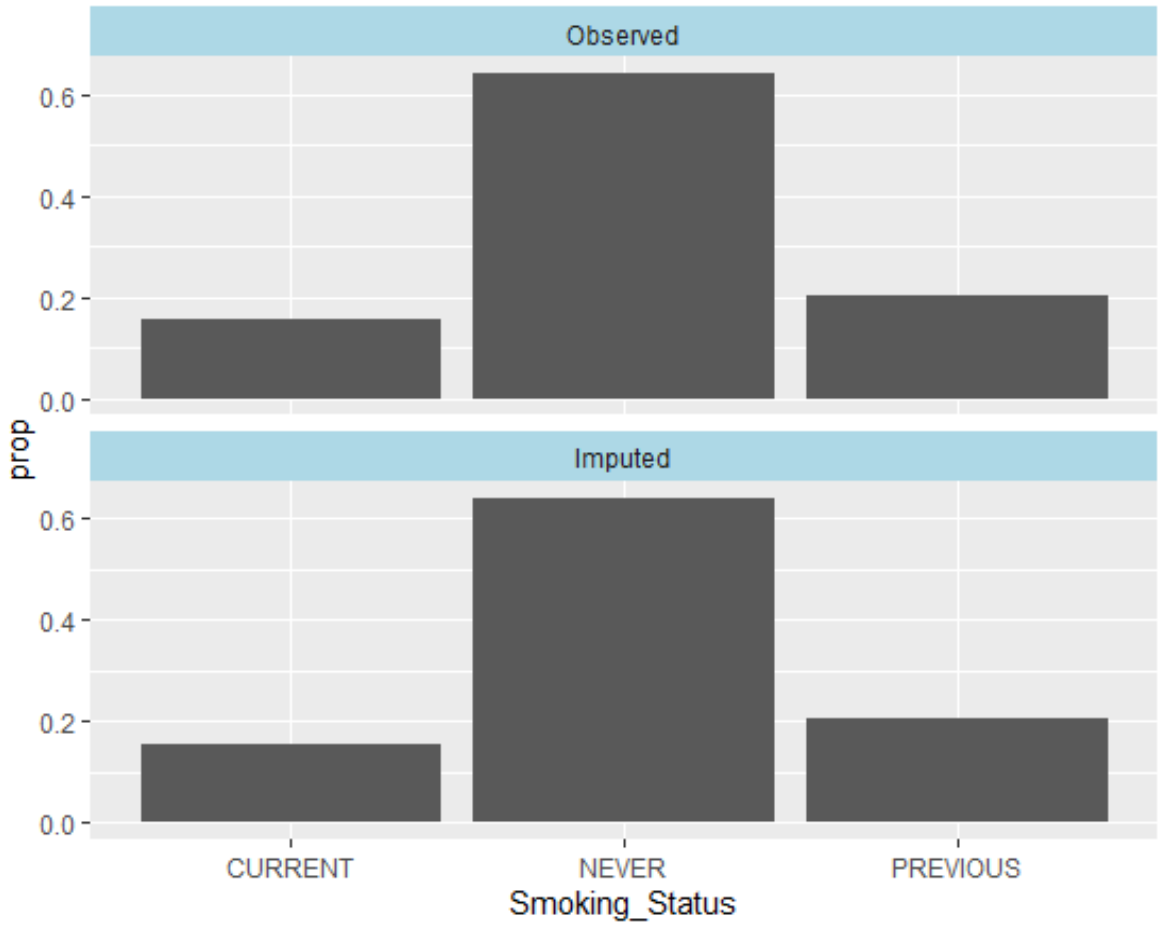


Figure 3.22. Barplot for the variable: 'Smoking Status', in content heterogeneity problem 1. Top figure shows the observed values and bottom figure shows the imputed data for each imputation in FCS.

In Table 3.6 we show the coefficients (estimate, standard error, t statistic and p-values) for linear regression model obtained from equation 3.11 after applying FCS in SLE data to solve the systematically missing values problem. Based on the results presented in tables 3.5 and 3.6, we see agreement between FCS and complete case analysis in inclusion of gender and treatment in prediction models. FCS's results showed a positive effect on outcome response when platelets increased and lymphocytes decreased. In both methods we see that BMI and creatinine showed no significance.

Table 3.6. Coefficients (estimate, standard error, t statistic and p-values) for linear regression model from equation 3.11 after applying FCS in SLE data.

	estimate	**standard error**	**t statistic**	**p-value**
(Intercept)	5.4191	1.9003	2.8517	0.0047 **
Ethnicity				

Caucasian	1.0537	0.7428	1.4185	0.1567
Other	-0.8130	0.8046	-1.0104	0.3128
Gender				
Male	-1.9558	0.8653	-2.2602	0.0243 *
Smoking Status				
Never	-0.8280	0.8644	-0.9579	0.3390
Previous	-1.0076	1.0362	-0.9724	0.3321
BMI	0.0484	0.0433	1.1180	0.2641
Treatment				
Placebo + MMF OR MMF	0.0795	0.7667	0.1037	0.9174
Placebo + MTX OR MTX	3.8444	1.5216	2.5266	0.0118 *
RITUX + AZA	1.9177	1.1503	1.6672	0.0963
RITUX + MMF	1.1892	0.8445	1.4082	0.1601
RITUX + MTX	2.5187	1.2096	2.0824	0.0378 *
Creatinine	0.4962	0.9051	0.5482	0.5838
Lymphocytes	-2.1851	0.5653	-3.8654	0.0002 ***
Platelets	0.0089	0.0037	2.3775	0.0191 **

*Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

3.6 Discussion

The problems of heterogeneous data integration produce a large gap between the potential of Big Data and its realisation in biomedical and public health research. In this chapter, we introduced the first problem of content heterogeneity; this is the existence of variables in at least one dataset but not in other datasets. Traditionally, statisticians, data scientists and informaticians have sought to solve this problem by applying complete case analysis.

The goal of this research was to argue that a data integration problem - systematically missing values problem - could be regarded as a missing data problem, and therefore can be solved with established probabilistic methods such as multiple imputation. We described our suggested probabilistic solution employing FCS and reported the results of a comprehensive series of simulation studies with different model errors, number of studies, individuals per study and study size to investigate the validity of the imputation method when we want to ask research questions that include systematically missing values in the data model. We further applied our probabilistic data integration approach to real-world lupus data to identify variables that best predict the outcome - BILAG response.

The simulation studies and the application of our approach in real-world data showed that systematically missing values is not a problem if we apply established imputation methods like FCS in order to use as much as data as possible and to answer different research questions. The closer the data generation mechanism is to the imputation model, the greater the gain in information through multiple imputation, with excellent results when they are the same [97].

In a stacked dataset, it is common to assume that observations in a single, individual dataset are more similar than observations from different individual datasets. In statistical terms, we should not assume that the observations within a single dataset are *entirely independent (i.i.d.)* but that they are, to some degree, correlated. Multilevel models allow us to do exactly that. As far as we are aware, multilevel imputation is not commonly used but it would be the better options for sporadically missing data. However, when for example we have two datasets and data are systematically missing in one of these, then we cannot use multilevel imputation because there is then not sufficient information to estimate higher-level coefficients. So, in theory this would be better, but because of the nature of the problem that we are trying to solve (systematically, not sporadically, missing data), we cannot use it. What this means is that we have to assume that all the observations are i.i.d. This is a strong assumption, we have to acknowledge that we make that assumption, and it is a challenging one. But without making the assumption, there would be no inference.

Therefore, for future work an important key challenge that we need to address patients from a single centre/study are more likely to be similar to each other than to patients from different centres/studies. This phenomenon is known in the statistical literature as 'clustering' and is apparent from many multi-centre studies and meta-analyses. This is likely due to variation across centres/studies that arises from residual, unmeasured confounding. The common approach to account for clustering is using multi-level regression analysis [113] in which random effects are used to model centre-level and study-level variation.

Ideally, shared data models would be implemented at source, enabling uniform data collection at different sites and studies. But in reality, data standardisation is always imperfect, and our approach embraces this imperfection rather than trying to extinguish it. Future work includes expanding of the general applicability of the method by creating a statistical package and applying the methods to real-world biomedical and health datasets such as asthma data. To our own knowledge there is very limited research on multilevel multiple imputation [103], [114]–[120] and very few recent papers that focus on the issue of systematically missing values [97], [105], [114], [121]–[123]. They agree that simulated

data under a correct specified data generation model, they have no bias, high and correct coverage and minor information loss compared with the complete case analysis.

Our results suggest our approach can be used to impute completely systematically missing values in some studies and agree with other recent research work [97], [114]. We recommend our approach implemented, for multiple imputation of healthcare structured data integration like that illustrated in Section 3.5 with the MASTERPLANS data. We suggest that the researchers and users in general should take advantage of the data that they have and choose to investigate probabilistic methods to solve systematically missing values problems. We explored different scenarios and the results indicated very good results and precise estimates.

Thus, we are very motivated to suggest FCS as a very promising solution to solve content heterogeneity problem resulting from data integration in healthcare. Multiple imputation has been explored by many researchers and have been gaining popularity for handling missing data due to its flexibility and practicality. In the case of imputing systematically missing values, we have described and evaluated a probabilistic approach. This alternative solution extends and overcomes current approaches such as complete case analysis. This probabilistic approach allows answering of research questions when one or more variables were not collected in one/some studies. Our simulation evaluation of this approach and its application to real data show promising results, and we hope it will be a useful addition to biomedical research and health data science.

Chapter 4: Varying granularity of categorical variables

4.1 Introduction

In Chapter 3, we focused on the first type of content heterogeneity that may exist in integrated datasets, this being systematically missing variables. In this chapter we address a second type of content heterogeneity; one that arises from differences in a variable's granularity. First, this specific problem is identified and defined (Section 4.2). The second part of this chapter outlines the theoretical solution using probabilistic approaches to data harmonisation in generic terms (Section 4.3). Afterwards, we evaluate the suggested methods for addressing the granularity problem through simulation studies (Section 4.4). Then, we apply our methods in the MASTERPLANS data, and discuss the obtained results (Section 4.5). Finally, the utility of the probabilistic data integration approaches to solve granularity issues is discussed and suggestions are made for future research (Section 4.6).

4.2 Problem identification of varying granularity of categorical variables

Data integration would be easier if the same data standards were applied in the data sources. A pre-alignment solution like this is suggested by Clinical Data Interchange Standards Consortium [124] about data collection, capture, reusability, and interoperability of data. Ideally, shared data models would be implemented at source, enabling uniform data collection at different sites and studies. In an ideal world, information would be a pro-alignment harmonisation where emerging studies would use integrated studies' questionnaires and standard operating processes procedures [125], [126]. But in reality, it is also significant to make use of the existing vast available information and increase the utility of 'post-alignment' harmonised data [127], [128]. One of the most significant problems when it comes to structured data integration and harmonisation is granularity.

There are many different definitions of granularity (in the context of data). We will focus on a specific kind. Granularity, is the level of detail at which data are stored in a data source. When variables representing the same information, across the different data sources are represented in multiple different levels/categories, then we have inconsistent granularity. For example, in ethnicity, some datasets include more categories and subsets than others, or age is captured as categories (i.e. 0-20, 21-40, 41-60,>60) in one study but slightly different categories (0-20, 21-40, 41-60,61-80, >80) in another; Therefore, we must deal with content heterogeneity due to differences in granularity.

The recent years there have been some efforts in health and biomedical sciences to solve this integration challenge when sharing clinical information across settings [129]. Research

revealed that only 17% was compatible among the data sources and those data differences in granularity and missing data cause a higher risk in medical errors, increased costs in integration and not useful usage of clinical time [130].

Traditional approaches for resolving granularity map all source datasets to a common data model that includes only high-level items, and thus omit all items that vary between datasets. For example, in the case described above, integration would be achieved by only including the lowest levels that are common in the harmonised dataset (i.e. 0-20, 21-40, 41-60,>60). Aligning variables to a common data model based on similarity takes too much valuable time and requires a manual approach which needs human knowledge, experience and critical thinking [129]. Therefore, this type of data harmonisation is not a standardised process and could lead to a potential analyst bias due to the need of human interpretation and absence of documentation and extensive data dictionaries. Le Sueur et al. [129] had this inconsistent granularity levels while integrating SLE data from different cohort studies. A limitation of their approach is the inevitable loss of information either due to differences in granularity or data capture.

Additionally, based on the example on ethnicity's granularity, researchers have focused on the importance of demographic data to capture population heterogeneity to identify health needs of diverse groups, to detect and address inequities in healthcare provision and outcomes. However, little is known about current methods of ethnic classification internationally and, in countries where ethnicity data are collected, about what level of granularity is employed in their ethnicity categorisation [131]. A recent study [132] revealed that research teams working on informatics, data science, public health projects should work on approaches to organise, store and analyse complex data like ethnicity granular data. They insisted on the urgent need to use information about diverse racial and ethnic population groups while increasing availability of meaningful and usable data [132].

A clarification at this point is that our research does not focus on solving inconsistencies on varying ethnicities, but it uses it as an example to build the methodology and illustrate methods' applications with a clear impact in biomedical research.

4.3 Theoretical solution

Assumptions

As in section 3.3 we have the same assumptions. In the beginning, we assume that we have more than one study datasets. These study datasets are assumed to be non-overlapping in terms of individuals included. Therefore, there is no need for record linkage. We hypothesise that the observations within a single dataset are i.i.d. We have systematic missingness

because data are missing in specific categories only. What this means is that we have to assume that all the observations are i.i.d. We also assume that they do not contain missing values to focus on the granularity problem to keep things feasible. Each study dataset is a flat table, and naming heterogeneity has already been resolved.

A probabilistic approach to harmonisation inference across data studies

Similarly to Chapter 3, we would like to use all the available information across the datasets being integrated. Our integration method integration method, converts the problem of content heterogeneity – varying granularity problem – to a systematically missing value problem (in specific categories). Therefore, gold standard methods i.e multiple imputation that solve data missingness could be used to explore and evaluate this concept.

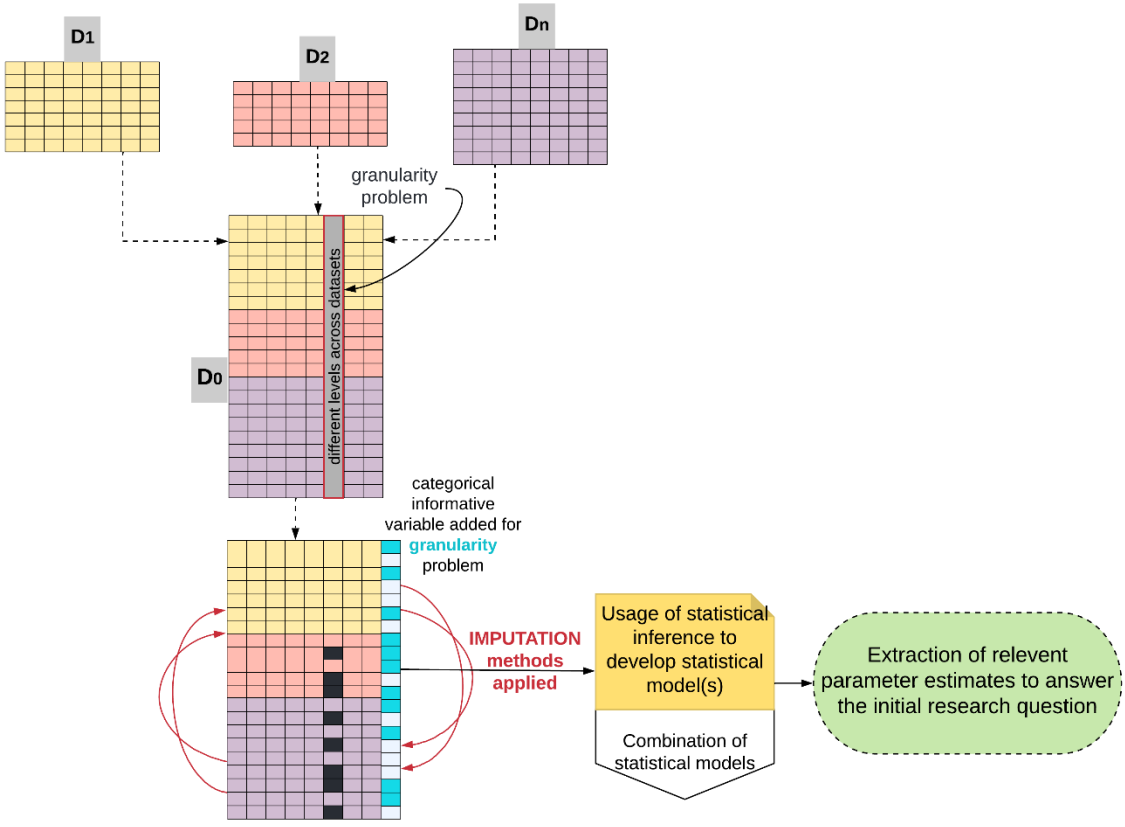


Figure 4.1. Main tasks of our probabilistic data integration process to solve granularity problem.

As shown in figure 4.1, we propose an approach that comprises some tasks to solve the granularity problem. After deciding on the research question, we integrate data from D_1, D_2 and D_n (figure 4.1 – yellow, orange, purple datasets shown respectively in the first step) that will help us answer the research question, in D_0. We choose which variables from D_0 will be included in the model. Let's say that we want to fit a model where the outcome Y is predicted

based on biomarkers X_1, X_2 and X_3. The content heterogeneity problem that results after data integration is that one categorical variable – for instance X_3 – (that needs to be included in the model) has varying levels across the studies and therefore a granularity problem exists (figure 4.1 – grey and red box in integrated dataset D_0). In this case, a traditional solution would be to keep the lowest level that is common.

However, we insist on taking advantage of all the available information. In D_0, we add a 'group' categorical informative variable that groups individuals based on X_3's levels (figure 4.1 – light blue column with different shades that represent the different X_3's levels). The tasks for handling the second content heterogeneity type are similar to the theoretical solution that solves systematically missing variables presented in Chapter 3.3. Therefore, we keep the highest number of levels of that categorical variable. As a result, systematically missing data (figure 4.1 – black squares) are introduced for some individuals in specific levels. We approach the granularity problem as a missing values problem; we solve it by applying *imputation* (figure 4.1 – red arrows) – a method well established to solve data missingness – to the integrated dataset D_0. We solve it using FCS with two different ways. More details in imputation steps, can be found in Figure 3.1.

In the first way, we apply FCS imputation to answer the research question and solve systematically missing values that arose from the granularity problem. In FCS model X'_3 will be imputed based on X_1, X_2, and Y. In the second way, we suggest an additional step to FCS's imputation model in order to eliminate misclassification in imputed values. The 'group' variable, that we introduced before, will be included in the imputation model FCS, so we have FCSgroup imputation model. In FCSgroup model X'_3 will be imputed based on X_1, X_2, X_3group and Y.

MI is implemented in most statistical software under the MAR assumption and provides unbiased and valid estimates of associations based on information from the available data. The method affects not only the coefficient estimates for variables with missing data but also the estimates for other variables with no missing data [133]. In case of MAR data, our suggested framework would slightly change so the data missingness could be explained by variables on which we have full information. Our suggested probabilistic solutions are based on FCS imputation method which can handle MAR data.

4.4 Simulation studies

We agree that the most sufficient and reasonable approach to test our idea on solving granularity issues on nominal variables after data integration, is to use realistically simulated data. We take this approach for evaluation as in synthetic data the true associations between

the predictors and the outcome are known and can be used reliable to quantify our integration methods' performance. Hence, in this section, we present the results of simulation studies designed to evaluate our approach to handle the second content heterogeneity problem.

4.4.1 Simulation study design

We first defined synthetic data that contained enough variables to illustrate the content heterogeneity problem of granularity. To simplify things and to be able to recognise easier any errors, we specified the bare minimum required to capture this problem. We performed a series of simulation studies designed to investigate our probabilistic methods in a simpler and generalised setting. The process described in figures 4.2 and 4.3 is repeated 1000 times, to obtain different datasets under the specified parameters, that were then used for analysis.

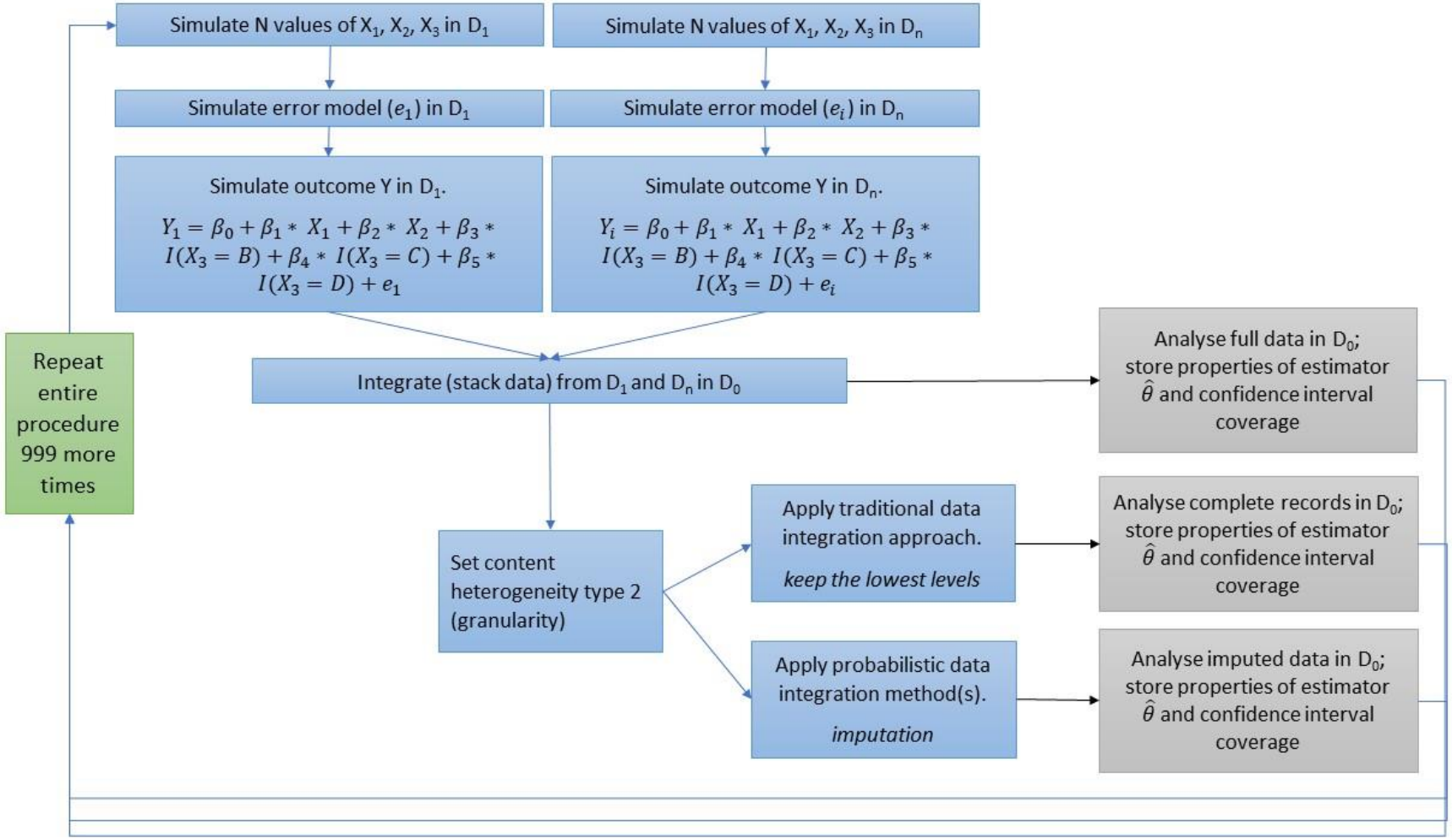


Figure 4.2. Granularity's simulation procedure: A pictorial representation of the simulation procedure for granularity.

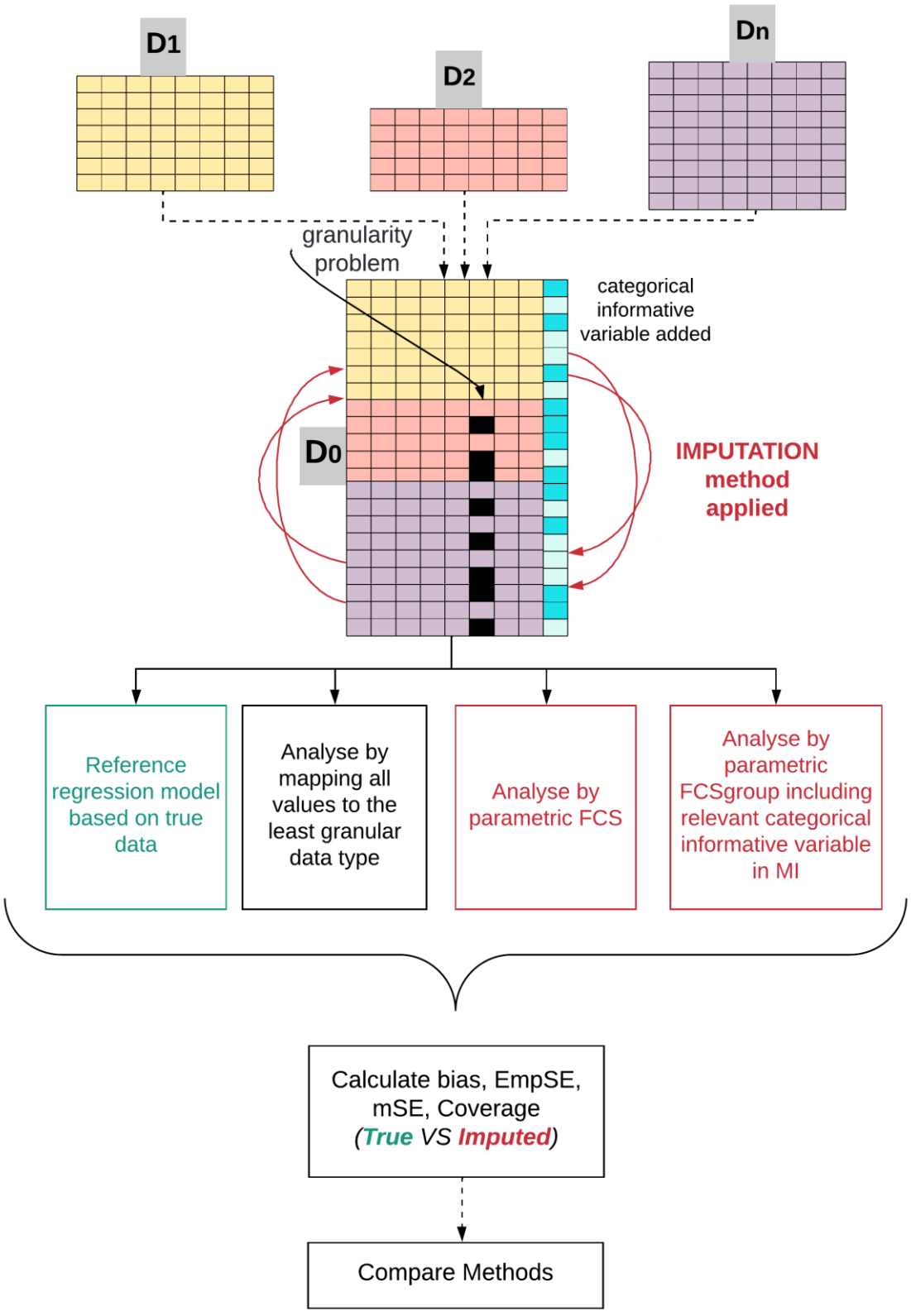


Figure 4.3. Simulation's procedure to show how the data are integrated, how granularity problem is solved through different methods and their comparison.

FCS: Fully Conditional Specification; *FCSgroup:* FCS including informative 'group' variable.

The simulation procedure (figures 4.2 and 4.3) was similar with the evaluation shown in Chapter 3.4. We simulated X_1, X_2, X_3 and we used the same data generating mechanism as in the previous chapter to simulate outcome Y (equation 4.1) with $\beta_0 = 1.4938$, $\beta_1 = -1.7483$, $\beta_2 = 0.4254$, $\beta_3 = -0.7876$, $\beta_4 = -0.9554$, $\beta_5 = 0.4844$.

$$Y_i = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * I(X_3 = B) + \beta_4 * I(X_3 = C) + \\ + \beta_5 * I(X_3 = D) + e_i \tag{4.1}$$

Afterwards, we integrated the study datasets in D_0 (figure 4.3 – yellow, orange, purple datasets), we fit the regression model to true Full Data (green box)and stored estimates $\hat{\theta}$ and standard errors $\widehat{se}(\hat{\theta})$.

We applied the second content heterogeneity problem. For example, in D_0, we had three levels ('AB', 'C', 'D') in X_3 for individuals from D_1 and four levels ('A,' 'B', 'C', 'D') in X_3 for individuals from D_2.

Traditional solution

We solved the content heterogeneity problem of granularity by applying the traditional approach where we kept the lowest level that is common (which is higher level than the most specific level available) in the problematic variable – Complete Records analysis (figure 4.3 – black box). Therefore, X'_3 had three levels ('AB', 'C', 'D') in D_0. Then, we fit a linear regression model: $Y \sim X_1 + X_2 + X'_3$ in D_0, analysed Complete Records and stored estimates $\hat{\theta}$ and standard errors $\widehat{se}(\hat{\theta})$.

Probabilistic solutions

After, we introduced the categorical informative variable X_3group (figure 4.3 – light blue column with different blue shades). In D_0, we created 'X_3group' that groups individuals based on X_3's levels (figure 4.3 – different blue shades in the light blue column). If $I(X_3 = A)$ or $I(X_3 = B)$ or $I(X_3 = AB)$ then 'X_3group' = '0', otherwise 'X_3group' = '1'. If $I(X_3 = AB)$ (i.e., individuals from D_1) set X_3 to missing. So, X_3 had systematically missing data for individuals from D_1 in values 'AB'. In figure 4.3 'X_3group' is the light blue column and its different shades represents the different X_3's levels.

We wanted to check if by adding X_3group and therefore extra information in the imputation model, we achieved better estimates and less misclassification. So, we solved the second content heterogeneity problem by applying the probabilistic approaches FCS and FCSgroup as presented in figures 4.1 - 4.3 in (red arrows for imputation and red boxes). We fit linear

regression models for FCS and FCSgroup: Y ~ X_1 + X_2 + X''$_3$ in D_0, analysed imputed datasets using Rubin's rules and stored estimates $\hat{\theta}$, standard errors $\widehat{se}(\hat{\theta})$ and confidence interval.

All imputation analyses were performed with R package *mice* freely available on CRAN [98]. We used the seed function () and set starting seed to 975392. In the next section, we present the simulations' results.

4.4.2 Performance measures and scenarios

To examine the simulations' results we chose the same performance measures as in chapter 3, *bias, EmpSE, mSE and Coverage*. As mentioned in chapter 3.4.3, all performance measures were calculated between the estimates coefficients of each (probabilistic/traditional) integration model and the generating coefficients and then averaged across iterations. In this chapter, estimand θ was the estimate of $\hat{\theta}_i$ of X_3 coefficients (and mainly $X_{3=B}$) in each model fit.

Table 4.1 shows the different scenarios to generate data following figures 4.1- 4.3 and their aforementioned tasks. We explore different simulation scenarios by varying the number of individuals (N) per study, number of studies (D_n), imputations (m), model errors (e_i). The number of m and iterations (it) is set to five (default).

Table 4.1. Scenarios 1 - 5 used to generate data from Figure 4.2.

	Scenario 1	**Scenario 2**	**Scenario 3**	**Scenario 4**	**Scenario 5**
Number of individuals per study (N)	200	1000	200	1000	D_1: 200, D_2:150, D_3:50, D_4:75, D_5:100
Model error e_i: (same for each study)	N~(0,0.2) N~(0,2) N~(0,20)	N~(0,0.2) N~(0,2) N~(0,20)	N~(0,0.2) N~(0,2) N~(0,20)	N~(0,0.2) N~(0,2) N~(0,20)	N~(0,0.2) N~(0,2) N~(0,20)
Number of studies (D)	2	2	5	5	5
Imputations (m)	5	5	5	5	5
Missingness applied to:	X_3 from D_1	X_3 from D_1	X_3 from D_4, D_5	X_3 from D_2, D_5	X_3 from D_4, D_5

Table 4.2. Scenarios 6 - 10 used to generate data from Figure 4.2.

	Scenario 6	**Scenario 7**	**Scenario 8**	**Scenario 9**	**Scenario 10**
Number of individuals per study (N)	D_1:800, D_2:150, D_3:50, D_4:75, D_5:350, D_6: 00, D_7:150, D_8:500, D_9:750, D_{10}:100	100	200	500	1000
Model error e_i: (same for each study)	N~(0,0.2) N~(0,2) N~(0,20)	e_1: N~(0,1.2), e_2: N~(0,1.3)	e_1: N~(0,1.2), e_2: N~(0,1.3)	e_1: N~(0,1.2), e_2: N~(0,1.3)	e_1: N~(0,1.2), e_2: N~(0,1.3)
Number of studies (D)	10	2	2	2	2
Imputations (m)	5	5	5	5	5
Missingness applied to:	X_3 from D_3, D_6, D_9, D_{10}	X_3 from D_1	X_3 from D_1	X_3 from D_1	X_3 from D_1

4.4.3 Results

In this section, we present the results of a series of simulation studies. The aim is to compare the Complete Records analysis with imputation in the concept of granularity and mainly evaluate our probabilistic data integration approaches FCS and FCSgroup against true Full Data - before we apply any content heterogeneity. All results in detail are presented in Appendix B.

Simulations with studies of same sizes (Scenarios 1 - 4)

Scenario 1

Here, we simulated data for **two studies**, each with **200 patients** and m = 5, it = 5. For each simulated dataset, we applied the granularity problem to X_3 from D_1. We present the simulation results in Figure 4.4 we see a graphical representation of the main results with e_i: (N ~ (0, 0.2)), (N ~ (0, 2)), (N ~ (0, 20)) respectively. We see that both FCS and FCSgroup

gave unbiased results and a very good coverage of the confidence interval for all the three model errors. It seems that when e_i: N ~ (0, 2), FCSgroup outperformed FCS. When the error varied extremely (e_i: N ~ (0, 20)), it led to overestimation of model and empirical standard errors. Complete Records seemed to be an incompatible technique and led to large biases.

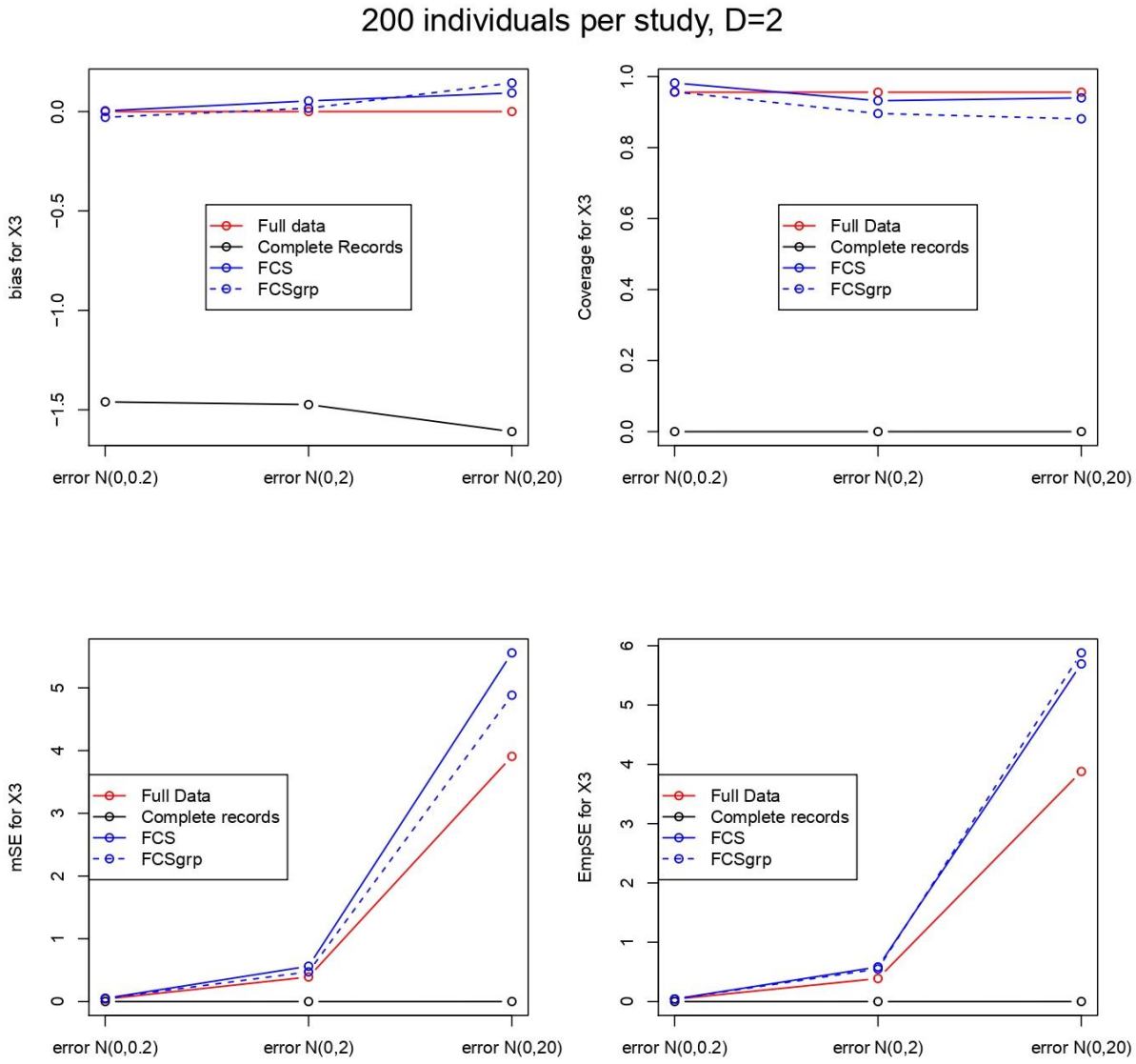


Figure 4.4. Main results from scenario 1's simulation study: Comparison of Bias, Coverage level, mSE and EmpSE for $X_{3=B}$ after 1000 simulations with Full Data (red line), handling granularity with Complete Records (black line), FCS (blue line), FCSgroup (broken blue line) for three model errors.

FCS: Multiple imputation by fully conditional specification; *FCSgroup:* FCS imputation model included X'_3; *D:* number of studies, *mSE:* mean model standard error; *EmpSE:* mean empirical standard error.

Scenario 2

This scenario was like scenario 1 with a difference of **1000 individuals per study**. We present the simulation results in figure 4.5 with e_i: (N ~ (0, 0.2)), (N ~ (0, 2)), (N ~ (0, 20)) respectively. When we increased patients from 200 to 1000 per study, imputation methods performed slightly better. Both FCS and FCSgroup gave unbiased results and good coverage. However, FCSgroup outperformed FCS. The Complete Records approach appeared to be not a good solution to solve the granularity issue and should be avoided.

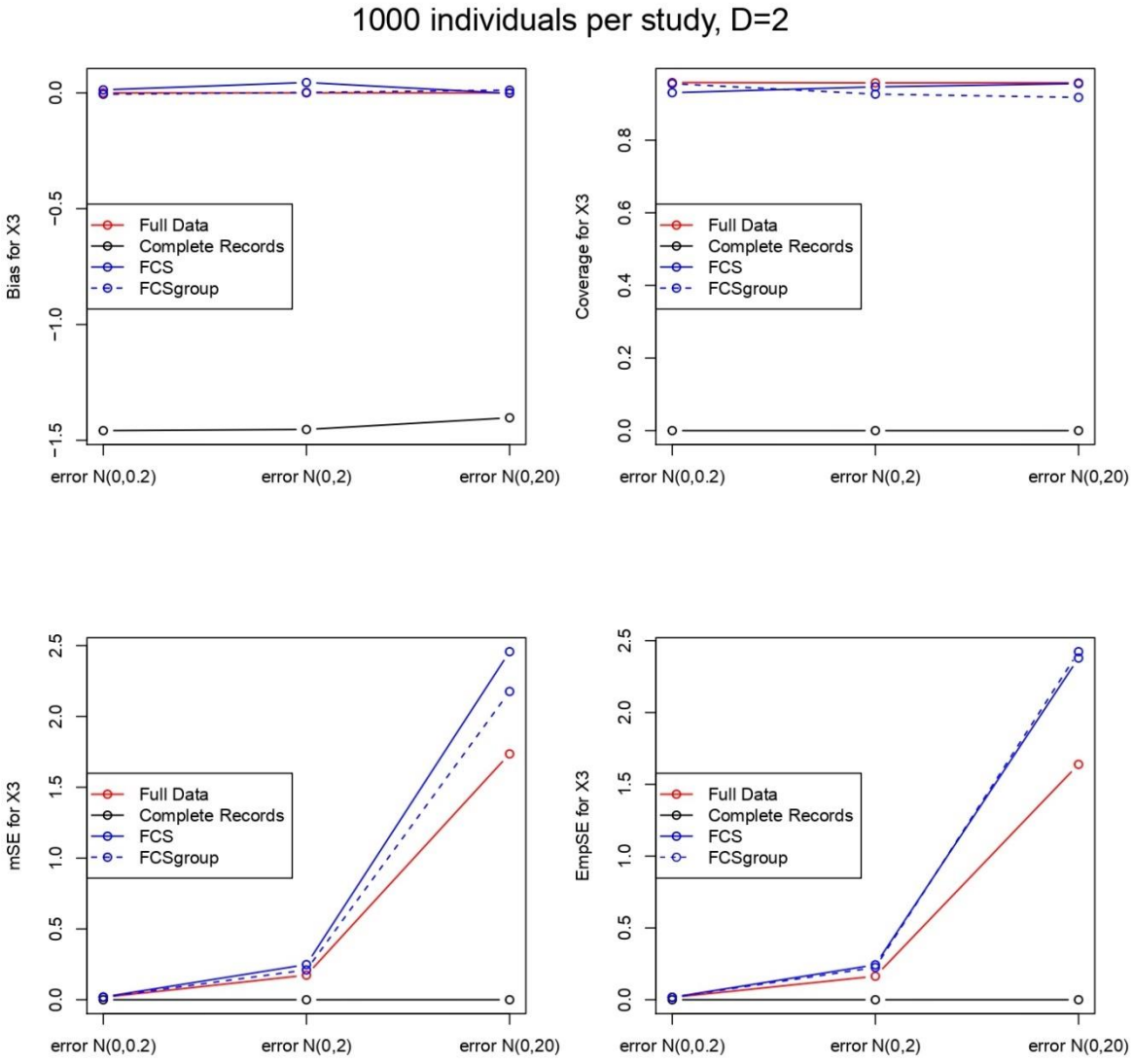


Figure 4.5. Main results from scenario 2's simulation study: Comparison of Bias, Coverage level, mSE and EmpSE for $X_{3=B}$ after 1000 simulations with Full Data (red line), handling granularity with Complete Records (black line), FCS (blue line), FCSgroup (broken blue line) for three model errors.

FCS: Multiple imputation by fully conditional specification; *FCSgroup:* FCS imputation model included X'$_3$; *D:* number of studies, *mSE:* mean model standard error; *EmpSE:* mean empirical standard error.

Simulations with studies of different sizes (Scenarios 3 - 4)

Scenarios 3 - 4

In scenarios 3-4, we simulated data for **five studies**, each with the same number of individuals per study. In Scenario 3 each study had **200 patients** (so D_0 had 1000 patients) and in scenario 4 each study had **1000 individuals** (so D_0 had 5000 patients). For each simulated dataset, we chose two random studies (D_4 and D_5 for scenario 3 and D_2 and D_5 for scenario 4) to apply the granularity issue in X_3 in D_0. We present the simulation results in figures 4.6 and 4.7. In scenario 3, analysis of datasets imputed with FCS and FCSgroup model gave good results both in terms of bias, precision, and confidence interval coverage when the model error was small and medium (figure 4.6). We see an indication of low bias in both FCS and FCSgroup when the model error was large (simulated as e_i:N~(0,20)). In scenario 4, as shown in figure 4.6, FCSgroup was very close to the true Full Data. FCS had a coverage of 84.9% when the error was small (simulated as e_i:N~(0,0.2)). In general, in both scenarios, when applying the Complete Records approach, the results were highly biased, resulting in large standard errors and under-coverage. The results show that FCSgroup performed slightly better than FCS when we had 200 individuals per study, and when we had 1000 individuals per study and small model error. In general, in both scenarios FCSgroup outperformed FCS.

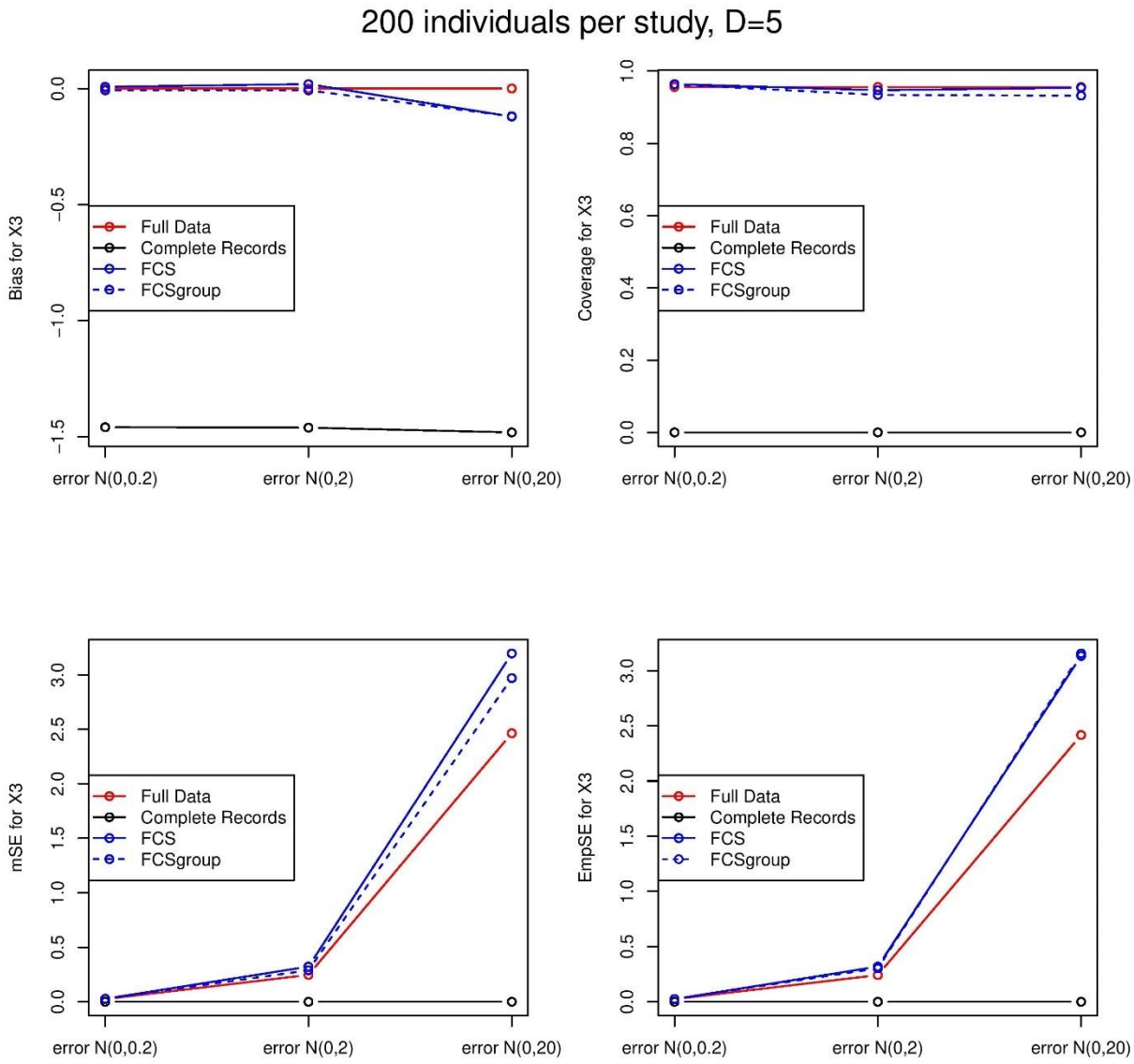


Figure 4.6. Main results from scenario 3's simulation study: Comparison of Bias, Coverage level, mSE and EmpSE for $X_{3=B}$ after 1000 simulations with Full Data (red line), handling granularity with Complete Records (black line), FCS (blue line), FCSgroup (broken blue line) for three model errors.

FCS: Multiple imputation by fully conditional specification; *FCSgroup:* FCS imputation model included X'$_3$; *D:* number of studies, *mSE:* mean model standard error; *EmpSE:* mean empirical standard error.

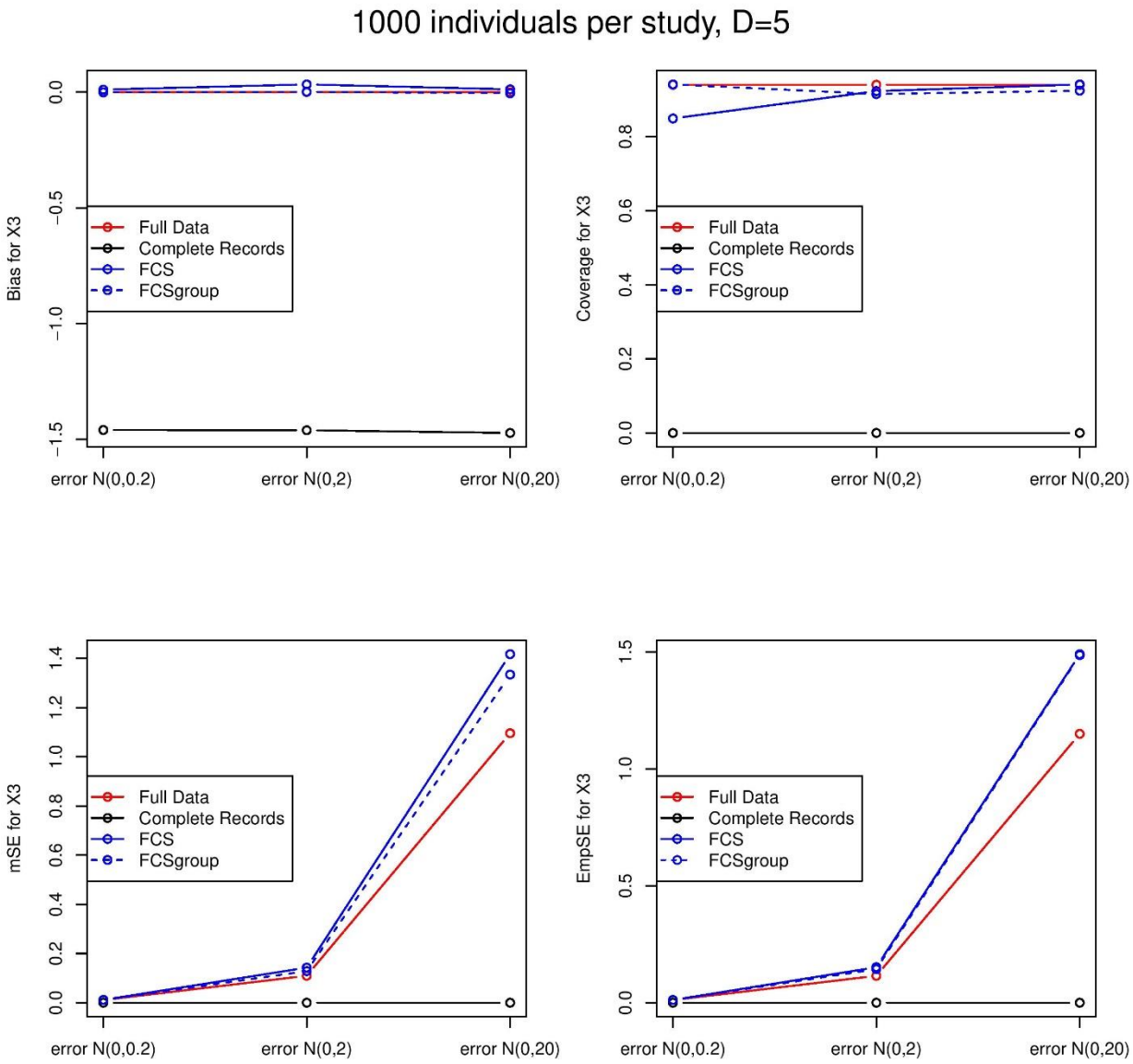


Figure 4.7. Main results from scenario 4's simulation study: Comparison of Bias, Coverage level, mSE and EmpSE for $X_{3=B}$ after 1000 simulations with Full Data (red line), handling granularity with Complete Records (black line), FCS (blue line), FCSgroup (broken blue line) for three model errors.

FCS: Multiple imputation by fully conditional specification; *FCSgroup:* FCS imputation model included X'$_3$; *D:* number of studies, *mSE:* mean model standard error; *EmpSE:* mean empirical standard error.

Overall, in all scenarios both methods produced unbiased or very little biased results (Figure 4.8). FCSgroup produced better mean estimate than FCS *except* the following cases: e_i:N~(0,20)), D2_N200 and D2_N1000. With regards to coverage (figure 4.9), both imputation models produced similar coverage to true Full Data in almost all cases. We observe under-coverage (84.9%) in FCS in scenario 4 when model error was small. In figures 4.10 and 4.11 we see mSE and EmpSE for scenarios 1 to 4. We see that as the model error increased the mSE and EmpSE increased as well. mSE and EmpSE were similar per scenario and they decreased as the integrated dataset's size increased.

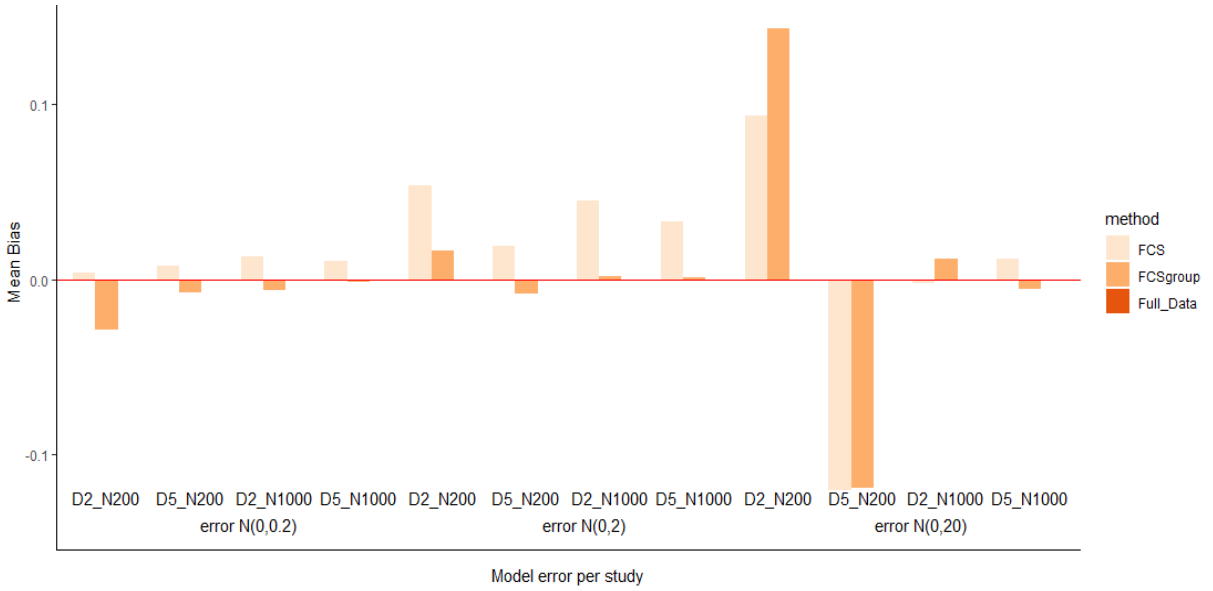


Figure 4.8. Mean bias for $X_{3=B}$ for '2 datasets, N=200 per dataset' (D2_N200), '5 datasets, N=200 per dataset' (D5_N200), '2 datasets, N=1000 per dataset' (D2_N1000), and (D5_N1000) '5 datasets, N=1000 per dataset' for three model errors.

FCS: Multiple imputation by fully conditional specification; *FCSgroup:* FCS imputation model included X'$_3$.

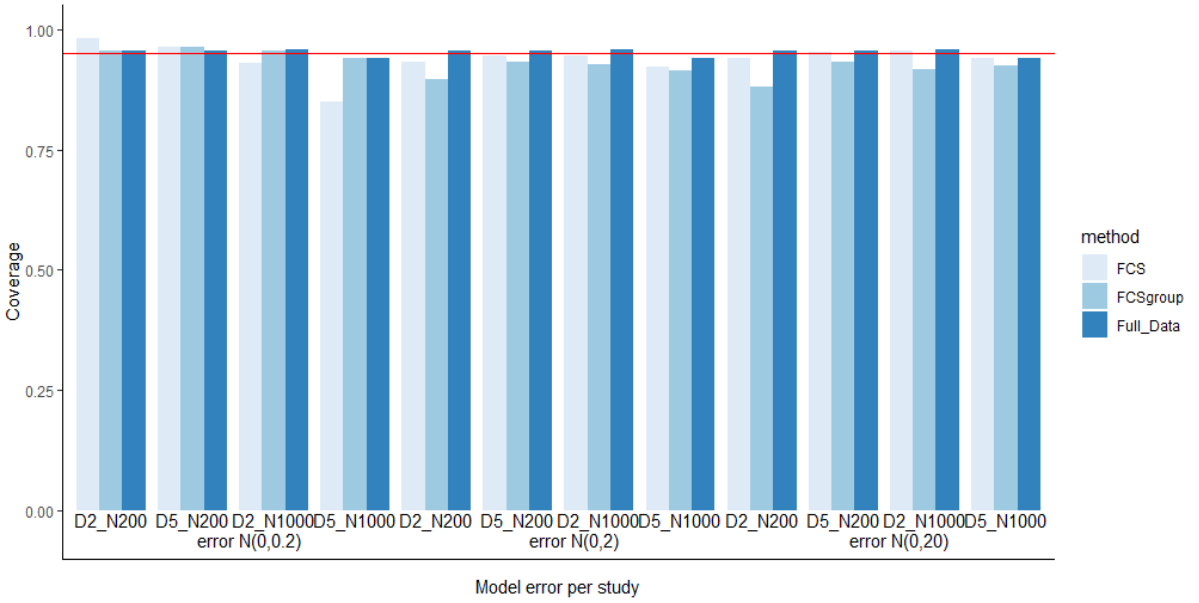


Figure 4.9. Coverage for $X_{3=B}$ for '2 datasets, N=200 per dataset' (D2_N200), '5 datasets, N=200 per dataset' (D5_N200), '2 datasets, N=1000 per dataset' (D2_N1000), and (D5_N1000) '5 datasets, N=1000 per dataset' for three model errors.

FCS: Multiple imputation by fully conditional specification; *FCSgroup:* FCS imputation model included X'$_3$.

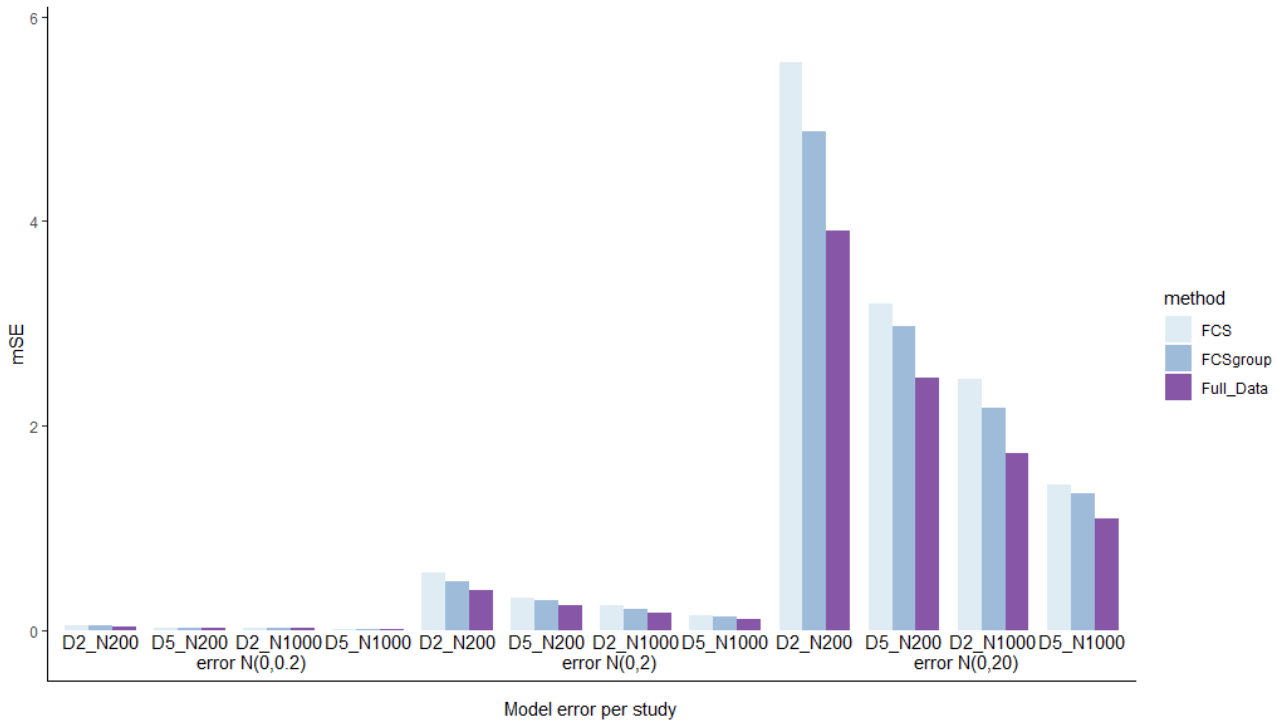


Figure 4.10. mSE for $X_{3=B}$ for '2 datasets, N=200 per dataset' (D2_N200), '5 datasets, N=200 per dataset' (D5_N200), '2 datasets, N=1000 per dataset' (D2_N1000), and (D5_N1000) '5 datasets, N=1000 per dataset' for three model errors.

FCS: Multiple imputation by fully conditional specification; *FCSgroup:* FCS imputation model included X'_3; *mSE:* mean model standard error.

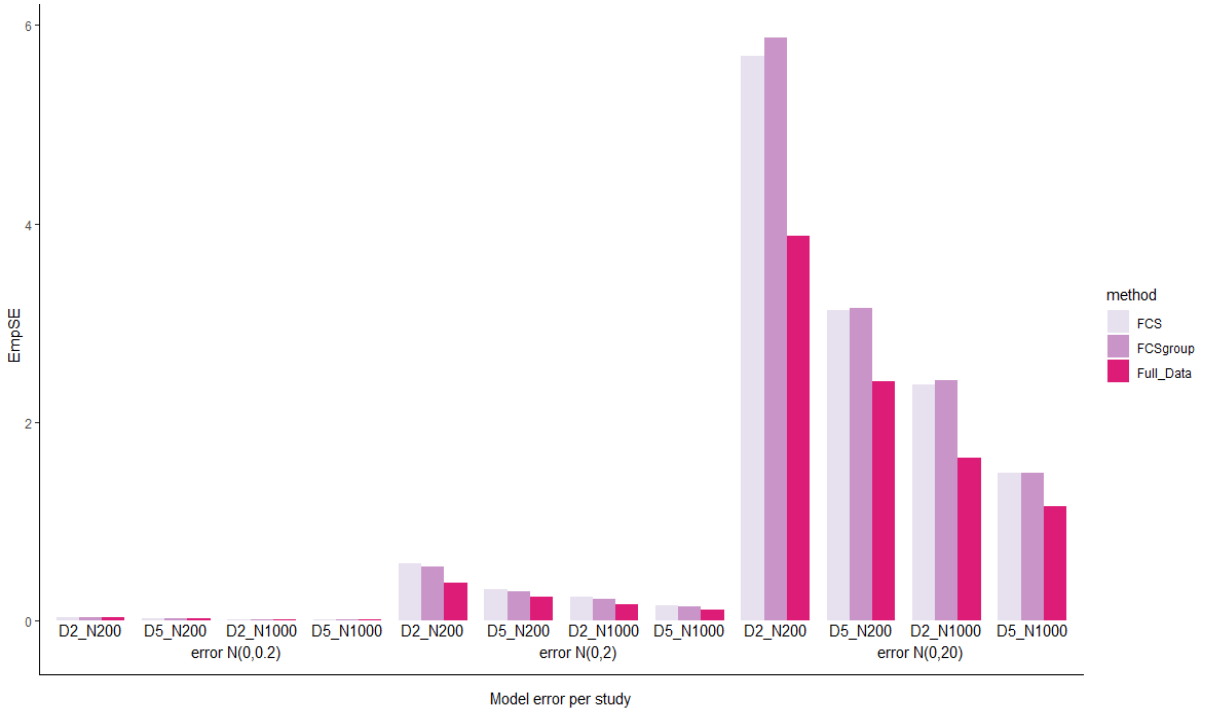


Figure 4.11. EmpSE for $X_{3=B}$ for '2 datasets, N=200 per dataset' (D2_N200), '5 datasets, N=200 per dataset' (D5_N200), '2 datasets, N=1000 per dataset' (D2_N1000), and (D5_N1000) '5 datasets, N=1000 per dataset' for three model errors.

FCS: Multiple imputation by fully conditional specification; *FCSgroup:* FCS imputation model included X'_3; *EmpSE:* mean empirical standard error.

Simulations with studies of different size and same model error per study (Scenarios 5 - 6)

Scenario 5

We simulated data for **five studies**, each with **different number of individuals** (D_1:200, D_2:150, D_3:50, D_4:75, D_5:100). We chose X_3 to have three levels in two studies (D_4 and D_5). See results in figure 4.12. FCS and FCSgroup produced almost identical results with true data. FCS had slightly better results than FCSgroup in terms of bias when the model error was high (e_i: N ~ (0, 20)). Both probabilistic models outperformed Complete Records.

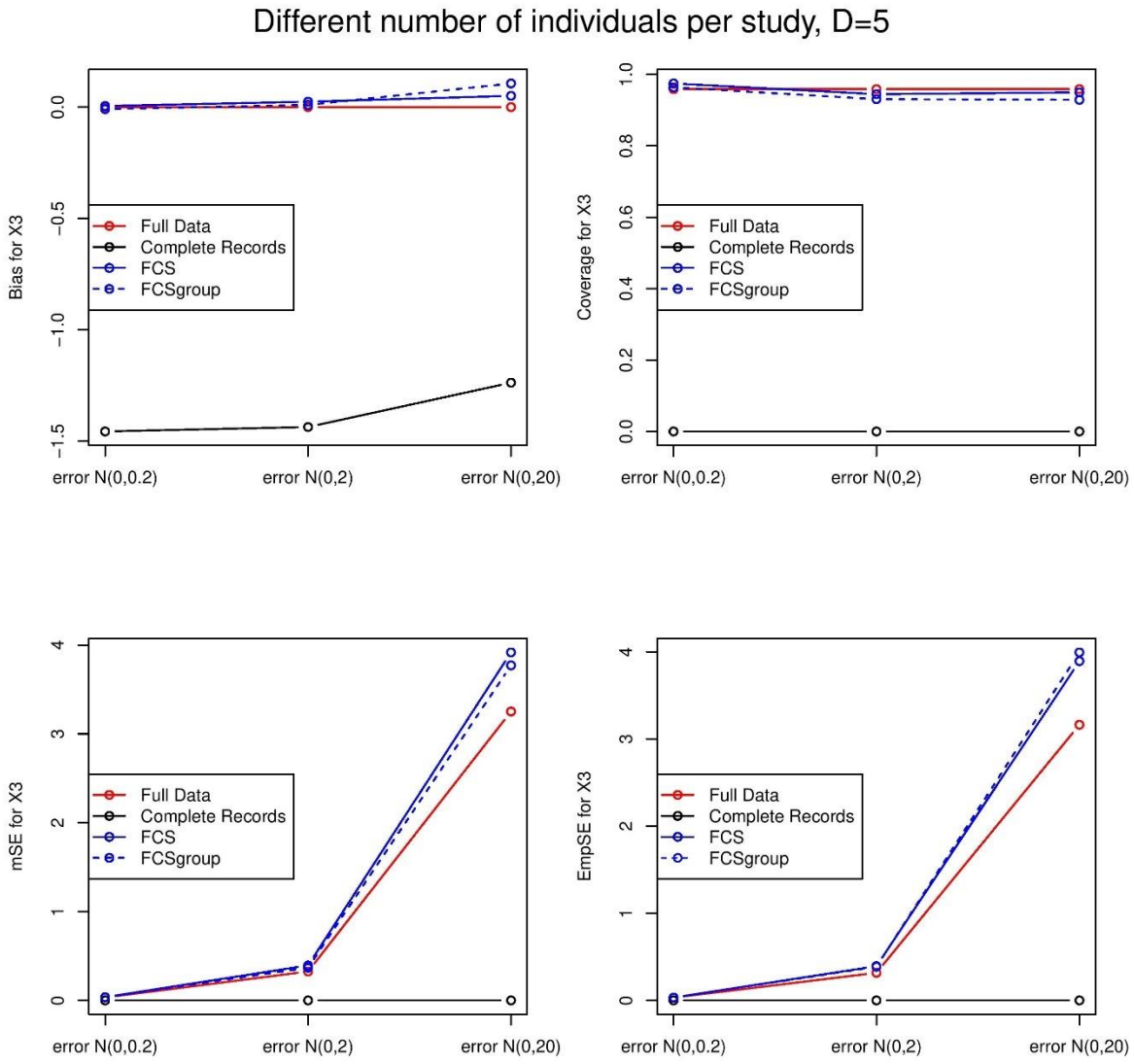


Figure 4.12. Main results from scenario 5's simulation study: Comparison of Bias, Coverage level, mSE and EmpSE for $X_{3=B}$ after 1000 simulations with Full Data (red line), handling granularity with Complete Records (black line), FCS (blue line), FCSgroup (broken blue line) for three model errors.

FCS: Multiple imputation by fully conditional specification; *FCSgroup:* FCS imputation model included X'$_3$; *D:* number of studies, *mSE:* mean model standard error; *EmpSE:* mean empirical standard error.

Scenario 6

We simulated data from **ten studies**, each with **different number of individuals** (D_1:800, D_2:150, D_3:50, D_4:75, D_5:350, D_6: 200, D_7:150, D_8:500, D_9:750, D_{10}:100). We decided X_3 to have three levels in four studies (D_3, D_6, D_9, and D_{10}). Results for this scenario are shown in figure 4.13. As in scenario 5, FCS and FCSgroup produced identical results with true data. FCSgroup had slightly better results in terms of bias when the model error is medium (e_i: N ~ (0, 2)).

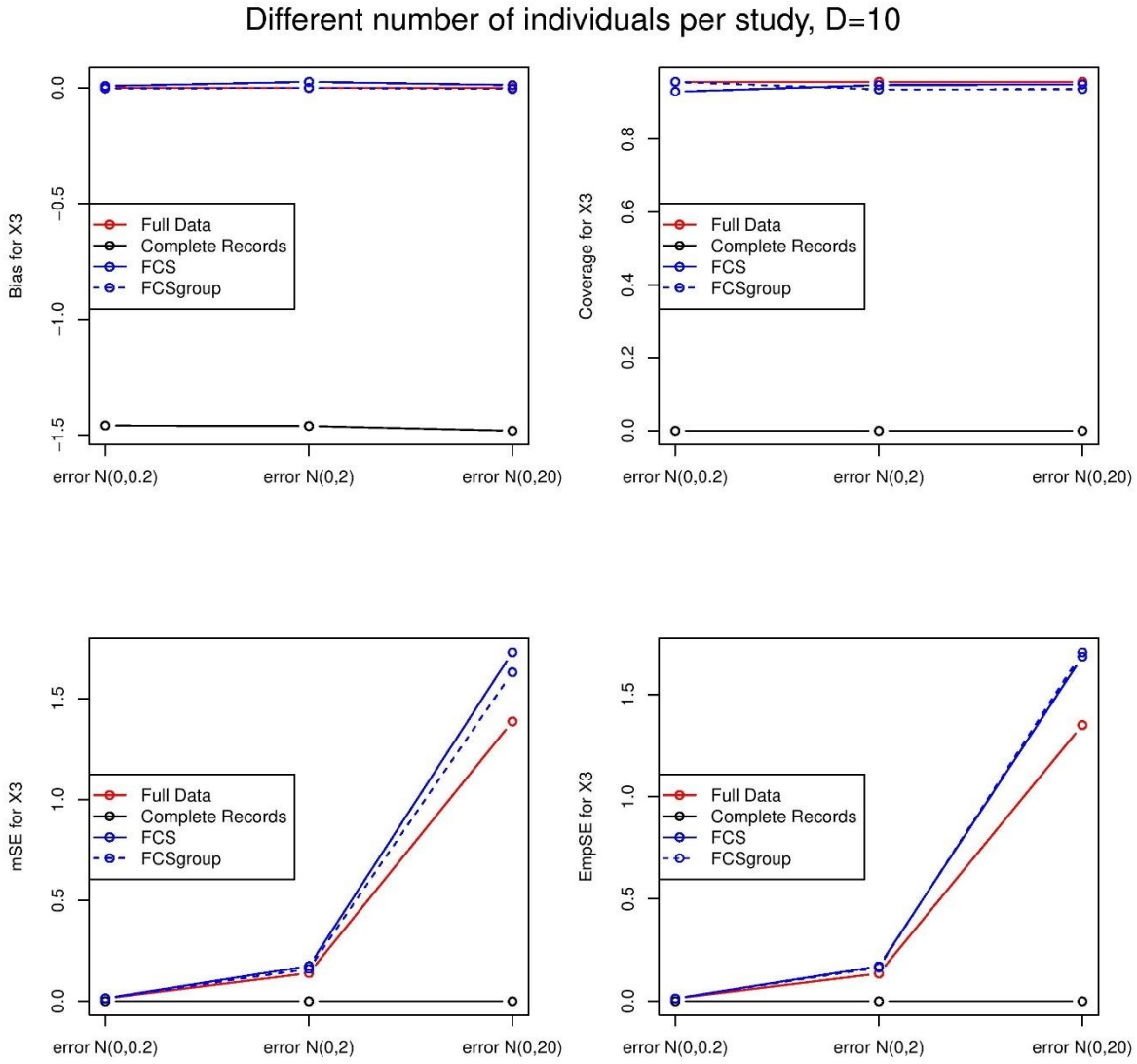


Figure 4.13. Main results from scenario 6's simulation study: Comparison of Bias, Coverage level, mSE and EmpSE for X_3 after 1000 simulations with Full Data (red line), handling granularity with Complete Records (black line), FCS (blue line), FCSgroup (broken blue line) for three model errors.

FCS: Multiple imputation by fully conditional specification; *FCSgroup:* FCS imputation model included X'$_3$; *D:* number of studies, *mSE:* mean model standard error; *EmpSE:* mean empirical standard error.

Overall, as we have seen in figures 4.14 - 4.17 in both scenarios, we had good coverage and there was either with no bias or very small bias (FCSgroup, D5_diffsize, large model error). When we had 10 datasets with different sizes both imputation models provided unbiased

results and FCSgroup had slightly better estimates than FCS. However, in the 5 different size datasets example FCS showed better mean estimate than FCSgroup when model errors were small and large. In scenarios 5 and 6, both probabilistic models outperformed Complete Records. In figures 4.16 and 4.17 we see mSE and EmpSE for scenarios 5 and 6. We see that as the model error increased the mSE and EmpSE increased as well. mSE and EmpSE are similar per scenario, and they decreased as the integrated dataset's size increased.

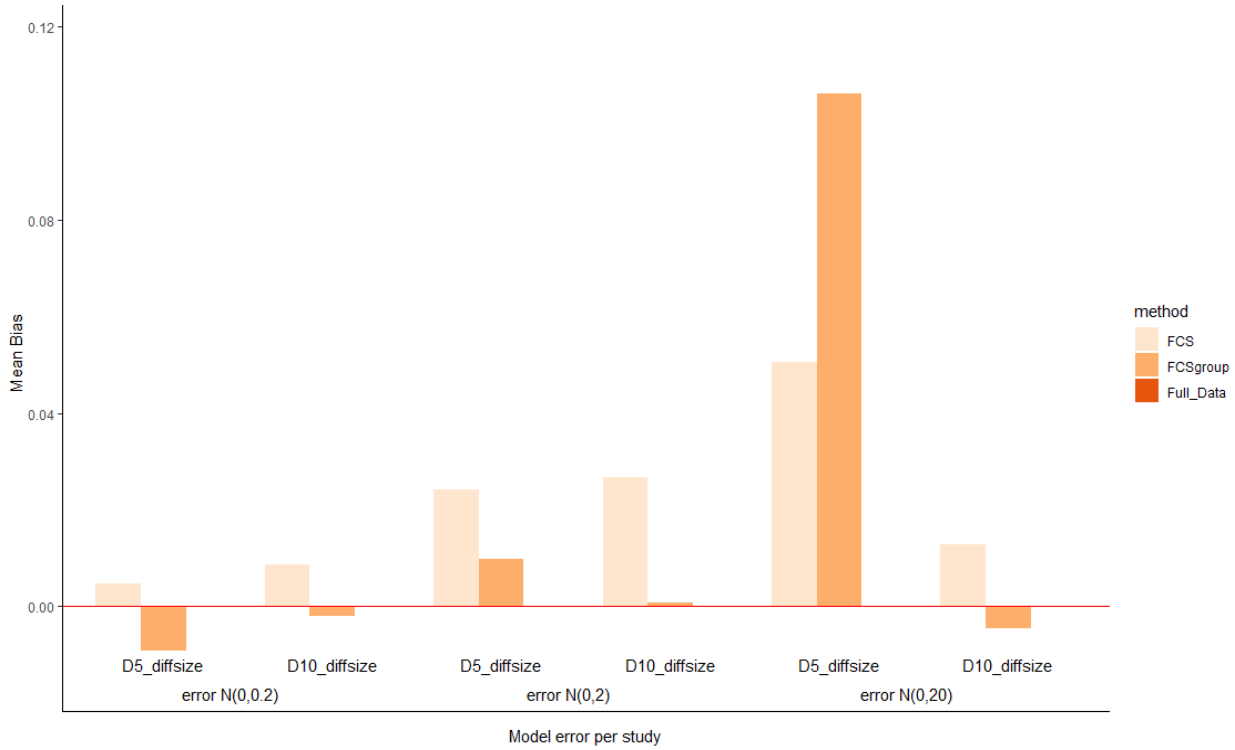


Figure 4.14. Mean bias for $X_{3=B}$ for '5 datasets, N=different per dataset' (D5_diffsize), '10 datasets, N=different per dataset' (D10_diffsize) for the three model errors.

FCS: Multiple imputation by fully conditional specification; *FCSgroup:* FCS imputation model included X'_3.

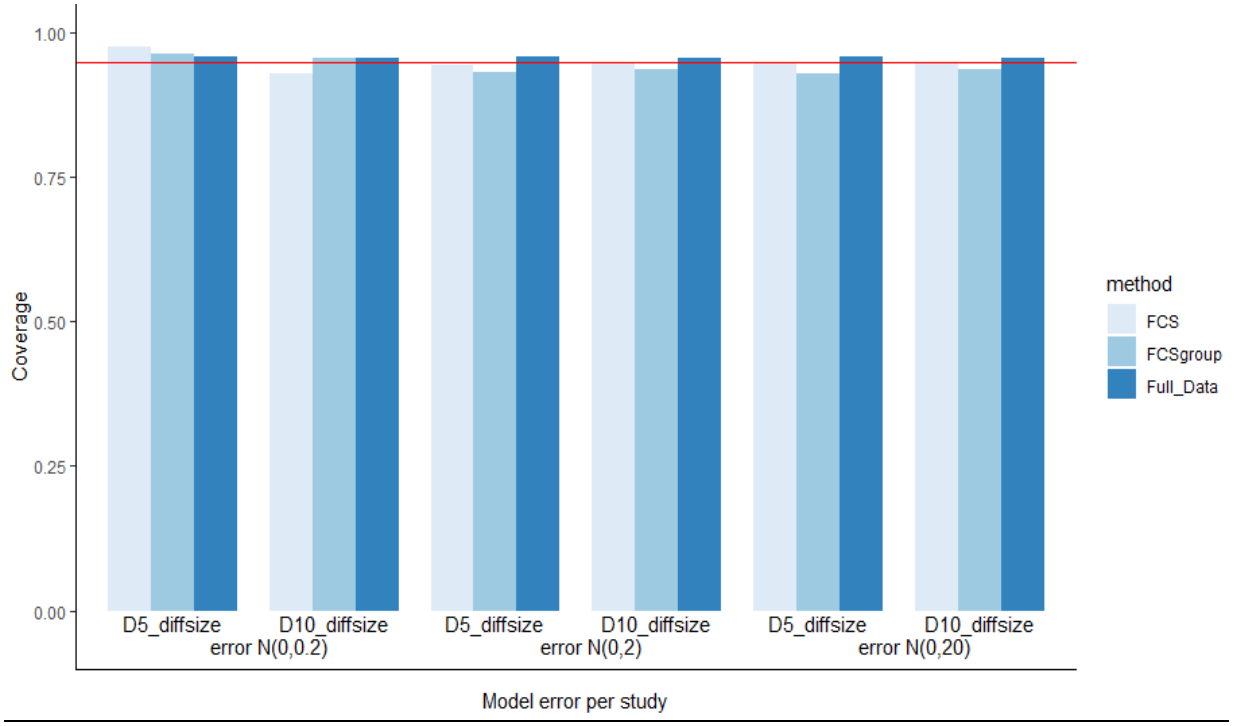


Figure 4.15. Coverage for $X_{3=B}$ for '5 datasets, N=different per dataset' (D5_diffsize), '10 datasets, N=different per dataset' (D10_diffsize) for the three model errors.

FCS: Multiple imputation by fully conditional specification; *FCSgroup:* FCS imputation model included X'_3.

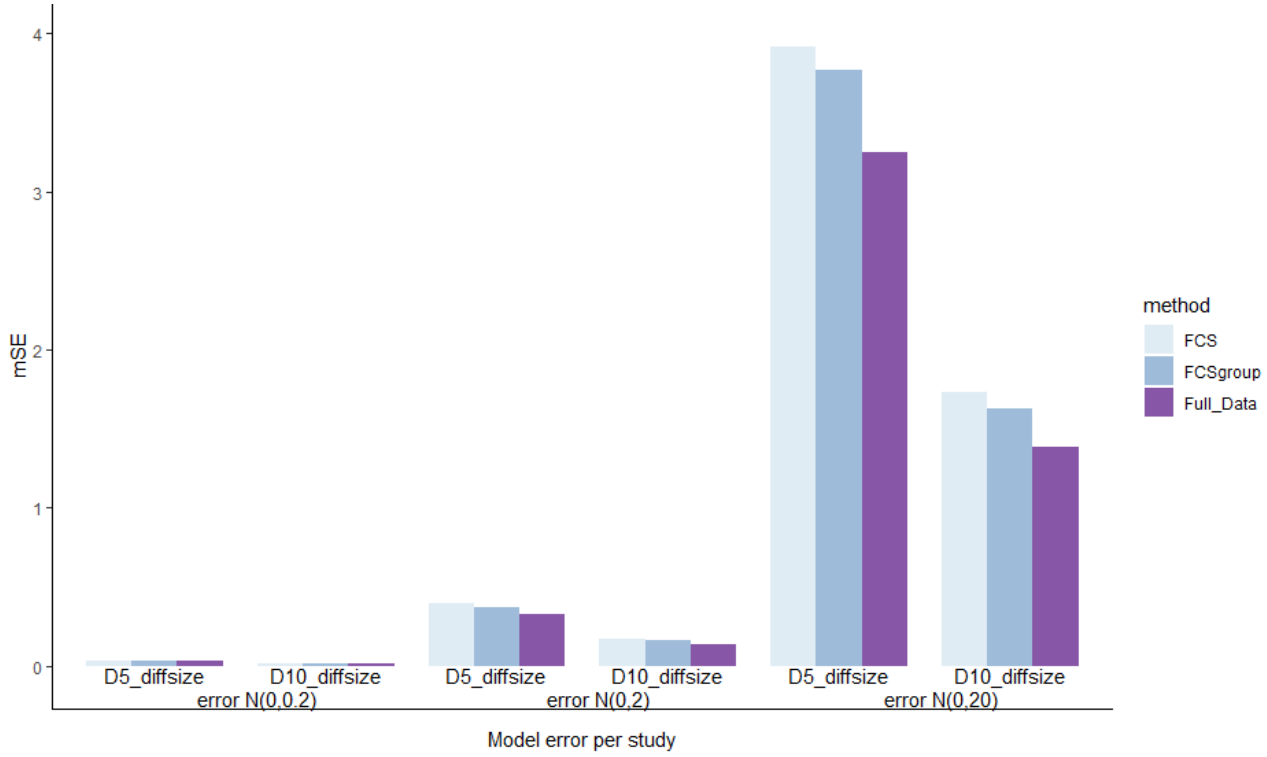


Figure 4.16. mSE for $X_{3=B}$ for '5 datasets, N=different per dataset' (D5_diffsize), '10 datasets, N=different per dataset' (D10_diffsize) for the three model errors.

FCS: Multiple imputation by fully conditional specification; *FCSgroup:* FCS imputation model included X'$_3$; *mSE:* mean model standard error.

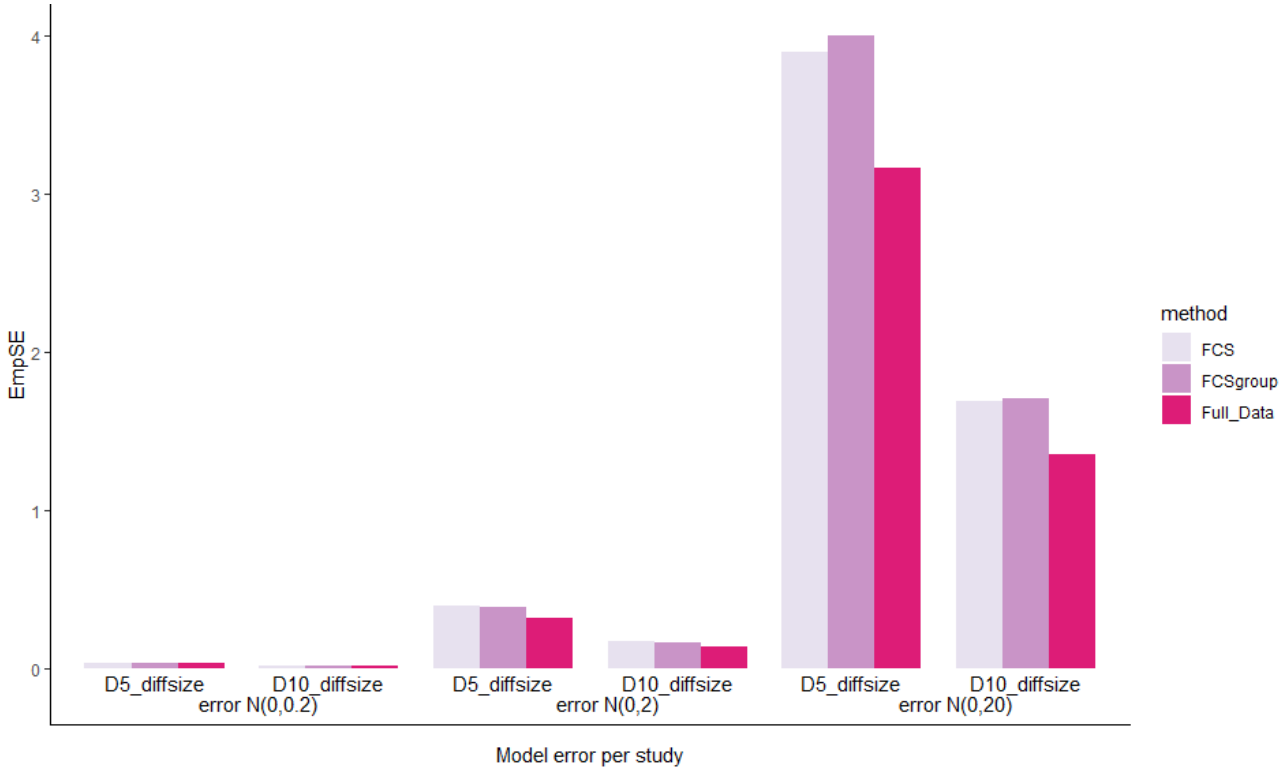


Figure 4.17. EmpSE for $X_{3=B}$ for '5 datasets, N=different per dataset' (D5_diffsize), '10 datasets, N=different per dataset' (D10_diffsize) for the three model errors.

FCS: Multiple imputation by fully conditional specification; *FCSgroup:* FCS imputation model included X'$_3$; *EmpSE:* mean empirical standard error.

Simulations with equal size studies, different model error per study (7 - 10)

Scenarios 7 – 10

Here, we again simulated data from **2 studies**, each with either 100, 200, 500 or 1000 individuals (with **size varying between scenarios**). Study 1, had model error e_i:N~(0,1.2), whereas study 2, had model error e_i:N~(0,1.3). For each simulated dataset, we applied the granularity problem to $X_{3=B}$ from D_1. We present the simulation results in figure 4.18. The results indicate that even when the FCS produced slightly biased results (Scenarios 9-10) the results were closer to the truth than Complete Records. In all cases we see an overestimation of the standard error in FCS and in FCSgroup. However, the standard error reduced as the number of individuals and therefore the number of individuals in total) per study increased. FCSgroup shows lower and closer mSE and EmpSE to the true model's. FCSgroup shows virtually some bias and small under-coverage of the confidence interval. FCS shows no bias and very good coverage levels, almost identical with true model's, in all

the examined size studies. Both FCS and FCSgroup outperform Complete Records which is not suggested as a solution on this case either.

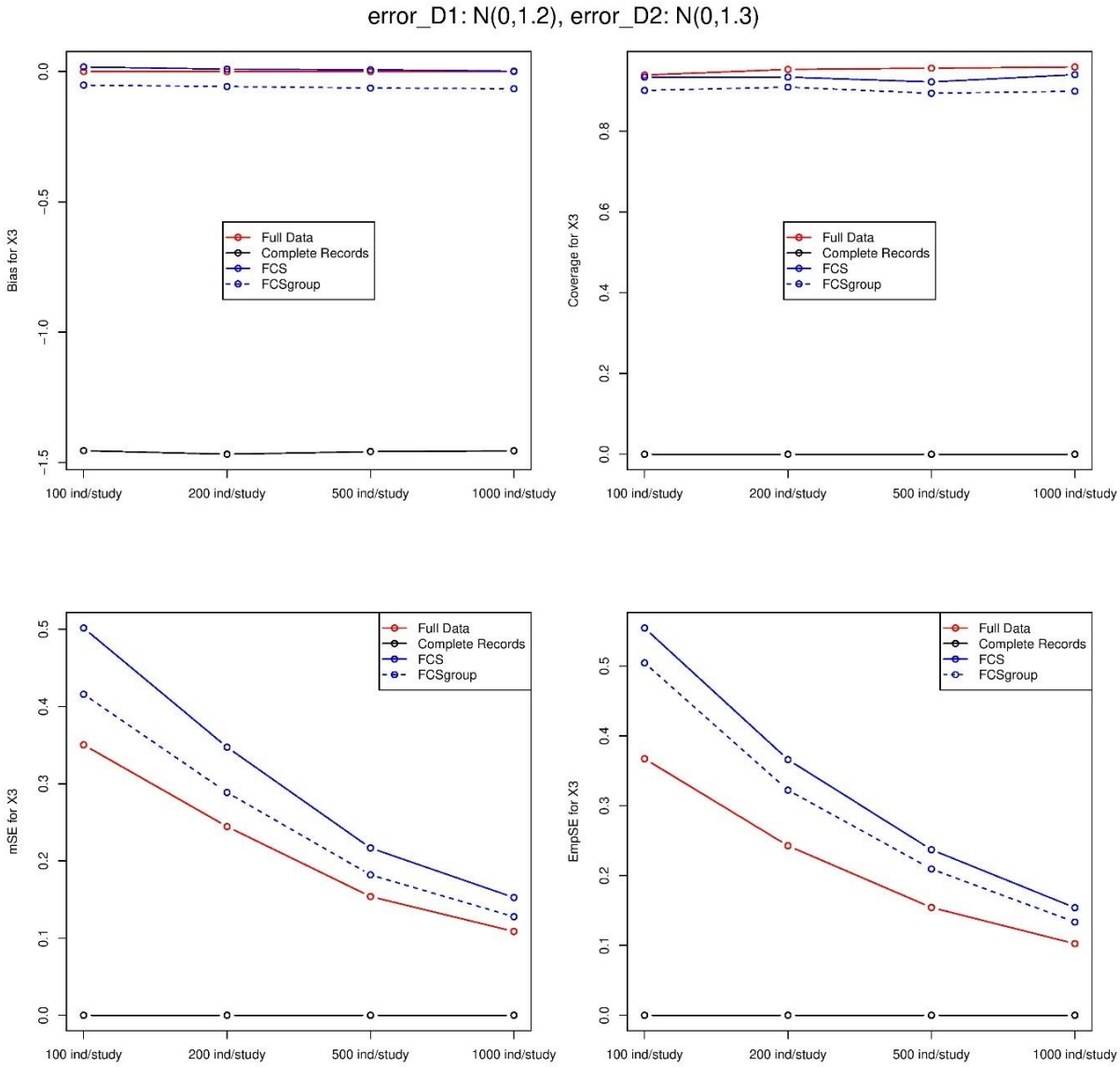


Figure 4.18. Main results from scenarios 7-10's simulation study: Comparison of Bias, Coverage level, mSE and EmpSE for X_3 after 1000 simulations with Full Data (red line), handling granularity with Complete Records (black line), FCS (blue line), FCSgroup (broken blue line) for three model errors.

FCS: Multiple imputation by fully conditional specification; *FCSgroup:* FCS imputation model included X'$_3$; *D:* number of studies, *mSE:* mean model standard error; *EmpSE:* mean empirical standard error.

4.4.4 Summary of findings from simulation studies

Drawing together the results from the simulation studies, we conclude the following:

Studies with the same model errors

In most cases, FCS and FCSgroup performed really well and gave unbiased and close to reality results, therefore they solved granularity problem successfully. In all scenarios as the model error increased, mSE and EmpSE increased as well. We can also say that FCSgroup produced slightly more similar estimates to true Full Data.

Studies with different model errors

In these simulation studies, we may conclude that if the number of imputations and iterations was larger, it achieved better convergence. Smaller datasets provided better estimates probably because of less variability. When we had 500 or 1000 individuals per study, we see smaller standard errors and very mild under-coverage.

Studies of same sizes

For both probabilistic methods, bias remained negligible, and coverage was very good and there was a lot more gain in information compared with a Complete Records analysis. FCSgroup was slightly better than FCS in most cases (with the exclusion of e_i:N~(0,20)), D2_N200 and D2_N1000).

Studies of different sizes

In both scenarios, there was good coverage and there was mostly no bias *except* in FCSgroup, D5_diffsize, e_i:N~(0,20). When we had 10 datasets with different sizes both imputation models provided unbiased results and FCSgroup had slightly better estimates than FCS. However, in the 5 different size datasets example FCS shows better mean estimate than FCSgroup when model errors were small and large.

Size of the model error

The smaller the model error per study, the smaller the EmpSE and mSE observed. We also see an indication that as the model error increased, the difference between EmpSE and mSE slightly increased.

Overall

Based on the results from the simulations, a probabilistic approach is a precise way and useful addition to structured healthcare data integration toolkit. When the data integration is more complex, a higher number of imputations may be needed as it may take more iterations to reach the appropriate convergence [134]. Results agree with the ones presented in chapter 3 as FCS gave unbiased estimates with good coverage confidence intervals. The difference with chapter 3 is that here we applied FCSgroup which seems to help categorical variables to achieve better estimate classification. The probabilistic approaches gave precise results

across a range of scenarios and in all cases outperformed the complete case analysis and gave estimates close to real data. They did not introduce practically important bias in any of the scenarios considered, although it was a little conservative in larger datasets with different model errors. We suggest that the worst that can happen is that estimates are very little biased (worst scenario – difference between true mean estimator and mean estimate - around 0.1) and we had around 90% coverage level. However, the vast majority of results indicates that our probabilistic solutions are well evaluated and are expected to provide valid results. Our new suggestion about including a categorical informative variable in the imputation model seems to help classification and may outperform FCS. It also proves that using the available extra information may help having better estimations and imputed granularity levels.

4.5 Application – MASTERPLANS exemplar

Here, we illustrate the granularity problem and its possible solutions (through traditional and probabilistic approaches) in the MASTERPLANS cohort data.

4.5.1 Data characteristics and granularity problem example

To demonstrate the utility of the developed approaches (figure 4.1) we applied to real-world biomedical and health datasets such as the studies in Lupus. For datasets D_1, D_2, D_3, we had datasets that contain data from ALMS, LUNAR and EXPLORER respectively. Suppose we were interested in the overall effect of ethnicity in drug response (BILAG score) on patients with SLE. For this purpose, we wanted to fit a linear regression model with BILAG score as the outcome, adjusting for ethnicity, age, gender, and creatinine (equation 4.2).

$$BILAG\ Score = Ethnicity + Age + Gender + Creatinine \quad (4.2)$$

For this research question, a granularity problem occured in ethnicity variable (see table 4.3). In one hand, ethnicity in ALMS had 14 levels, i.e., ‘Algerian’, ‘Asian’, ‘Black or African American’, ‘Caucasian’, ‘Cape Coloured’, ‘East Indian’, ‘Eritrean’, ‘Hispanic’, ‘Mexican Mestizo’, ‘Middle Eastern’, ‘Mixed’, ‘Moroccan’, ‘Native American’, ‘Nicaraguan’). On the other hand, ethnicity’s levels in LUNAR and EXPLORER were 3 i.e., ‘Caucasian’ and ‘Black or African American’ and ‘Other’. Therefore, in ALMS, ethnicity’s granularity was very high.

Table 4.3. Mapping between ethnicity's levels. Traditional VS Probabilistic data integration.

Traditional data integration	**Probabilistic data integration**
Ethnicity's levels	
'Caucasian'	'Caucasian'
'Black or African American'	'Black or African American'
'Other'	'Algerian'
	'Asian'
	'Cape Coloured'
	'East Indian'
	'Eritrean'
	'Hispanic'
	'Mexican Mestizo'
	'Middle Eastern'
	'Mixed'
	'Moroccan'
	'Native American'
	'Nicaraguan'

Table 4.4 summarises the information on granularity problem of ethnicity before and after lupus data sets' integration and gives us a summary of the ethnicity's data characteristics for the 530 patients included in the final analyses.

Table 4.4. Ethnicity's data characteristics after integrating lupus studies ALMS, LUNAR, EXPLORER.

Data characteristics	D_0, N=530 (%)	ALMS, N=204 (38.50)	LUNAR, N=107 (24.00)	EXPLORER, N=199 (37.50)
Ethnicity (%)				
Algerian	1 (0.19)	1 (0.50)		
Asian	67 (12.59)	67 (32.80)		
Black or African American	100 (18.90)	22 (10.80)	33 (26.00)	45 (22.60)

Cape Coloured	1 (0.19)	1 (0.50)		
Caucasian	251 (47.38)	90 (44.10)	41 (32.30)	120 (60.30)
East Indian	1 (0.19)	1 (0.50)		
Eritrean	1 (0.19)	1 (0.50)		
Hispanic	2 (0.38)	2 (1.00)		
Mexican Mestizo	14 (2.64)	14 (6.90)		
Middle Eastern	1 (0.19)	1 (0.50)		
Mixed	1 (0.19)	1 (0.50)		
Moroccan	1 (0.19)	1 (0.50)		
Native American	1 (0.19)	1 (0.50)		
Nicaraguan	1 (0.19)	1 (0.50)		
Other	87 (16.40)		53 (41.70)	34 (17.10)

4.5.2 Results

4.5.2.1 Traditional Data Integration – mapping common levels to the least granular

Traditionally, such problems are resolved by mapping all datasets to a common data model. This model would only include variables' levels that are present in all datasets, and, in case of granularity problem, levels that exist in all datasets (figure 4.1, table 4.3) i.e., 'Black or African American', 'Caucasian' and 'Other'. Le Sueur et al [129], came across this content heterogeneity issue in two variables in their data integration approach. In particular, ethnicity in some datasets included more categories and subsets than others (the same problem that we address here) and *visit time* was captured either as sequential visit numbers (i.e., 'visit 1', 'visit 2' etc) in one study or as time from baseline (in days, weeks, or months) in another study. They solved the granularity issue by harmonising to lowest granularity present e.g. in ethnicity across datasets were reduced to 3 levels ('Black or African American', 'White', 'Other').

Table 4.10 shows the coefficients for the linear regression model that estimated the drug response by applying equation 4.2 after applying complete case analysis in SLE data. We see that 'Other' is significant (p-value = 0.0393) and it implies that ethnicity is associated with drug response. We also see that gender is significant (p-value = 0.0219) and is associated with drug response. Concerning age (p-value = 0.0900) there is weak evidence to show significance and we may need a larger sample.

Table 4.5. Coefficients (estimate, standard error, t statistic and p-values) for linear regression model from equation 4.2 after applying complete case analysis in SLE data.

	estimate	**standard error**	**t statistic**	**p-value**
(Intercept)	5.5780	1.2099	4.6100	0.0000
Ethnicity				
Caucasian	0.9739	0.7332	1.3280	0.1847
Other	-1.5937	0.7713	-2.0660	0.0393 *
Age	0.0403	0.0237	1.6980	0.0900 .
Gender				
Male	-1.9346	0.8417	-2.2990	0.0219 *
Creatinine	0.4629	0.8915	0.5190	0.6038

*Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

4.5.2.2 Probabilistic Data Integration – multiple imputation

Le Sueur et al mention [129] in their paper that their traditional data integration approach to SLE data has a limitation as there is unavoidable loss of information either due to differences in granularity or data capture. For example, excluding specific groups changes the attributes of the patients who are being analysed and excluding patients and rows of data due to missingness, it makes smaller the sample size and introduces potential biases.

We chose to illustrate how our probabilistic approaches would answer the research question in equation 4.2. Therefore, we applied the suggested probabilistic data integration approaches FCS and FCSgroup to answer the research question shown in equation 4.2. We retained the highest number of levels in ethnicity which was 14 levels (table 4.3). For FCSgroup's imputation model, we added the informative categorical variable 'ethnicity$_{group}$' that classified patients based on the 'Other' ethnicity. With this additional variable we tried to eliminate misclassification in imputation. For both approaches FCS and FCSgroup, the chosen parameters were 20 imputed datasets, 20 iterations, and seed number 384839.

First, we applied FCS imputation to answer to equation 4.2's research question. In Table 4.6 we see the coefficients (estimate, standard error, t statistic and p-values) for linear regression model obtained from equation 4.2 after applying FCS in SLE data to solve the granularity problem. In figure 4.19, we can see the barplot for the variable: 'Ethnicity', in second content

heterogeneity problem. The top figure shows the observed values, and the bottom figure shows the imputed data in FCS. In table 4.12 we find equation 4.2's coefficients after applying FCS imputation in SLE data.

Secondly, we apply FCSgroup imputation to answer to equation 4.2's research question. In Table 4.7 we see the coefficients for linear regression model obtained from equation 4.2 after applying FCSgroup in SLE data to solve the granularity problem. In figure 4.20, the top subfigure shows the observed values, and the bottom subfigure shows the imputed data in FCSgroup.

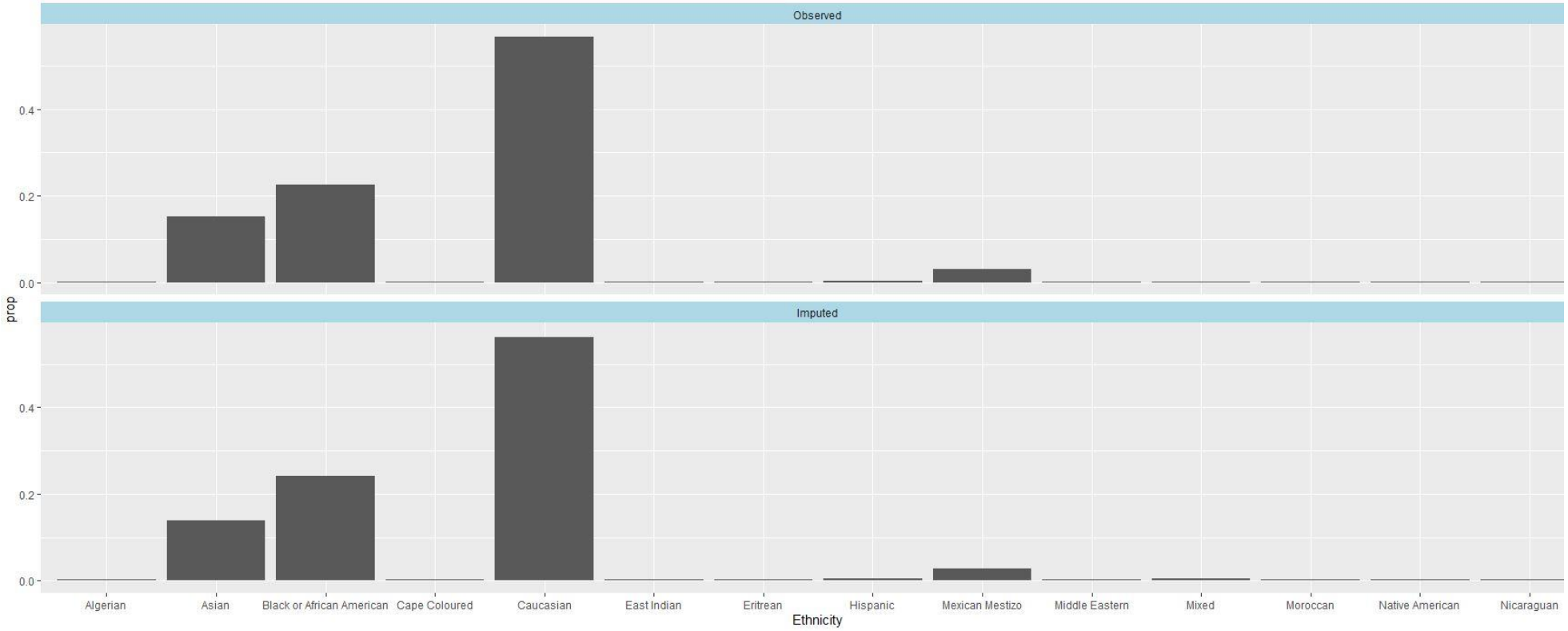


Figure 4.19. Barplot for the variable: 'Ethnicity', in content heterogeneity problem 2. Top figure shows the observed values and bottom figure shows the imputed data for each imputation in FCS.

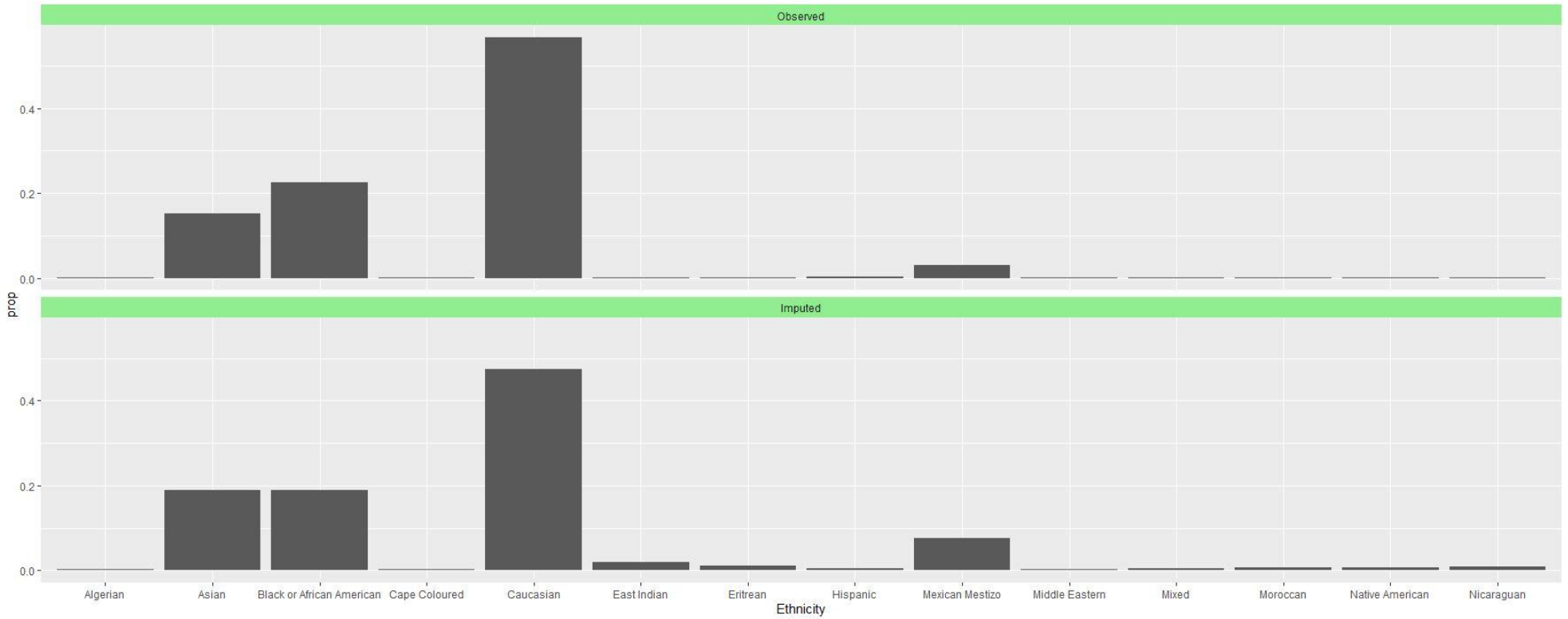


Figure 4.20. Barplot for the variable: 'Ethnicity', in content heterogeneity problem 2. Top figure shows the observed values and bottom figure shows the imputed data for each imputation in FCSgroup.

Table 4.6. Coefficients (estimate, standard error, t statistic and p-values) for linear regression model from equation 4.2 after applying FCS in SLE data.

	estimate	**standard error**	**t statistic**	**p-value**
(Intercept)	-0.4413	5.5811	-0.0791	0.9370
Ethnicity				
Asian	3.5605	5.5269	0.6442	0.5198
Black or African American	5.6116	5.4971	1.0208	0.3079
Cape Coloured	-1.2059	7.9282	-0.1521	0.8792
Caucasian	6.6572	5.4796	1.2149	0.2250
East Indian	5.3604	8.6922	0.6167	0.5381
Eritrean	9.5691	7.7793	1.2301	0.2194
Hispanic	3.6259	6.7893	0.5341	0.5935
Mexican Mestizo	1.3876	5.7023	0.2433	0.8078
Middle Eastern	-1.4825	7.8027	-0.1900	0.8494
Mixed	1.5852	7.3992	0.2142	0.8305
Moroccan	4.5089	7.7296	0.5833	0.5599
Native American	1.4248	7.9580	0.1790	0.8580
Nicaraguan	10.9610	7.8240	1.4010	0.1619
Age	0.0441	0.0239	1.8431	0.0659 .
Gender				
Male	-1.8161	0.8417	-2.1576	0.0314 *
Creatinine	0.5338	0.9548	0.5591	0.5764

*Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Table 4.7. Coefficients (estimate, standard error, t statistic and p-values) for linear regression model from equation 4.2 after applying FCSgroup in SLE data.

	estimate	**standard error**	**t statistic**	**p-value**
(Intercept)	0.1907	5.2396	0.0364	0.9710
Ethnicity				
Asian	3.7827	5.1608	0.7330	0.4640
Black or African American	5.6627	5.1161	1.1068	0.2690
Cape Coloured	-1.3142	7.1144	-0.1847	0.8535
Caucasian	6.5507	5.1010	1.2842	0.1997

East Indian	8.7870	5.4468	1.6133	0.1075
Eritrean	12.7096	6.4423	1.9728	0.0500 *
Hispanic	3.6611	6.8804	0.5321	0.5951
Mexican Mestizo	2.1342	5.1624	0.4134	0.6795
Middle Eastern	-2.1434	7.6921	-0.2787	0.7806
Mixed	2.2200	6.3032	0.3522	0.7248
Moroccan	4.4725	6.4906	0.6891	0.4912
Native American	2.6655	6.0514	0.4405	0.6598
Nicaraguan	7.5501	6.0377	1.2505	0.2118
Age	0.0528	0.0240	2.2033	0.0280 *
Gender				
Male	-1.7109	0.8501	-2.0125	0.0447 *
Creatinine	-0.3564	1.1341	-0.3142	0.7536

*Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

4.6 Discussion

This work was motivated by a scope to explore probabilistic solutions such as multiple imputation to situations where for some studies a certain variable has not been collected in the same level of granularity. In this context, we introduced the second problem of content heterogeneity which is that of granularity problem. We gave an overview of this specific content heterogeneity problem from a theoretical perspective and argued that as with the missing variable problem, the granularity problem too could be translated to systematically missing data problem and therefore solved with established methods like imputation. The motivation was that we could ask a research question including the most of information and that it would allow us borrow information across studies. We described our suggested probabilistic method and how to impute data based on methods that already exist in the literature. This work included an exploration of the general applicability of the suggested method, and comparison of results of the proposed integration techniques with gold standard results through statistical simulation studies. Moreover, to demonstrate their utility, we applied them to real-world biomedical and health datasets such as the studies in SLE that were used for our examples.

This study offers an alternative to Complete Records analysis with cutting-edge probabilistic approaches to solve granularity issues. The probabilistic approaches give mostly valid results across a range of model errors and at all times outperformed the traditional data integration approach. To the best of our knowledge, there is no current research that has

applied imputation methods like FCS to solve a content heterogeneity problem such as the granularity problem. In addition, this study suggests an innovative idea, to use the traditional data integration approach in the imputation model in the form of a categorical informative group variable. The FCSgroup model takes advantage of the given information concerning subgroups and levels in categorical levels in imputation.

Summary of the main findings

In the simulated studies for handling missing data, in the form of granularity problem after integrating data from different biomedical data sources, parametric imputations produced estimates with no material bias for a linear model in data artificially introduced MCAR missingness. FCS produced unbiased estimates and almost 95% confidence intervals in most cases. In very few cases - where we had different model errors between studies - the coverage probability was smaller than 95% suggesting than confidence intervals may have been conservative. Overall, our results suggest that a probabilistic method such as multiple imputation by chained equation methods is useful for imputing complex health/biomedical datasets in which there is a granularity problem in the common categorical variables in the datasets after integration and especially for linear regression models.

Imputation method including the extra informative variable (FCSgroup)

The simulation studies showed that the usage of the extra informative variable in the imputation model helped the imputation itself and improved data classification. FCSgroup showed accuracy and in some cases outperformed standard FCS.

Strengths

This study has numerous strengths that are worth mentioning. It has been one of the first research studies that talk about and investigate solutions to granularity problems of categorical variables after structured data integration in biomedical data sources. It has been the first study in a biomedical concept that describes this content heterogeneity problem in detail, solves it with a theoretical solution, evaluates and compares traditional and probabilistic approaches using a series of simulation studies and illustrates the paradigm in real-world data. This work also extends past research on data integration challenges in SLE by providing an alternative solution to differences in granularity thus addressing the limitation of inevitable loss of information mentioned by Le Sueur et al. [129]. Different methods were applied in real-world data (MASTERPLANS) and the results show that probabilistic approaches do offer unbiased results especially in comparison to complete record analysis.

More specifically about our illustration to real-world lupus data. Results show a practically useful gain in information over Complete Records with FCS and FCSgroup. However, we need to be careful with the analysis as individuals in those lupus studies were asked about their ethnicity and they clarified that they belong to the '*Other'* category and not in '*Caucasian'* or '*Black or African American'*. MICE package does not have an option to exclude categories in imputation in nominal/categorical variable. Our additional step in imputation - FCSgroup – tries to do exactly that, eliminate misclassification and it shows that it can do it.

Limitations

Although, this extended study has strengths, it was based on real-world problem, analyses were realistically complex and many scenarios were investigated, it has various drawbacks. The most important is that we chose for simplicity reasons to ignore practice/study level clustering at the imputation and analysis stages. If patients from the same practice/study are more similar than patients from different practices/studies, the variance of parameter estimates may have been underestimated and parameter estimated may also be biased. Whereas we presented promising results in two simulation studies where model errors were different in the two datasets. This may need further investigation and we could properly choose other imputation methods such as joint modelling multiple imputation [102] and multilevel imputation methods [114], [135].

Another limitation in producing general application of the proposed methods is that their evaluation was based on specific data generating mechanisms and scenarios. However, our simulation studies were complex enough and included many situations. We extended the simulation study to a larger number of data sources, a variety of study sizes and model errors.

Another thought for future work is to compare methods' performance when the fitted model is logistic, or a multinomial logistic model. In broad terms, linear regression is used to estimate the dependent variable in case of a change in independent variable whereas logistic regression is used to calculate the probability of an event. In order to perform logistic regression successfully we need to choose the correct variables into the model, avoid the use of highly correlated variables, probably restrict the number of variables, be careful how to handle continuous variables (chosen categories and loss of information), check the assumptions regarding the relationship between input and output variables, and interpret the results carefully (odds VS risk) [136]. Nevertheless, we expect to see that our suggested probabilistic methodologies that solve data integration problems lead to similar results and conclusions when using a logistic model in our study. We consider FCSgroup to be

promising, but it should be tested on a larger range of data sets and in simulations to explore whether it gives unbiased estimates where there are nontrivial nonlinearities or interactions in imputation models. However, if we see the illustration in the lupus data, we need to be careful as if the majority of data are captured in very few levels, then the imputation follows the observed probability and that means that levels with very few observations may not have as many values after imputation as it would be in reality. Therefore, we need to be careful when our research interest is about finding out specific things about clustered/subgrouped individuals. Thus, inclusion of the extra informative variable in the imputation model (FCSgroup) may be useful when researchers are interested in limiting the imputed values to levels or ranges and focusing on clustering/subgroups.

In summary, the results of our comprehensive set of simulation studies show that researchers can use probabilistic methods for solving content heterogeneity presented as granularity problem when studies are collected from similar cohorts. FCS is one of the most recommended methods for multiple imputation in health and biomedical data and we have shown that FCS and FCSgroup work reasonably well under artificially introduced missingness completely at random for granularity issues in realistically complex data sets. FCS and FCSgroup should be further investigated in some extreme case scenarios.

Chapter 5: Mixed numeric and non-numeric data types

5.1 Introduction

In Chapters 3 and 4, we focused on two types of content heterogeneity that may exist in integrated datasets, systematically missing values and varying granularity. In this chapter we address a third type of content heterogeneity, one that arises from representational differences of a common variable across different datasets. Similar to previous methodological chapters, in this chapter we describe the content heterogeneity problem using an example (Section 5.2), and then our generic probabilistic approach to data harmonisation (Section 5.3). Subsequently, we evaluate the suggested approach through simulation studies (Section 5.4), and we apply and evaluate it to real-world data (Section 5.5). We finish the chapter with a discussion about utility of the methods on mixed type problem, limitations, and future steps for improvement (Section 5.6).

5.2 Problem identification of mixed numeric and non-numeric data types

We define content heterogeneity caused by mixed numeric and non-numeric data types as the situation where a numerical attribute is represented numerically in some datasets and non-numerically (in categories) in other datasets. An example is where participants' age is recorded as a categorical variable with categories '0-20', '21-40', '41-60', '>60' in dataset A, and recorded as an integer number in dataset B. In such cases there would be uncertainty regarding the specific ages of patients from dataset A, if we try to express them as integer numbers (for integration with dataset B). Traditionally, this problem has been solved by mapping all values to the least granular data type (in the example the age grouping by 20 years). The evidence instance is linked to the age range defined in the database by the minimum and the maximum values of the range [137]. This is likely to result in severe loss of information as more datasets are integrated.

In this chapter, we propose an alternative, probabilistic approach for integrating datasets with content heterogeneity caused by mixed numeric and non-numeric data types.

5.3 Methodology

As in previous chapters we assume that the variable in question is relevant to a defined research question, and which we aim to address using a regression model. As depicted in figure 4.1, we stack the extracted structured data from D_1, D_2 to D_n in one large integrated dataset D_0. To address the research question, we select the relevant variables from D_0 that will be included in the regression model. The problem that we have here is

that at least one variable that will be included in the model is represented using different data types across D_1, …, D_n (figure 5.1 – yellow, orange, purple datasets).

From a probabilistic point, we have already showed in previous chapters that imperfect harmonisation of different data sources is not problematic if we can determine which information each of the data sources brings in answering our research question. The tasks for handling the third content heterogeneity problem are similar to the theoretical solutions that solve systematically missing values and granularity presented in Chapters 3.3 and 4.3. As in previous chapters we argue that by translating the mixed type problem to missing data problem, we can solve it using the common and gold standard FCS method. Likewise, we add an informative categorical variable (blue column in figure 5.1) that keeps the information to which mixed type variable's level each individual belongs (different shades in blue column in figure 5.1).

The mixed type variable has both categories and continuous values. The traditional data integration approach would be to convert the mixed type to a categorical variable – mapping common levels (black box in figure 5.1). But we would like to have the most granular data, therefore to change the mixed type variable to continuous. Hence, we introduce missingness in the mixed type variable – categorical data are removed (black squares in figure 5.1). Therefore, in the integrated dataset we have missing data which we impute based on all the available information in D_0, excluding (FCS) or including (FCSgroup) the categorical informative 'group' variable in the imputation model (red boxes in figure 5.1).

Our goal ideally was on the creation of a single, very large table that obtained all the information that was described by the stacked tables that we have integrated (figure 5.1 – yellow, orange, purple datasets). Moreover, that integrated table had systematically missing values in the form of the mixed type content heterogeneity problem.

5.4 Simulation studies

In this section, we show various simulation studies' results designed to evaluate the use of our probabilistic methods to solve content heterogeneity in the form of a mixed type problem. These simulation studies had similar design, scenarios, and data generating mechanisms to simulations presented in chapters 3 and 4. We start with a reminder of the simulation design and the scenarios with explicitly showing how they changed to adapt to mixed type problem. We begin in Subsection 5.4.1 with the general design of the simulation and in 5.4.2 is the description of the performance measures. Then, we continue

in other subchapters and investigate what happens with different number of individuals (N) per study, number of studies (Dn) and model errors (e_i). The number of imputations (m) and the number of iterations (it) are set to five (default). All simulation results can be found in Appendix C.

5.4.1 Simulation study design

Simulation studies followed similar design presented in chapter 4, figure 4.1. Simulation's procedure that shows how the data were integrated, how mixed type variable problem was solved through different methods and their comparison is presented in figure 5.1.

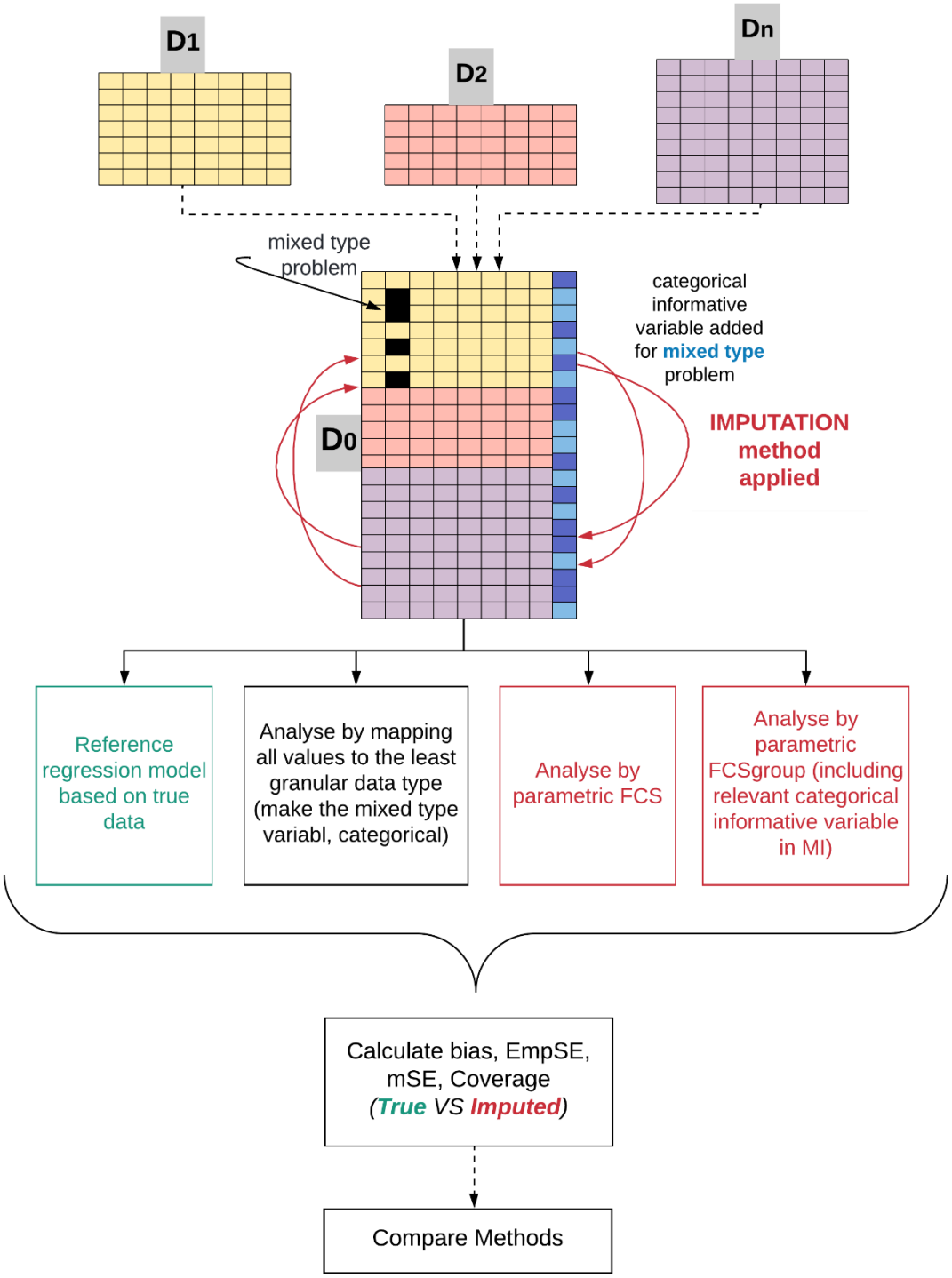


Figure 5.1. Simulation's procedure to show how the data are integrated, how mixed type variable problem is solved through different methods and their comparison.

FCS: Fully Conditional Specification; *FCSgroup:* FCS including informative 'group' variable.

As we see in figure 5.1, we followed similar steps with previous chapters' simulations (3.4.1 and 4.4.1). Differences are found in betas used to simulate outcome Y, starting seed number and application of content heterogeneity problem. We simulated same continuous

variables X_1, X_2, and categorical X_3 and we used similar data generating mechanism as in the previous chapter to simulate outcome Y (equation 5.1). We had different betas in every chapter, particularly in this one we had: $\beta_0 = 1.4938$, $\beta_1 = -1.7483$, $\beta_2 = 0.4255$, $\beta_3 = 1.4589$, $\beta_4 = -0.9554$, $\beta_5 = 0.4845$. Y is complete, had no missing data and had known dependency on X variables.

$$Y_i = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * I(X_3 = B) + \beta_4 * I(X_3 = C) + \beta_4 * I(X_3 = D) + e_i \quad (5.1)$$

After study datasets' integration in D_0, we fit a linear regression model $Y \sim X_1 + X_2 + X_3$ to the true Full Data and stored estimates $\hat{\theta}$, standard errors $\widehat{se}(\hat{\theta})$ and coverage level (green box in figure 5.1). Afterwards, we applied the third content heterogeneity problem. In D_0, there was one numeric representation and multiple other datasets with a categorical representation. For example, X_2 was a categorical variable (two levels: [min-mean), [mean-max]) for individuals from D_1 and a continuous variable for individuals from D_2.

Traditional solution

We solved mixed type problem applying the traditional approach where we kept the least common denominator in X_2. So, X'_2 had two levels [min-mean), [mean-max] in D_0. Then, we fit a linear regression model: $Y \sim X_1 + X'_2 + X_3$ in D_0, analysed Complete Records and stored estimates $\hat{\theta}$, standard errors $\widehat{se}(\hat{\theta})$ and coverage level (black boxes in figure 5.1).

Probabilistic solutions

In D_0, we created the categorical informative 'group' variable 'X_2group' that grouped individuals in two levels: [min-mean) and [mean-max] based on their X_2 values (figure 5.1 – blue column with different blue shades). Afterwards we introduced data missingness due to mixed type problem. For example, for the last 1000 individuals in D_0 (individuals from D_2), we set their X_2 values to missing. So, X_2 had systematically missing data (black squares in figure 5.1) for individuals from D_2 and complete continuous data for individuals from D_1 (colourful squares in the relevant column in figure 5.1). We then solved the third content heterogeneity problem by applying the probabilistic approaches FCS, and FCSgroup as presented in figures 4.1 and 5.1 using predictive mean matching as a method (red boxes in figure 5.1). Afterwards, we fit linear regression models $Y \sim X_1 + X''_2 + X_3$ after imputations in D_0 where X''_2 contains the imputed values. We analysed imputed datasets using Rubin's rules and store estimates $\hat{\theta}$, standard errors $\widehat{se}(\hat{\theta})$ and

coverage level (last two black boxes in figure 5.1). All imputation analyses were performed with R package *mice* freely available on CRAN [98]. We used 0605215 as starting seed number to generate a sequence of random numbers. In the next section, we present the simulations' results.

5.4.2 Performance measures and scenarios

Similar to previous chapters, across all simulation scenarios, we present the results of the analyses with different performance measures in terms of bias, mSE estimated from the models, EmpSE (i.e., standard deviation of the simulation estimates) and confidence coverage level. A valid method should yield unbiased results, similar model and empirical standard errors, and coverage levels close to 95%. In this chapter, our estimand θ was the estimate $\hat{\theta}_i$ of X_2 coefficient in each model fit. In tables 5.1 and 5.2 we see ten scenarios used to generate data from figure 5.1 for mixed type problem. In each simulation, we have data missingness, due to the mixed type problem. For example when missingness was applied to X_2 from D_2, it means that in D_2, the numerical variable X_2 had been mutated to a categorical variable with two levels.

Table 5.1. Scenarios 1 - 5 used to generate data from Figure 5.1 for content heterogeneity type 3.

	Scenario 1	**Scenario 2**	**Scenario 3**	**Scenario 4**	**Scenario 5**
Number of individuals per study (N)	200	1000	200	1000	D_1: 200, D_2:150, D_3:50, D_4:75, D_5:100
Model error e_i: (same for each study)	N~(0,0.2) N~(0,2) N~(0,20)	N~(0,0.2) N~(0,2) N~(0,20)	N~(0,0.2) N~(0,2) N~(0,20)	N~(0,0.2) N~(0,2) N~(0,20)	N~(0,0.2) N~(0,2) N~(0,20)
Number of studies (D)	2	2	5	5	5
Imputations (m)	5	5	5	5	5
Missingness applied to:	X_2 from D_2	X_2 from D_2	X_2 from D_4, D_5	X_2 from D_2, D_5	X_2 from D_4, D_5

Table 5.2. Scenarios 6 - 10 used to generate data from Figure 5.1 for content heterogeneity type 3.

	Scenario 6	**Scenario 7**	**Scenario 8**	**Scenario 9**	**Scenario 10**
Number of individuals per study (N)	D_1:800, D_2:150, D_3:50, D_4:75, D_5:350, D_6:200, D_7:150, D_8:500, D_9:750, D_{10}:100	100	200	500	1000
Model error e_i: (same for each study)	N~(0,0.2) N~(0,2) N~(0,20)	e_1: N~(0,1.2), e_2: N~(0,1.3)	e_1: N~(0,1.2), e_2: N~(0,1.3)	e_1: N~(0,1.2), e_2: N~(0,1.3)	e_1: N~(0,1.2), e_2: N~(0,1.3)
Number of studies (D)	10	2	2	2	2
Imputations (m)	5	5	5	5	5
Missingness applied to:	X_2 from D_3, D_6, D_9, D_{10}	X_2 from D_2	X_2 from D_2	X_2 from D_2	X_2 from D_2

5.4.3 Results

Simulations with studies of same sizes (Scenarios 1 - 4)

Scenario 1

Here, we simulated data for two studies, **each with 200 patients**. For each simulated dataset, we had data missingness, due to the mixed type problem, to X_2 from D_2. This means that in D_2, the numerical variable X_2 had been converted to a categorical variable with two levels. In Figure 5.4 we see a graphical representation of the main results. Simulation results indicate that FCS and FCSgroup gave unbiased results and good coverage of the confidence interval for medium and large model errors whereas coverage was around 90% and unbiased estimates when the model error was small. The Complete Records approach appeared to be not a good solution to solve the mixed type issue and should be avoided.

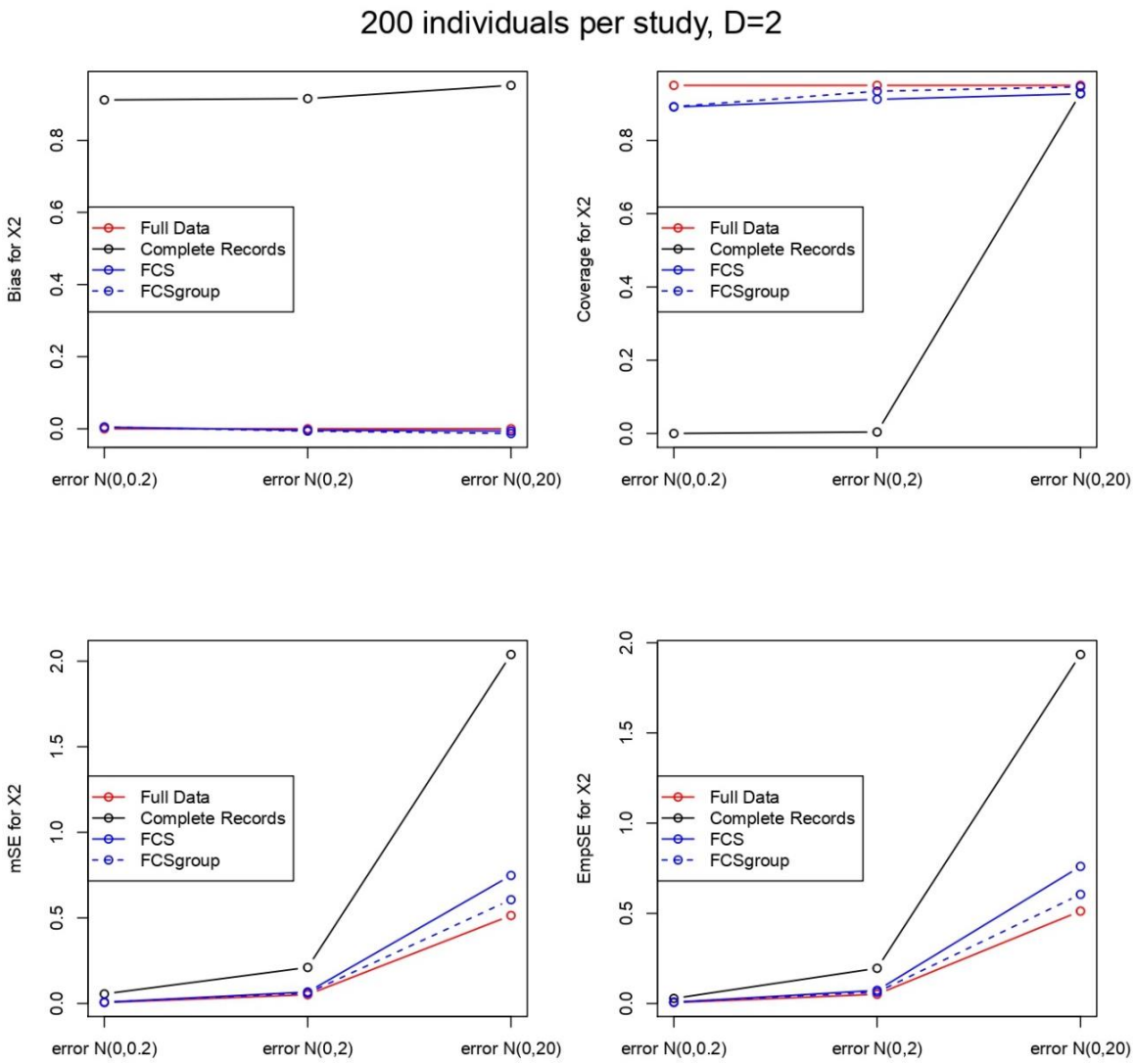


Figure 5.2. Main results from scenario 1's simulation study: Comparison of Bias, Coverage level, mSE and EmpSE for X_2 after 1000 simulations with Full Data (red line), handling mixed type with Complete Records (black line), FCS (blue line), FCSgroup (broken blue line) for three model errors.

FCS: Fully Conditional Specification; *FCSgroup:* FCS imputation model included X'$_2$; *mSE:* mean model Standard Error; *EmpSE:* mean Empirical Standard Error

Scenario 2

This scenario was like scenario 1 with a difference of **1000 individuals per study**. We present the simulation results in figure 5.5 with e_i: (N ~ (0, 0.2)), (N ~ (0, 2)), (N ~ (0, 20)). Estimates were unbiased for both imputation models when e_i: N ~ (0, 0.2). FCS had good coverage in comparison with FCSgroup whose coverage is around 90% when e_i: N ~ (0, 0.2). For e_i: N ~ (0, 2), both imputation methods had unbiased estimates and good coverage. For e_i: N ~ (0, 20), FCS and FCSgroup had good coverage and very good estimates. Complete Records seemed to be an incompatible technique and led to large biases.

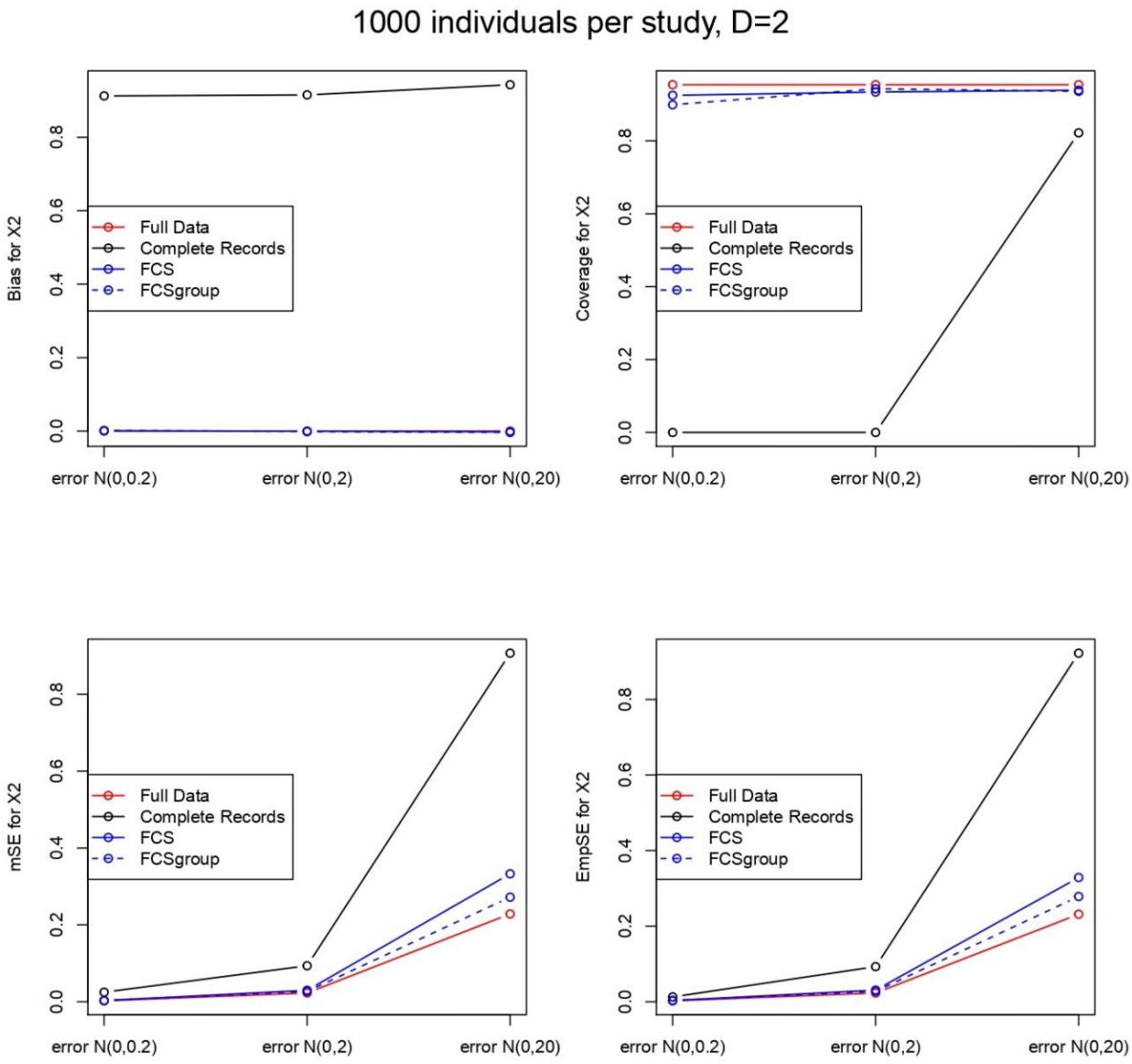


Figure 5.3. Main results from scenario 2's simulation study: Comparison of Bias, Coverage level, mSE and EmpSE for X_2 after 1000 simulations with Full Data (red line), handling mixed type with Complete Records (black line), FCS (blue line), FCSgroup (broken blue line) for three model errors.

FCS: Fully Conditional Specification; *FCSgroup:* FCS imputation model included X'$_2$; *mSE:* mean model Standard Error; *EmpSE:* mean Empirical Standard Error

Scenarios 3 – 4

In scenarios 3-4, we simulated data for **five studies**, each with the same number of individuals per study. In Scenario 3 **each study has 200 individuals** (so D_0 had 1000 individuals in total) and in scenario 4 **each study had 1000 individuals** (so D_0 has 5000 individuals in total). For each simulated dataset, we chose two random studies (D_4 and D_5 for scenario 3 and D_2 and D_5 for scenario 4) to apply data missingness due to mixed type issue in X_2 in D_0. We present the simulation results in figures 5.4 and 5.5 for scenarios 3 and 4 respectively. In scenarios 3 and 4, analyses of datasets imputed with FCS and

FCSgroup models gave great results in terms of bias, precision, and confidence interval coverage in any examined model error (figures 5.6 – 5.9).

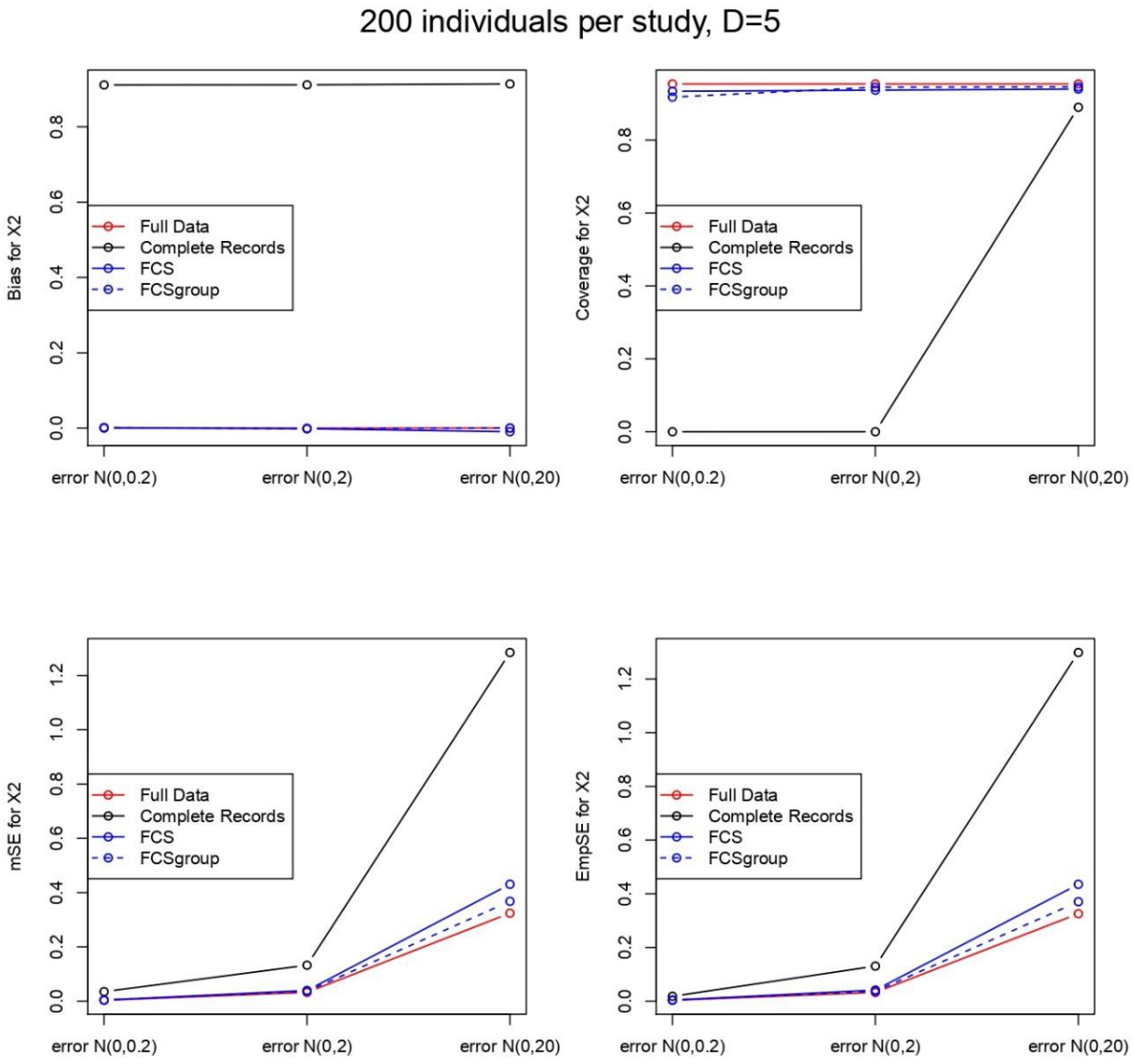


Figure 5.4. Main results from scenario 3's simulation study: Comparison of Bias, Coverage level, mSE and EmpSE for X_2 after 1000 simulations with Full Data (red line),

handling mixed type with Complete Records (black line), FCS (blue line), FCSgroup (broken blue line) for three model errors.

FCS: Fully Conditional Specification; *FCSgroup:* FCS imputation model included X'$_2$; *mSE:* mean model Standard Error; EmpSE: mean Empirical Standard Error

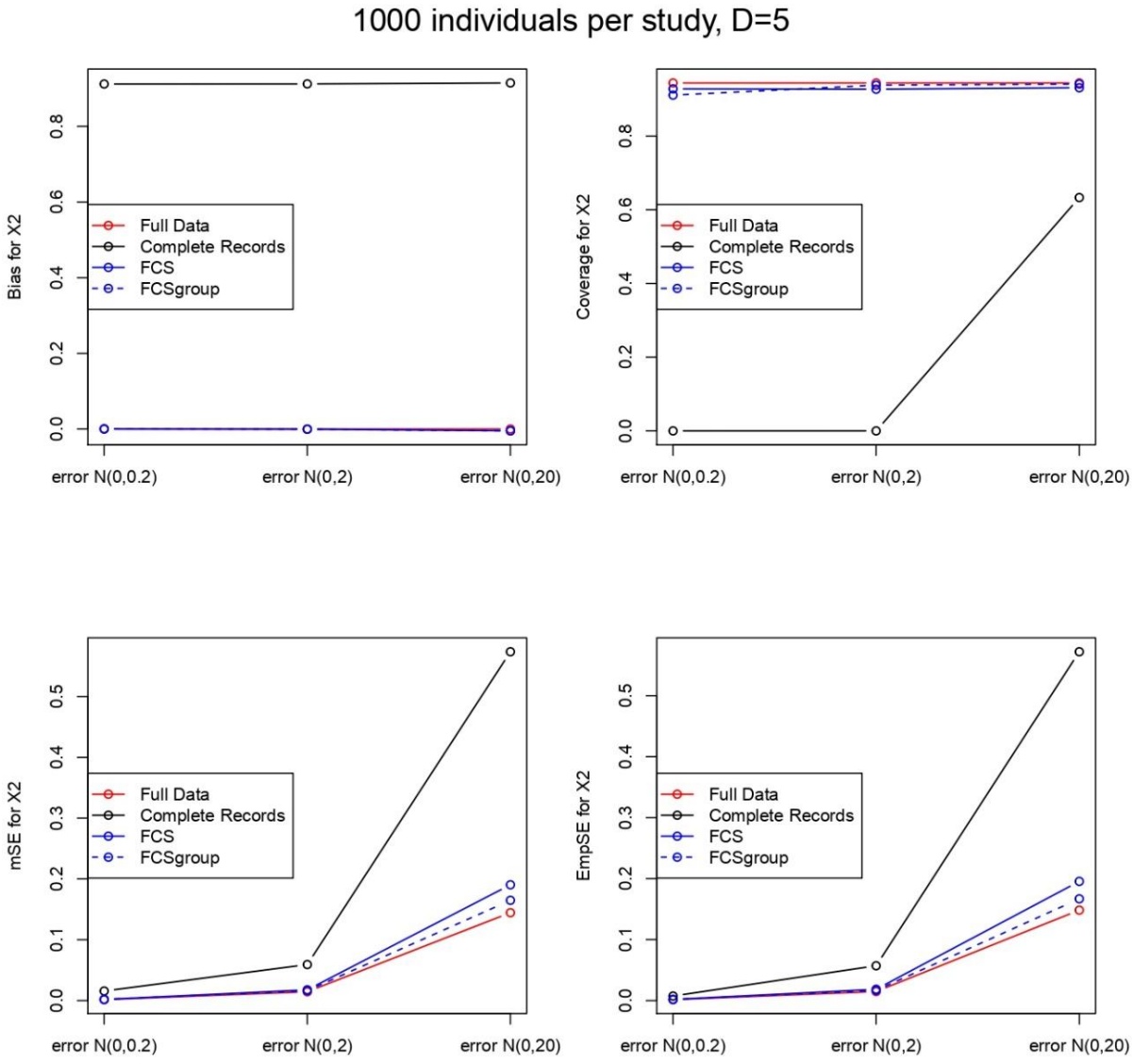


Figure 5.5. Main results from scenario 4's simulation study: Comparison of Bias, Coverage level, mSE and EmpSE for X_2 after 1000 simulations with Full Data (red line), handling mixed type with Complete Records (black line), FCS (blue line), FCSgroup (broken blue line) for three model errors.

FCS: Fully Conditional Specification; *FCSgroup:* FCS imputation model included X'$_2$; *mSE:* mean model Standard Error; *EmpSE:* mean Empirical Standard Error

Overall, in all scenarios when applying the Complete Records approach the regression estimates were highly biased which led to very large standard errors and great under-coverage. The results show that FCS and FCSgroup were valid methods to solve mixed type problem after data integration. In general, we cannot conclude if FCSgroup

performed better than FCS. mSE and EmpSE (figures 5.8 and 5.9) were similar per scenario, and they decreased as the integrated dataset's size increased. We also observe that FCSgroup had the closest mSE and EmpSE to Full Data. For all imputation methods the larger the integrated dataset, the smaller the bias.

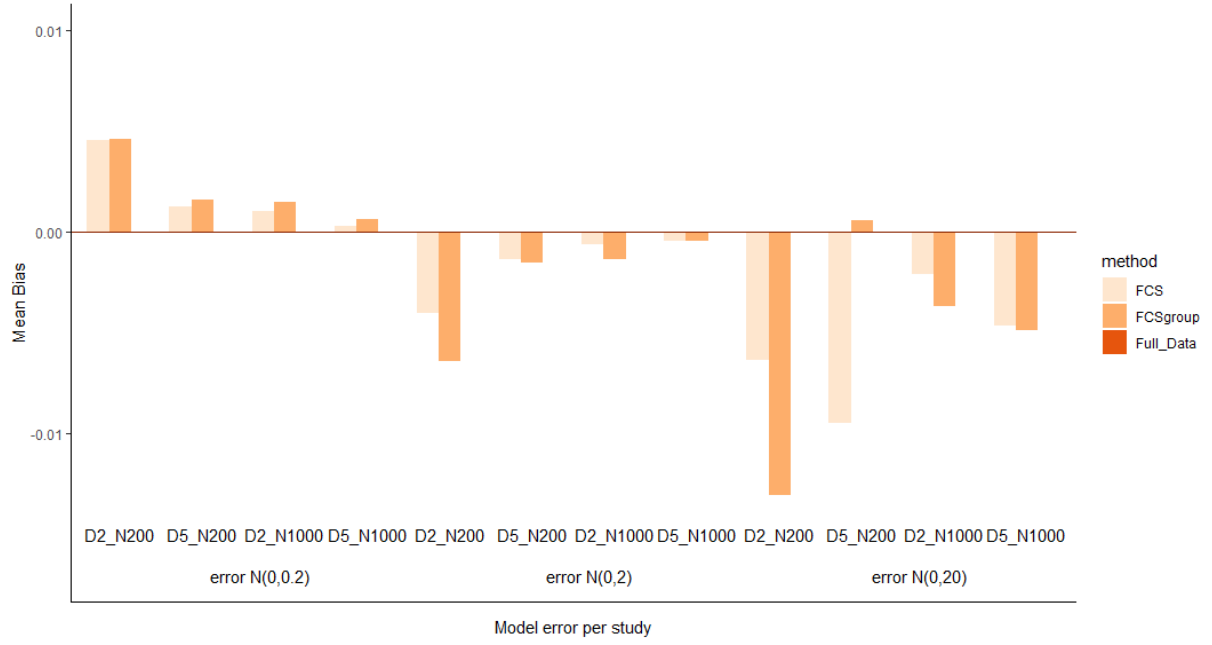


Figure 5.6. Mean bias for X_2 for '2 datasets, N=200 per dataset' (D2_N200), '5 datasets, N=200 per dataset' (D5_N200), '2 datasets, N=1000 per dataset' (D2_N1000) and '5 datasets, N=1000 per dataset' (D5_N1000) for the three model errors.

FCS: Multiple imputation by fully conditional specification; *FCSgroup:* FCS imputation model included X'_2.

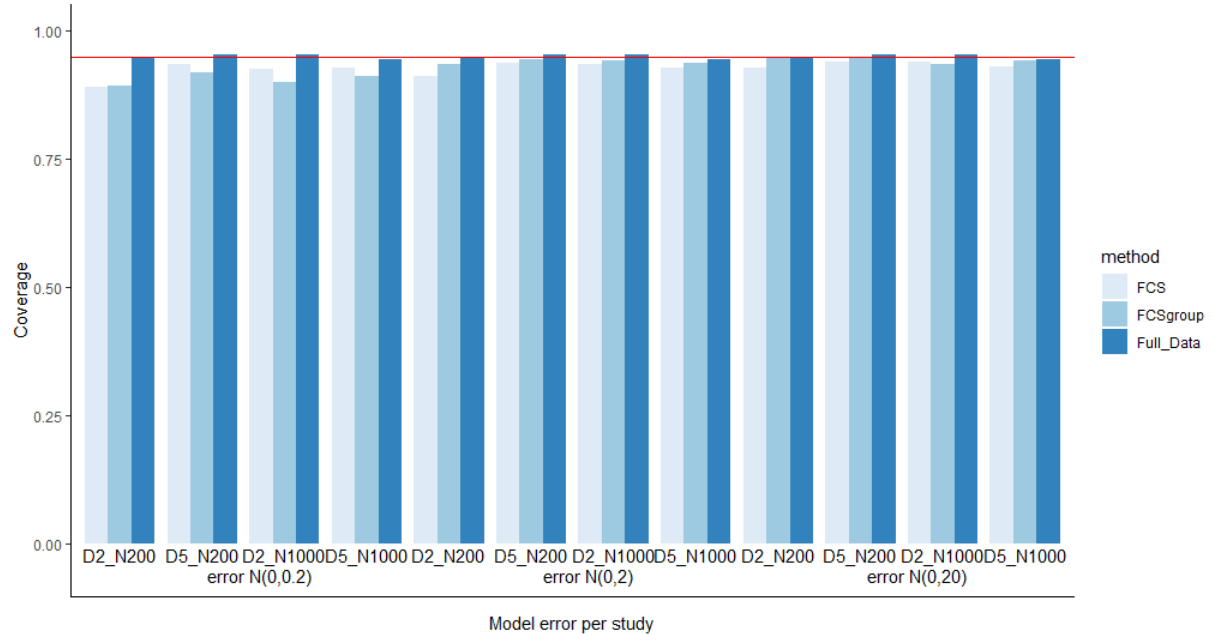


Figure 5.7. Coverage for X_2 for '2 datasets, N=200 per dataset' (D2_N200), '5 datasets, N=200 per dataset' (D5_N200), '2 datasets, N=1000 per dataset' (D2_N1000) and '5 datasets, N=1000 per dataset' (D5_N1000) for the three model errors.

FCS: Multiple imputation by fully conditional specification; *FCSgroup:* FCS imputation model included X'$_2$.

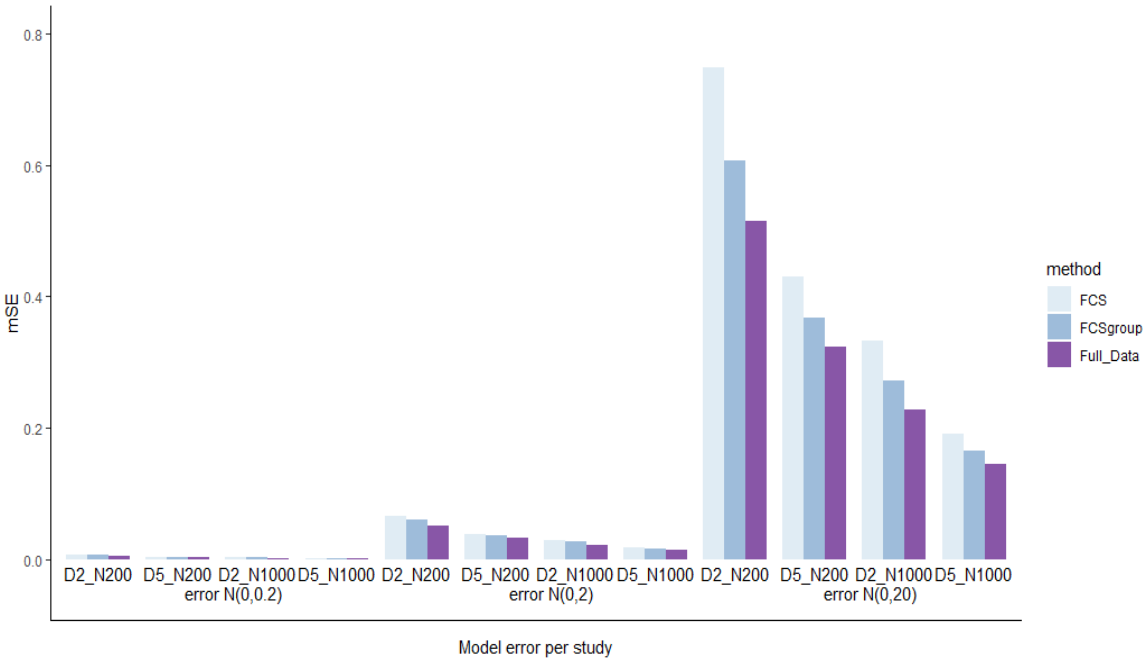


Figure 5.8. mSE for X_2 for '2 datasets, N=200 per dataset' (D2_N200), '5 datasets, N=200 per dataset' (D5_N200), '2 datasets, N=1000 per dataset' (D2_N1000) and '5 datasets, N=1000 per dataset' (D5_N1000) for the three model errors.

FCS: Multiple imputation by fully conditional specification; *FCSgroup:* FCS imputation model included X'$_2$; *mSE:* mean model standard error.

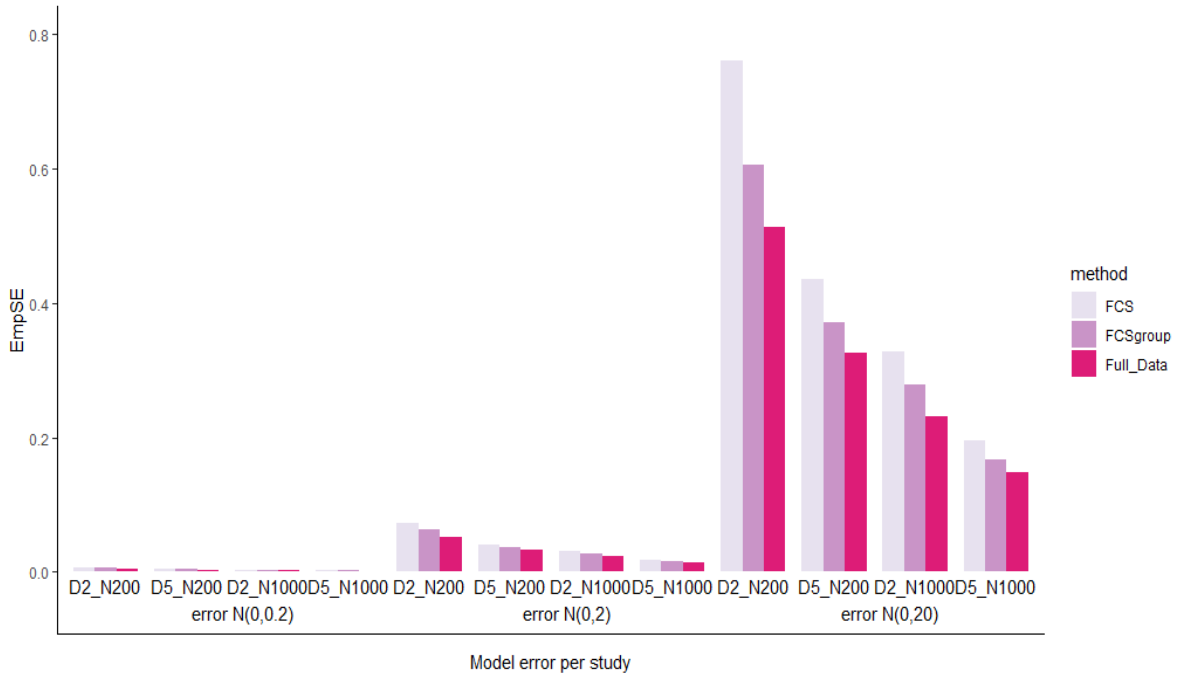


Figure 5.9. EmpSE for X_2 for '2 datasets, N=200 per dataset' (D2_N200), '5 datasets, N=200 per dataset' (D5_N200), '2 datasets, N=1000 per dataset' (D2_N1000) and '5 datasets, N=1000 per dataset' (D5_N1000) for the three model errors.

FCS: Multiple imputation by fully conditional specification; *FCSgroup:* FCS imputation model included X'$_2$; *EmpSE:* mean empirical standard error.

Simulations with studies of different size and same model error per study (scenarios 5-6)

Scenarios 5 – 6

In scenario 5, we simulated data for **five studies**, each with **different number of individuals** (D_1:200, D_2:150, D_3:50, D_4:75, D_5:100). We chose X_2 to have mixed type issue and therefore data missingness in two studies (D_4 and D_5). Simulation results for scenario 5 are in figure 5.9. FCS and FCSgroup produced almost identical results with true Full Data. FCS had slightly better results than FCSgroup in terms of bias but again almost the same. FCSgroup had higher coverage and lower mSE and EmpSE than FCS. Both probabilistic models outperformed completely Complete Records which once again was not suggested as an integration approach.

In scenario 6, we simulated data from **ten studies**, each with **different number of individuals** (D_1:800, D_2:150, D_3:50, D_4:75, D_5:350, D_6: 200, D_7:150, D_8:500, D_9:750, D_{10}:100). We decided X_2 to have data missingness due to mixed type issue in four studies (D_3, D_6, D_9, and D_{10}). Results for scenario 6 are shown in figure 5.10. As in scenario 5, FCS and FCSgroup produced identical results with true data. Complete Records produced large bias in estimates and under-coverage.

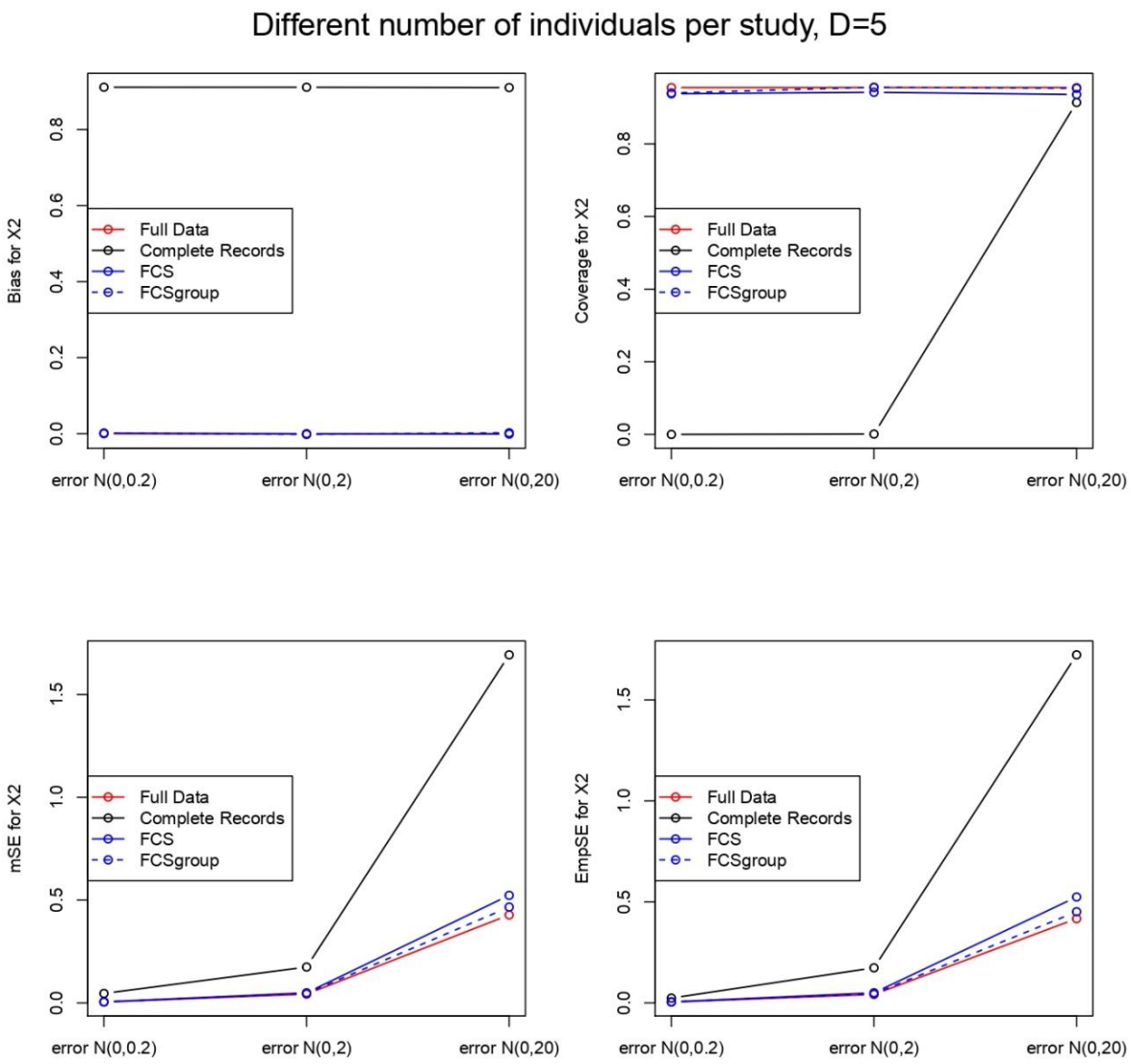


Figure 5.10. Main results from scenarios 5's simulation study: Comparison of Bias, Coverage level, mSE and EmpSE for X_2 after 1000 simulations with Full Data (red line), handling mixed type with Complete Records (black line), FCS (blue line), FCSgroup (broken blue line) for three model errors.

FCS: Fully Conditional Specification; *FCSgroup:* FCS imputation model included X'$_2$; *mSE:* mean model Standard Error; *EmpSE:* mean Empirical Standard Error

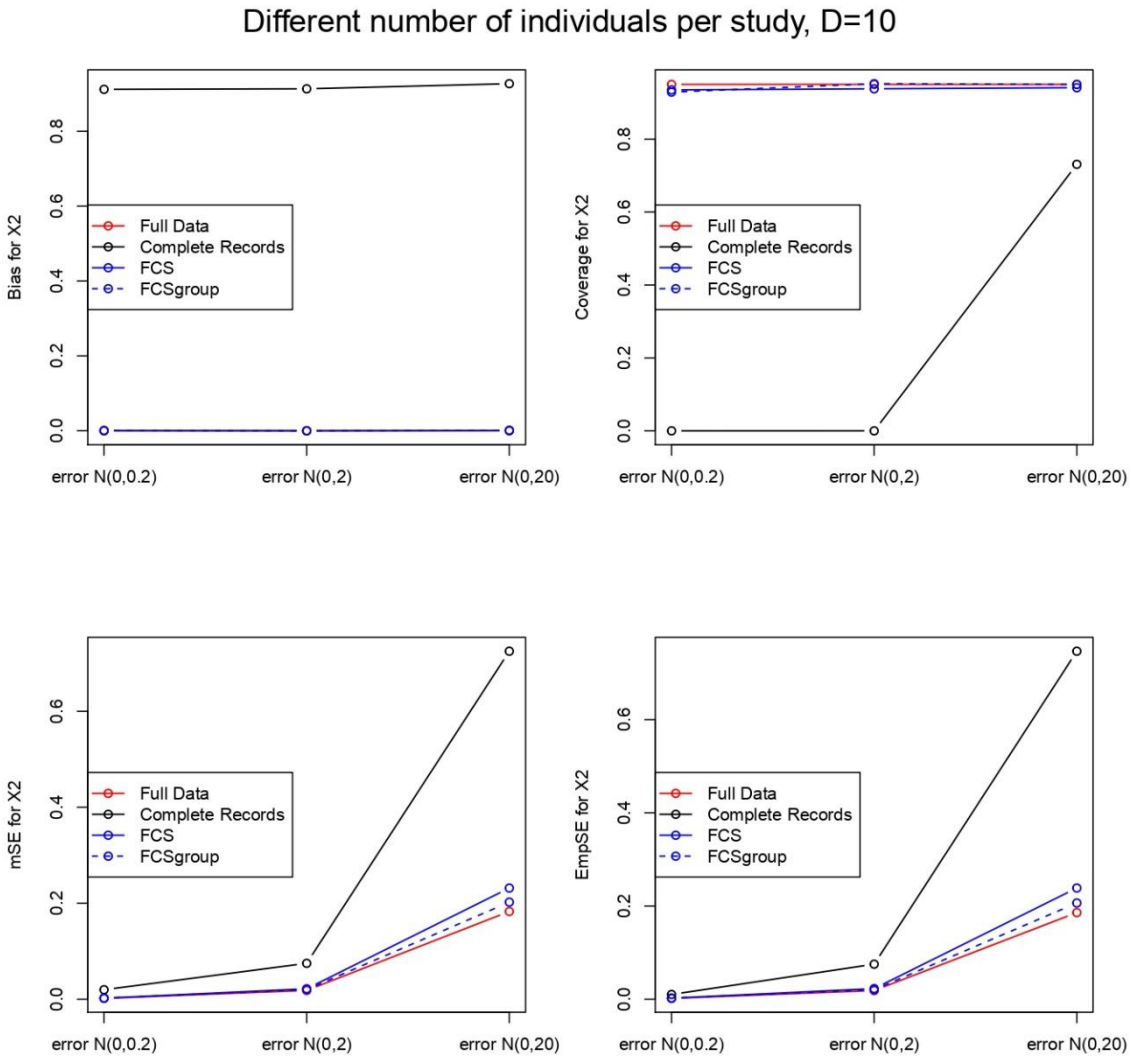


Figure 5.11. Main results from scenario 6's simulation study: Comparison of Bias, Coverage level, mSE and EmpSE for X_2 after 1000 simulations with Full Data (red line), handling mixed type with Complete Records (black line), FCS (blue line), FCSgroup (broken blue line) for three model errors.

FCS: Fully Conditional Specification; *FCSgroup:* FCS imputation model included X'$_2$; *mSE:* mean model Standard Error; *EmpSE:* mean Empirical Standard Error

Overall, both scenarios (figures 5.12 – 5.13) achieved good coverage and no bias for the two probabilistic models, FCS and FCSgroup. When we had 10 datasets with different sizes both imputation models provided unbiased results and FCSgroup had slightly better estimates than FCS. However, in the 5 different size datasets example FCS showed better mean estimate than FCSgroup when model errors were small and large. In figures 5.14 and 5.15 we see mSE and EmpSE for scenarios 5 and 6. We see that as the model error increased the mSE and EmpSE increased as well. mSE and EmpSE are similar per

scenario, and they decreased as the integrated dataset's size increased. In scenarios 5 and 6, all probabilistic models outperformed Complete Records.

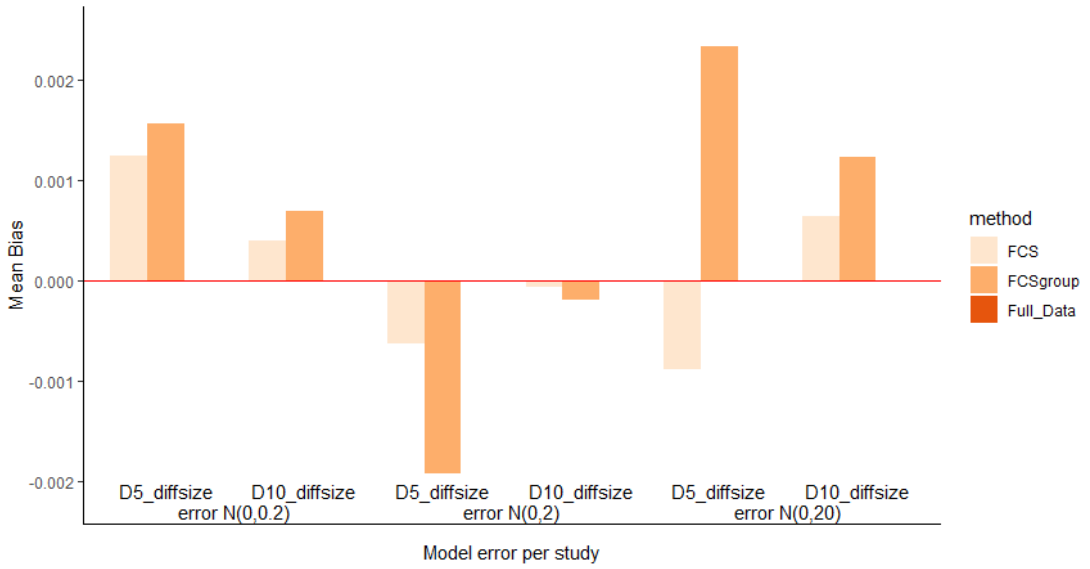


Figure 5.12. Mean bias for X_2 for '5 datasets, N=different per dataset' (D5_diffsize), '10 datasets, N=different per dataset' (D10_diffsize) for the three model errors.

FCS: Multiple imputation by fully conditional specification; *FCSgroup:* FCS imputation model included X'$_2$.

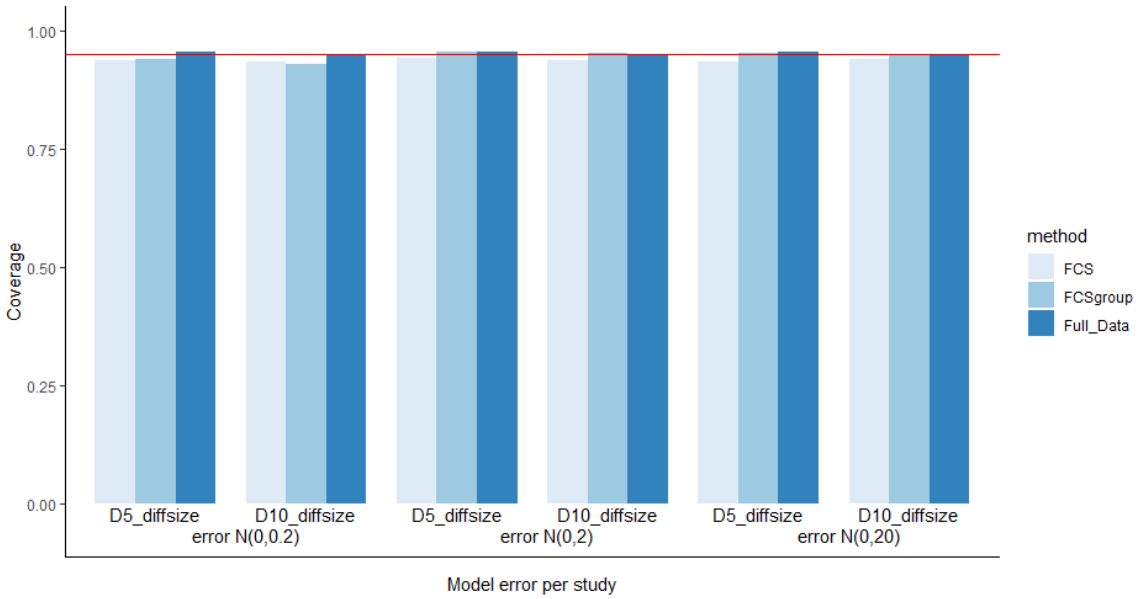


Figure 5.13. Coverage for for X_2 for '5 datasets, N=different per dataset' (D5_diffsize), '10 datasets, N=different per dataset' (D10_diffsize) for the three model errors.

FCS: Multiple imputation by fully conditional specification; *FCSgroup:* FCS imputation model included X'$_2$.

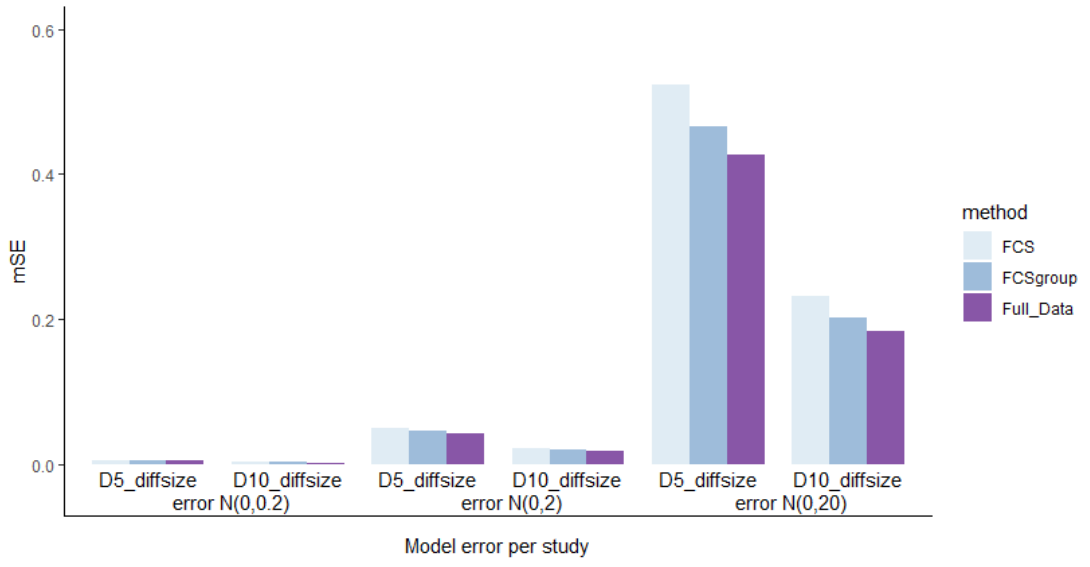


Figure 5.14. mSE for X_2 for '5 datasets, N=different per dataset' (D5_diffsize), '10 datasets, N=different per dataset' (D10_diffsize) for the three model errors.

FCS: Multiple imputation by fully conditional specification; *FCSgroup:* FCS imputation model included X'2; *mSE:* mean model standard error.

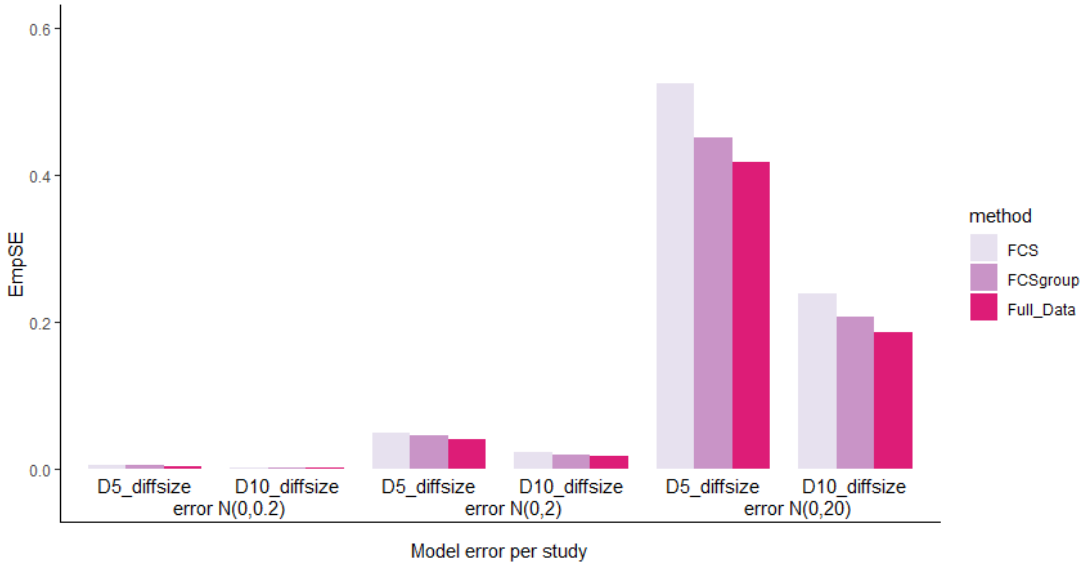


Figure 5.15. EmpSE for X_2 for '5 datasets, N=different per dataset' (D5_diffsize), '10 datasets, N=different per dataset' (D10_diffsize) for the three model errors.

FCS: Multiple imputation by fully conditional specification; *FCSgroup:* FCS imputation model included X'2; *EmpSE:* mean empirical standard error.

Simulations with equal size studies, different model error per study (7 - 10)

Scenarios 7 – 10

Here, we again simulated data from two equal size studies but with different model error per study. In Scenarios 7-10 each dataset had 100, 200, 500 and 1,000 individuals respectively. For each simulated dataset, we applied the mixed type problem to X_2 from D_2. We present the simulation results in figure 5.16.

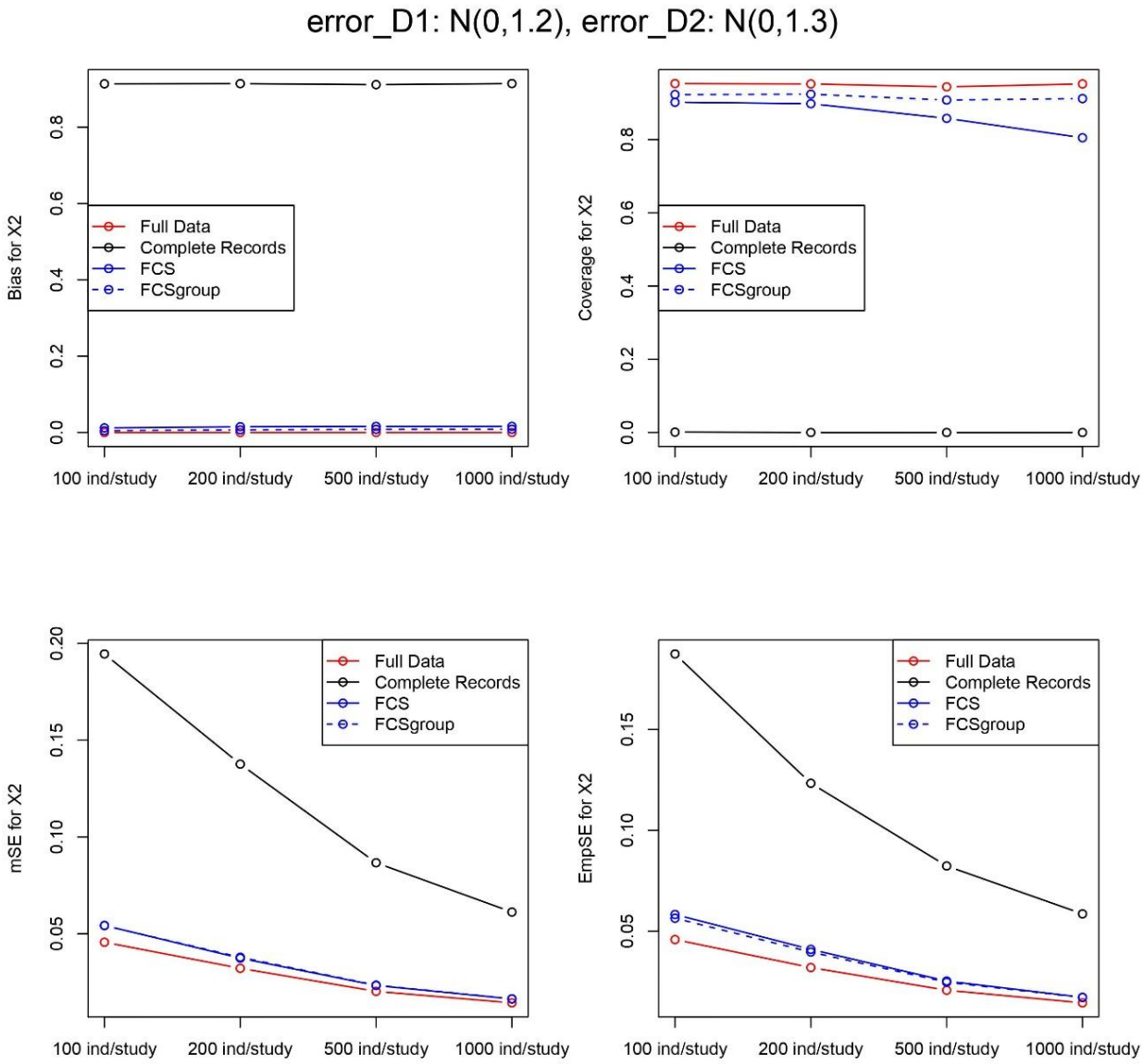


Figure 5.16. Main results from scenario 7-10's simulation study: Comparison of Bias, Coverage level, mSE and EmpSE for X_2 after 1000 simulations with Full Data (red line), handling mixed type with Complete Records (black line), FCS (blue line), FCSgroup (broken blue line) for three model errors.

FCS: Fully Conditional Specification; FCSgroup: FCS imputation model included X'$_2$; mSE: mean model Standard Error; EmpSE: mean Empirical Standard Error

5.4.1 Summary of findings from simulation studies

Bringing all the simulation results together we come to the following conclusions:

Studies with same model errors

We observe that FCS and FCSgroup provided unbiased estimates, almost identical with true Full Data and they solved mixed type successfully. In all scenarios as the model error increased, mSE and EmpSE increase by remaining equal between them.

Studies with different model errors

For probabilistic method FCSgroup had good coverage and no bias for the four study sizes. FCS's coverage started with 90% when we had 100 individuals per study and slowly reduced as the study size increased.

Studies of same sizes

FCS, and FCSgroup's biases remained negligible, and coverage was very good and there was a lot more gain in information compared with a Complete Records analysis.

Studies of different sizes

In both scenarios we see great results in FCS and FCSgroup. Simulation's results show that mSE and EmpSE decreased as the study size increased. The same happens with bias. Coverage was good in both scenarios, and it does not show any particular difference.

Size of model error

Similarly, with chapter 3.4 and 4.4 the smaller the model error per study, the smaller the EmpSE and mSE observed. We also observe that as the model error increased, the difference between EmpSE and mSE slightly increased.

Overall

The probabilistic approaches FCS and FCSgroup are suggested to solve the third type of content heterogeneity – mixed type variable. Results indicate a preference in including the 'group' variable in the imputation model as we have closer to reality mSE and EmpSE. In the scenarios tested we did not see a clear difference between FCS and FCSgroup. Therefore, more complex scenarios may be needed to test that. Most of the simulations' results come to the conclusion that probabilistic approaches are expected to give close to reality results. Our new suggestion about including a categorical informative variable in the imputation model seems to help classification but does not outperform FCS in all

cases. It shows that using the available extra information may help having better estimations and well imputed mixed type variables.

5.5 Application and evaluation – MASTERPLANS exemplar

Here, we illustrated and evaluated the probabilistic integration approach for the imputation of real, rather than simulated data, again comparing the results with those obtained using traditional data integration with FCS and FCSgroup.

5.5.1 Data characteristics and mixed type problem description

To illustrate the developed approaches (Figures 5.1-5.2) we have applied to real-world biomedical and health datasets such as the studies in Lupus. For datasets D_1, D_2, D_3, we have datasets that contain data from ALMS, LUNAR, and EXPLORER respectively. The following analysis used a dataset that included patients in total from all the three studies together, with multiple visits per subject. For the measure of response to treatment and creatinine levels, in order to limit to one per patient, we kept the value recorded at visit that had the least absolute difference from 90 days (3 months). The number of the patients were 597 after that filtering. The integrated dataset consists of the following 17 variables: gender, age, ethnicity, height, weight, BMI, creatinine, current treatment, and various BILAG disease activity scores (Total, Cardiorespiratory, General, Mucocutaneous, Musculoskeletal, Neurological, Renal, Vasculitis, and Haematology). Suppose we are interested in the effect of BMI to Renal response. For this purpose, we wanted to fit a linear regression model with BMI the main interest adjusting for age, ethnicity, creatinine, and gender. So, the linear regression model that answers the research question will look like equation 5.2.

$$Renal\ BILAG\ Score = Age + Ethnicity + Creatinine + Gender + BMI \quad (5.2)$$

Table 5.3 gives us a summary of the baseline characteristics for the 597 patients included in the final analyses. Since these are well-controlled trials, our dataset had no missing values and we were capable to answer the research question (equation 5.2) straight away. Patients' age was obtained as a continuous variable and was complete for all the three studies. We consider this to be an advantage because it allowed us to compare the imputed data against the true, raw data. Similar idea with the real-world example that Quartagno and Carpenter [102] presented in their research paper.

Table 5.3. MASTERPLANS' data characteristics after integrating lupus studies ALMS, LUNAR, EXPLORER: mixed type issue.

Data characteristics	**Integrated dataset N = 597 (%)**	**ALMS N = 248 (41.5)**	**LUNAR N = 138 (23.1)**	**EXPLORER N = 211 (35.4)**
Gender (%)				
Female	526 (88.1)	212 (85.5)	124 (89.9)	190 (90.0)
Male	71 (11.9)	36 (14.5)	14 (10.1)	21 (10.0)
Age, years				
Mean± SD	34.44 ± 11.52	31.5 ± 10.48	30.57 ± 9.34	40.42 ± 11.61
Mix – Max	12.0 – 71.0	12.0 – 64.0	17.0 – 56.0	18.0 - 71.0
Median (IQR)	34.0 (25.0 – 42.0)	31.5 (24.0 – 38.0)	29.0 (23.0 – 36.0)	41.0 (31.5 – 49.5)
Ethnicity (%)				
Black or African American	115 (19.3)	26 (10.5)	37 (26.8)	52 (24.6)
Caucasian	273 (45.7)	108 (43.5)	44 (31.9)	121 (57.3)
Other	209 (35.0)	114 (46.0)	57 (41.3)	38 (18.1)
Height, cm				
Mean ± SD	163.0 ± 9.07	161.6 ± 9.45	163.1 ± 8.81	164.69 ± 8.52
Min – Max	132.0 – 198.1	132.0 – 191.0	143.5 – 195.6	141.0 – 198.1

Median (IQR)	162.5 (157.0 – 168.0)	160.0 (156.0 – 166.0)	162.5 (157.5 – 167.9)	163.8 (159.8 – 170.2)
Weight, kg				
Mean ± SD	70.15 ± 19.53	63.21 ± 15.22	71.06 ± 16.76	77.72 ± 22.63
Min-Max	34.20 – 156.63	34.20 – 121.10	41.77 – 120.31	42.00 – 156.63
Median (IQR)	65.65 (55.84 – 80.54)	61.10 (52.50 – 70.75)	67.50 (58.57 – 82.13)	74.20(60.72 – 90.26)
BILAG score (total)				
Mean ± SD	13.68 ± 8.25	17.71 ± 7.58	10.33 ± 7.22	11.13 ± 7.68
Min – Max	0.0 – 54.0	1.0 – 54.0	0.0 – 52.0	1.0 – 44.0
Median (IQR)	13.0 (7.0 – 18.0)	17.0 (13.0 – 22.0)	8.50 (5.0 – 13.75)	10.0 (1.0 – 15.0)
BMI, kg/m^2				
Mean ± SD	26.28 ± 6.58	24.11 ± 5.01	26.64 ± 5.62	28.60 ± 7.85
Min – Max	13.85 – 56.98	13.85 – 49.13	16.69 – 42.65	16.55 – 56.98
Median (IQR)	24.69 (21.68 – 29.41)	22.96 (20.57 – 26.71)	25.16 (22.52 – 29.77)	26.81 (23.07 – 33.05)
Treatment				
Placebo + AZA OR	149 (25.0)	122 (49.2)		27 (12.8)

Placebo + MMF OR MMF	220 (36.9)	126 (50.8)	68 (49.3)	26 (12.3)
Placebo + MTX OR MTX	16 (2.6)			16 (7.6)
RITUX + AZA	45 (7.5)			45 (21.3)
RITUX + MMF	127 (21.3)		70 (50.7)	57 (27.0)
RITUX + MTX	40 (6.7)			40 (19.0)
Creatinine, mg				
Mean ± SD	0.89 ± 0.44	0.91 ± 0.53	0.97 ± 0.52	0.80 ± 0.21
Min – Max	0.31 – 5.27	0.34 – 5.27	0.30 – 3.50	0.40 – 1.70
Median (IQR)	0.8 (0.63 – 0.97)	0.8 (0.62 – 1.0)	0.8 (0.7 – 1.1)	0.8 (0.7 – 0.9)
Cardiorespiratory score				
Mean ± SD	0.46 ± 1.46	0.36 ± 1.34	0.23 ± 1.14	0.74 ± 1.72
Min – Max	0.0 – 12.0	0.0 – 12.0	0.0 – 12.0	0.0 – 12.0
Median (IQR)	0.0 (0.0 – 0.0)	0.0 (0.0 – 0.0)	0.0 (0.0 – 0.0)	0.0 (0.00 – 1.0)
General score				
Mean ± SD	1.11 ± 1.91	1.05 ± 2.19	0.81 ± 1.46	1.37 ± 1.79
Min – Max	0.0 – 12.0	0.0 – 12.0	0.0 – 12.0	0.0 – 12.0
Median (IQR)	1.0 (0.0 – 1.0)	0.0 (0.0 – 1.0)	1.00 (0.0 – 1.0)	1.0 (1.0 – 1.0)

Mucocutaneous score				
Mean ± SD	2.02 ± 2.85	1.72 ± 2.74	1.20 ± 1.84	2.89 ± 3.27
Min – Max	0.0 – 12.0	0.0 – 12.0	0.0 – 5.0	0.0 – 12.0
Median (IQR)	2.02 (0.0 – 5.0)	0.0 (0.0 – 5.0)	0.0 (0.0 – 1.0)	1.0 (1.0 – 5.0)
Musculoskeletal score				
Mean ± SD	1.65 ± 2.83	0.89 ± 1.75	0.93 ± 2.29	3.02 ± 3.58
Min – Max	0.0 –12.0	0.0 – 12.0	0.0 – 12.0	0.0 – 12.0
Median (IQR)	1.0 (0.0 –1.0)	0.0 (0.0 – 1.0)	0.0 (0.0 – 1.0)	1.0 (0.0 – 5.0)
Neurological Score				
Mean ± SD	0.29 ± 1.1	0.07 ± 0.48	0.28 ± 0.89	0.56 ± 1.6
Min – Max	0.0 –12.0	0.0 – 5.0	0.0 – 5.0	0.0 – 12.0
Median (IQR)	0.0 (0.0 – 0.0)	0.0 (0.0 – 0.0)	0.0 (0.0 – 0.0)	0.0 (0.0 – 0.0)
Renal Score				
Mean ± SD	5.84 ± 5.37	11.1 ± 2.60	4.68 ± 3.51	0.42 ± 1.7
Min – Max	0.0 – 12.0	1.0 – 12.0	0.0 – 12.0	0.0 – 12.0
Median (IQR)	5.0 (0.01 – 2.0)	12.0 (12.0 –12.0)	5.0 (1.0 – 12.0)	0.0 (0.0 – 0.0)
Vasculitis Score				

Mean ± SD	0.59 ± 1.52	0.37 ± 1.26	0.44 ± 1.06	0.94 ± 1.95
Min – Max	0.0 –12.0	0.0 – 12.0	0.0 – 5.0	0.0 – 5.0
Median (IQR)	0.0 (0.0 –1.0)	0.0 (0.0 – 0.0)	0.0 (0.0 – 1.0)	0.0 (0.0 – 1.0)
Haematology Score				
Mean ± SD	1.72 ± 2.24	2.15 ± 2.61	1.75 ± 2.28	1.19 ± 1.53
Min – Max	0.0 – 12.0	0.0 – 12.0	0.0 – 12.0	0.0 – 5.0
Median (IQR)	1.72 (0.0 – 5.0)	1.0 (0.0 – 5.0)	1.0 (0.0 – 5.0)	1.0 (0.0 – 1.0)

5.5.2 Probabilistic Data Integration – mixed type problem

In order to illustrate and evaluate the imputation methods as a tool to answer a research question (equation 5.2), after data integration, we sampled with replacement from the original data 1000 datasets with the same sample size and in each of these we introduced the mixed type problem to *Age* variable. More specifically about the mixed type problem, we chose one of the lupus studies and specifically LUNAR to have *Age* as a categorical variable (‘0-20’, ‘21-40’, ‘41-60’, ‘>60’) and the other two studies, ALMS and EXPLORER, to have *Age* as an integer. Therefore, we had an example of mixed type problem for patients in the integrated dataset. We then created a categorical informative age_{group} variable with four levels ‘0-20’, ‘21-40’, ‘41-60’, ‘>60’ and assigned to individuals based on their true age. We then removed patients’ *Age* records from LUNAR study, so data missingness was introduced. We knew patients’ actual age and therefore agegroup for ALMS and EXPLORER, but we only knew the agegroup for LUNAR’s patients.

Our goal was to answer our predefined research question (equation 5.2), so we needed to apply the suggested probabilistic data integration approaches that take advantage of available information. We kept the largest number of levels in Age which was the continuous variable. Therefore, in the integrated dataset we had missing data for the Age variable and would impute them based on some available information using FCS and FCSgroup. In both imputation models the following variables were used as predictors:

gender, age, ethnicity, BMI, creatinine, current treatment, BILAG disease activity scores (Cardiorespiratory, General, Mucocutaneous, Musculoskeletal, Neurological, Renal, Vasculitis, and Haematology). We set seed to 547683. BMI and total BILAG score variables in the MASTERPLANS dataset were combinations or recoded versions of other data, so were not included in the imputation models [138]. We used predictive mean matching as method, a different number of imputations (5, 10, 15, 20), and iteration number was set to 10. At the end, we fit linear regression models to the complete datasets that resulted from all multiple imputation methods.

Figures 5.17 and 5.18 show estimate and coverage level for the set of fixed effect Age parameter estimate from Full Data analysis and handling missing data with the two imputation models, FCS and FCSgroup. Figure 5.17 compares coverage levels for the set effect parameter estimated from Full Data analysis, and handling missing data with either FCS or FCSgroup. Coverage levels for all imputation methods were close to 95%, although on average FCSgroup was better and similar to the true Full Data. Again, the results showed very small bias in estimates for Age in FCS. In particular, estimates from FCSgroup were the most unbiased and identical to the true model (Full Data sample). The increasing number of imputed datasets indicates that it helped achieving a better coverage but with similar estimates. All imputation methods could be used to solve mixed type problem and we would have a clear preference in using the agegroup variable in the imputation model to achieve close to reality results. Hence, a comparison between the imputation strategies is consistent with the results from the simulation studies and confirms that FCSgroup specifically performed very similarly in real world data and when there was a larger number of variables. Application in real-world data shows a practically useful gain in information over Complete Records with multiple imputation using FCS and FCSgroup.

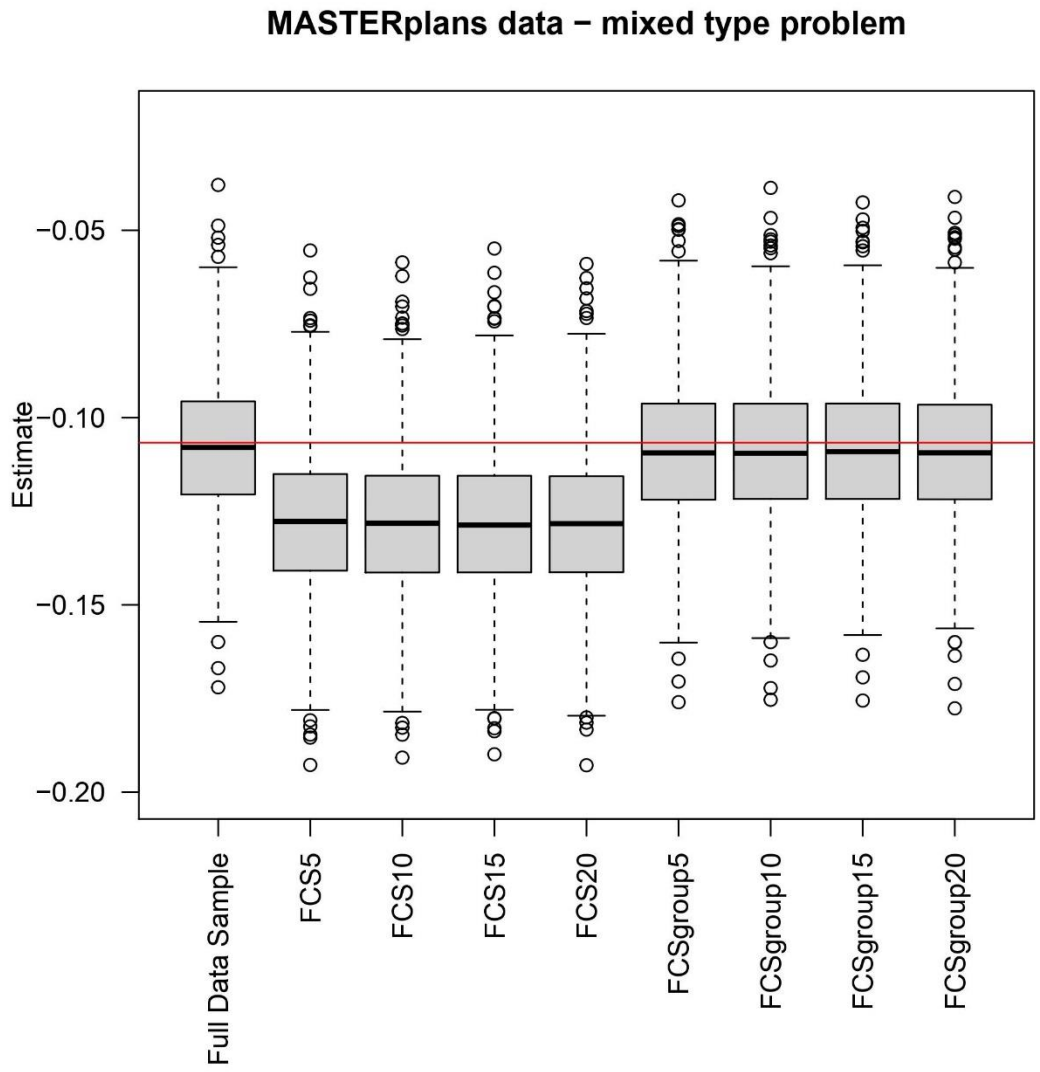


Figure 5.17. Results of the resampling study with the MASTERPLANS Data: boxplot of Age estimate in the linear model.

FCS: Fully Conditional Specification; *FCSgroup:* Fully Conditional Specification including informative 'group' variable. The numbers (5, 10, 15, 20) after the imputation method indicate the number of imputations (5, 10, 15, 20) - iterations are fixed to 10.

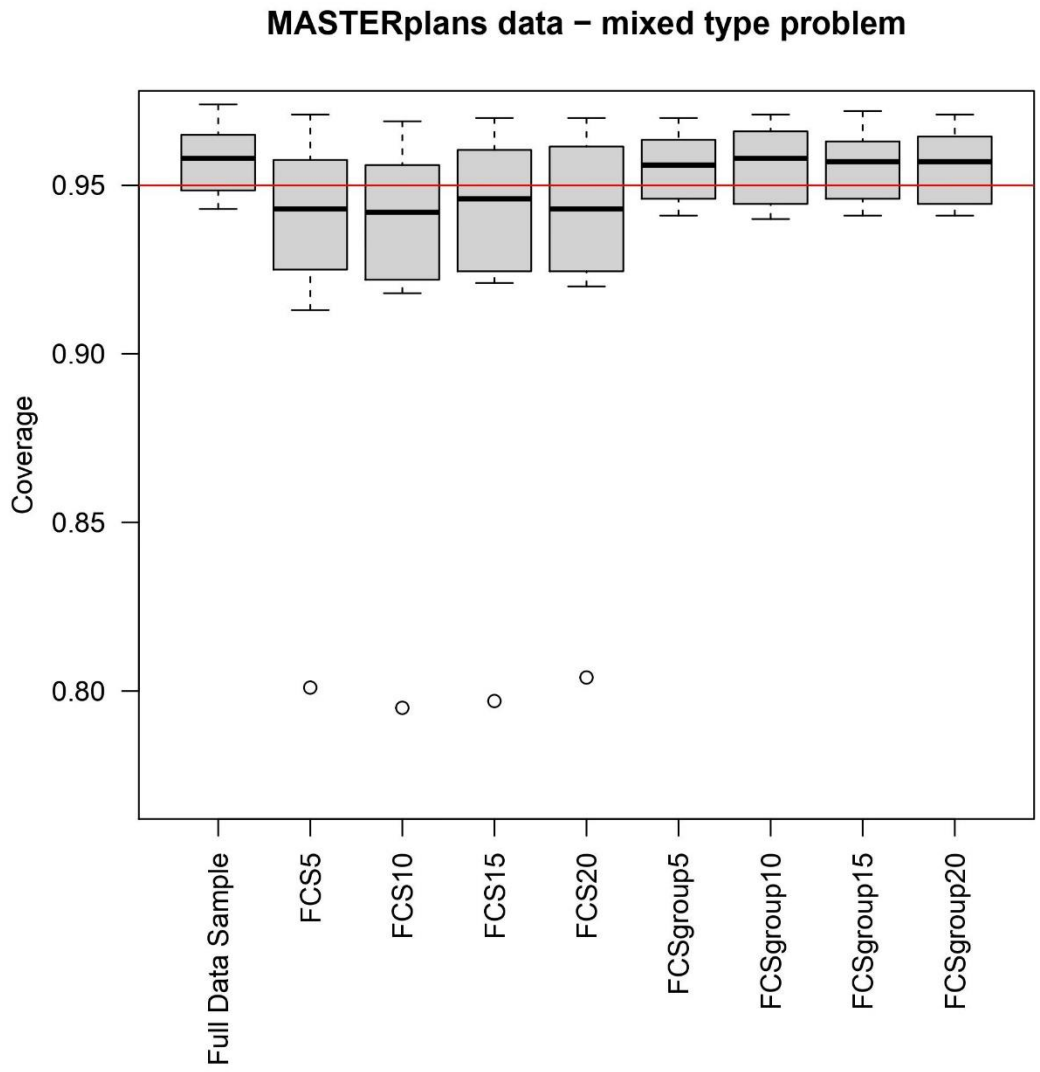


Figure 5.18. Results of the resampling study with the MASTERPLANS Data: boxplot of Age coverage level in the linear model.

FCS: Fully Conditional Specification; *FCSgroup:* Fully Conditional Specification including informative 'group' variable. The numbers (5, 10, 15, 20) after the imputation method indicate the number of imputations (5, 10, 15, 20) - iterations are fixed to 10.

Note: We also ran analyses using true SLE data without performing simulations through resampling method - only using the initial complete integrated dataset. In Appendix C, we can find tables which present coefficients (estimate, standard error, t statistic and p-values) for linear regression model from equation 5.2 in true data, after applying Complete Records analysis, FCS (5 imputations, 10 iterations) and FCSgroup (5 imputations, 10 iterations) in SLE data.

5.6 Discussion

Summary of main findings

Here we describe a new probabilistic approach for tackling mixed-type variable problem after data integration, from both a theoretical and a practical perspective. Our initial assessment revealed very promising results, and we see that our probabilistic method offers accurate results, and imputed data values that are close to the real observed data. Based on this, we would argue that the mixed type variable problem could be resolved to produce homogeneous type variable, without removing information; and multiple imputation could provide 'close to reality' results that will enable us to answer research question(s).

By default, a harmonised type of a variable could be produced, for example, if age captured as a continuous variable in one study is transformed into a categorical variable to match other datasets where it was captured as a category [139]. Our simulations using artificially and real-world data show that this traditional approach should not be recommended as large bias was introduced in the fitted models. Our alternative approach relies on probabilistic methodologies. Multiple imputation has also been suggested and evaluated as a method for individual patient data meta-analysis in cases of complete missing variables and heterogeneity among studies [97], [121].

In the simulation studies of methods for handling missing data, in the form of mixed type problem after integrating data from different biomedical data sources, most parametric imputations produced estimates with no material bias for a linear model in data artificially introduced MCAR missingness. FCS and FCSgroup produced unbiased estimates and almost 95% confidence intervals in most cases. FCS, in very few cases - where we had different model errors between studies - the coverage probability was a bit smaller than 95% suggesting than confidence intervals may have been conservative. Overall, our results suggest that a probabilistic method such as multiple imputation by chained equations methods is useful for imputing complex biomedical datasets in which there is a mixed type problem in the common after integration and especially for linear regression models. Probabilistic approaches have also been suggested and experimentally evaluated in real-world data (MASTERPLANS) in previously described problems of content heterogeneity i.e., missing variables (Chapter 3) and granularity (Chapter 4) [140]. To our knowledge, no published study to date has considered imputation as a method to solve data heterogeneity among studies when a variable presented in mixed-type i.e.,

categorical and integer. In addition, this study suggests an original idea that takes advantage of the current traditional solution which includes as a step in the imputation models in order to achieve better coefficients. Our goal is to take advantage of the given information concerning subgroups and levels in mixed type variables in multiple imputation. The suggested probabilistic approaches give mostly valid results across a range of model errors and always outperform the traditional data integration approach.

Imputation method including the extra informative variable (FCSgroup)

The simulation studies showed that the usage of the extra informative variable in the imputation model improved the imputation itself and improved data classification. FCSgroup showed accuracy and in most cases outperformed standard FCS. In MASTERPLANS exemplar, FCSgroup advantage over FCS was clear.

Limitations

Similarly with chapters 3 and 4 this study has advantages, and it is based on an everyday problem in data science and statistical analysis. The approaches that we have presented here were extensively evaluated and were further applied to real-world datasets. The application of probabilistic approaches to solve data integration's mixed type problem comprised a number of critical choices (such as the choice of predictive modelling algorithms [141], number of simulated datasets, number of imputations) that requires thorough methodological investigation. It also relies on assumptions (such as reasons for missingness) that can influence the end result and should therefore be further investigated through sensitivity analyses when it is applied in practice. Nonetheless, the suggested methods were tested in many different scenarios with artificial data and real-world data in which many variables were present and three different cohort studies were integrated.

Conclusions and further development

Existing methods for dataset integration lean on mapping to common data models, often resulting in a significant loss of information that occurs in the source datasets. Suggested traditional solutions that solve mixed numeric and non-numeric data types after data integration should be properly evaluated. This study offers the idea how the current universally accepted solution may not be the best. Ideally, shared data models would be implemented at source, enabling uniform data collection at different sites and studies. But data standardisation is always incomplete, and our approach grasps this weakness instead of suffocating it. Our data integration solution is based on probabilistic methodologies. This chapter included evaluation of the general applicability of the method and compared

results of the proposed integration techniques with gold standard results through statistical simulation studies. It has also illustrated this approach using a real-world example from lupus cohort studies. In summary, the results of our comprehensive set of simulation studies show that researchers can use FCSgroup and FCS, implemented in mice R package, for imputation to solve mixed type problem with confidence. They can also choose to apply FCS in case the integrated dataset is already very large and complicated and if by adding some extra steps and variables in imputation gives big delays in data analysis. Finally, our general suggestion is the inclusion of the informative variable in the imputation model, especially when computational time allows it, as it helped the algorithm achieve better results and estimates, and smaller standard errors.

Chapter 6: Combined types of content heterogeneity

6.1 Introduction

In previous chapters we investigated four different types of content heterogeneity problems that can occur in integrated datasets. The objective of this chapter is to address the combination of content heterogeneity problems. The motivation is that in reality we see that content heterogeneity problems often do not present in isolation but appear alongside each other when we stack multiple datasets. Therefore, we have to be able to address combinations of content heterogeneity problems that co-exist within a single, stacked dataset. Here, we focus on combinations of the problem of varying granularity (previously addressed in Chapter 4) and the problem of mixed numeric and categorical data types (previously addressed in Chapter 5).

As before we assume that we have multiple study datasets and want to answer research question(s) by performing regression analysis on the stacked combination of these study datasets. We assume again that the datasets are non-overlapping and that each dataset's observations are independent and identically distributed. Similar to previous chapters, in section 6.2) we explore the usage of probabilistic approaches to solve the combination of content heterogeneity problems (depicted in Figure 6.1 below), and compare them with traditional approaches that use a common data model consisting only of variables that are present in all datasets with similar granularity and as the same data type.

We investigate the combination of the two problems through simulation studies (Section 6.3) and through application of the method in real-world data (Section 6.4). Finally, we reflect on the utility of the methods to address this combination of problems, their limitations, and discuss potential further research (Section 6.5).

6.2 Probabilistic methods to solve combined content heterogeneity types

Similarly, to previous chapters we assume that the variables in question are relevant to the research question that we want to answer. In figure 6.1 we see the main tasks to solve two content heterogeneity problems probabilistically. The figure is similar to figure 4.1. After data integration, we select variables that will be included in the regression model. In the mixed type problem, we assume that there is at least one numeric representation and at least one other study with a categorical representation. In the granularity problem, we assume that at least one study has less granular categorical representation than the other studies. We also assume that the study with less granular categorical representation

does not have numeric representation, so data missingness (black squares) due to granularity and mixed type are not present in the same study.

Another difference with figure 4.1 is that due to the parallel existence of both content heterogeneity problems, we add one 'group' categorical informative variable for each problem respectively (figure 6.1 – light blue column for granularity and dark blue column for mixed type). Afterwards, to solve data missingness due to combined content heterogeneity problems in D_0 we apply multiple imputation (figure 6.1 – red arrows). We also want to evaluate our idea about the inclusion of 'group' variable(s). To evaluate the impact of the 'artificial cluster' within the imputation model we used four imputation models, either excluding or including the 'group' variables. It means that they were added as predictor(s) in the imputation models. In chapters 4 and 5 we showed that FCS and FCSgroup offered very good results. In more detail, in section 4.3, we saw that FCS and FCSgroup were almost identical whereas in section 5.3, FCSgroup outperformed FCS in some cases.

Thus, this chapter explores situations where two content heterogeneity probles are solved through multiple imputation simultaneously. In order to explore the practical consequences of imputation model differences, missing data are imputed using the following four imputation models:

- **FCS**, multiple imputation by chained equations – both categorical informative variables are excluded from imputation models, so they are not used as predictors during imputation modelling;
- **FCSgroup**, multiple imputation by chained equations - inclusion of relevant categorical informative variable in the relevant imputation model. In other words, each 'group' variable is used as predictor during imputation modelling only for the relevant content heterogeneity problem;
- **FCSgroups**, multiple imputation by chained equations - inclusion of both categorical informative variables in both imputation models. Therefore, both 'group' variables are used as predictors during imputation modelling for both content heterogeneity problems;
- **FCS3group2**, multiple imputation by chained equations – FCS imputation model is used in solving granularity problem, FCSgroup imputation model is used in solving mixed type problem.

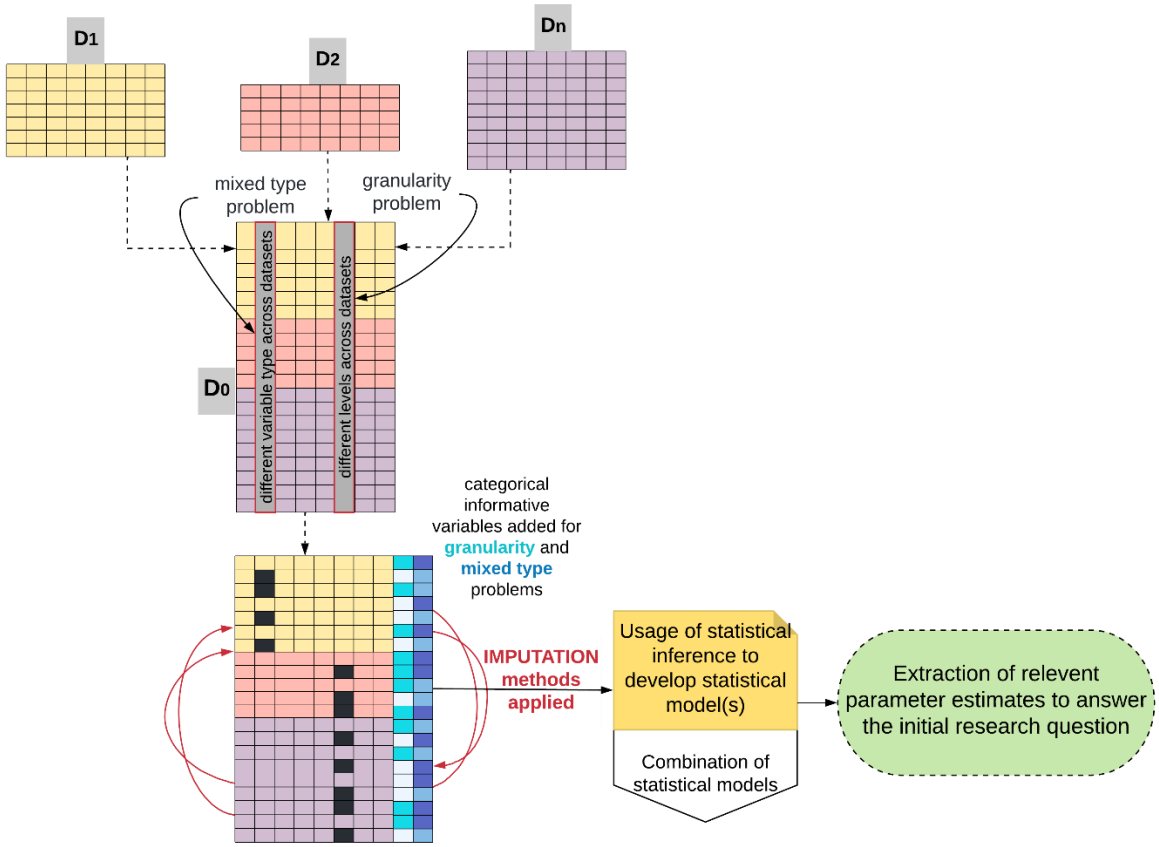


Figure 6.1. Main tasks of our probabilistic data integration processes to solve combined content heterogeneity problems.

6.3 Simulation studies

The methods proposed in section 6.2 suggest that using multiple imputation procedures and including/excluding categorical informative variables. We therefore conducted simulation studies aiming to explore and compare the performance of methods in a probabilistic procedure to solve combined types of content heterogeneity and explore whether the inclusion of 'group' variables improved imputation models. As mentioned in chapter 3.4.3, all performance measures were calculated between the estimates coefficients of each (probabilistic/traditional) integration model and the generating coefficients and then averaged across iterations. In this chapter, estimand θ was the estimate $\hat{\theta}_i$ of X_2 and X_3 coefficients in each model fit.

6.3.1 Simulation design and scenarios

The process of data simulation was similar to previous chapters, except for the simulation data that introduced content heterogeneities, betas used for generating outcome variable and starting seed number. We chose 240621 as a starting seed in the random number generator. We see a pictorial representation of the simulation procedure for the combined problems in figure 6.3. We simulated, two continuous variables with normal distribution

X_1, X_2, and one categorical X_3 with four levels ('A', 'B', 'C', 'D') as in previous chapters. We used the same data generating mechanism as in chapters 3 to 5 (equations 3.1, 4.1,5.1) to simulate outcome Y but with different betas than previous chapters. In particular, we have $\beta_0 = 1.4883$, $\beta_1 = -1.7553$, $\beta_2 = 0.4236$, $\beta_3 = 1.4656$, $\beta_4 = -0.9567$, $\beta_5 = 0.4977$. Y was complete, had no missing data and had known dependency on X variables. After joining the study datasets and fitting a linear regression model (true Full Data) to the combined dataset D_0:

1) We introduced content heterogeneity as mixed type problem as follows. We replaced X_2 with two levels LOW and HIGH to D_0 for some individuals, when $X_2<T$, then X_2 = LOW, and X_2 = HIGH otherwise, for a given threshold T. In this case, T was mean value. For example, for the last 1000 individuals in D_0 (i.e., the individuals from dataset D_2), we set X_2 to missing. Now, X_2 had systematically (MCAR) missing values for observations from D_2 but X'_2 had values across the entire combined dataset. It corresponds to the situation that we would have in practice, because the categorical values can always be derived from the numeric values but not vice versa.
2) We added a categorical informative variable X'_2 in D_0 with two levels LOW and HIGH, where X'_2=LOW when $X_2<T$, and X'_2=HIGH otherwise, for a given threshold T. In this case, T was mean value.
3) We added content heterogeneity as granularity as follows. We replaced values in X_3 for a specific study, in D_0. For example, the first 1000 individuals in D_0 (i.e., the individuals from dataset D_1), whenever X_3 was equal to 'A' or 'B', we replaced it by a new value, 'AB'. Thus, the levels in X_3 were inconsistent across D_0 and therefore granularity problem was introduced. In particular, the first 1000 individuals in D_0 (individuals from D_1) had the following values in X_3: 'AB', 'C', 'D' and the last 1000 individuals in D_0 (individuals from D_2) had the following values in X_3: 'A', 'B', 'C', 'D'.
4) We added a categorical informative variable X'_3 in D_0 with two levels '0', '1', '2', where X'_3 = '0' when X_3 = 'A' or X_3 = 'B' or X_3 = 'AB', X'_3 = '1' when X_3 = 'C' and X'_3 = '1' when X_3 = 'D'.

Afterwards, as we see in figures 6.1 - 6.2 we analysed 1,000 simulated datasets, solving missingness using the traditional data integration approach (Complete Records), and probabilistic approaches i.e., **parametric FCS** (excluded X'_2 and X'_3 as predictors from imputation), **FCSgroup** (FCS included as predictor X'_2 when imputed X_2 and included X'_3 as predictor when imputed X_3), **FCSgroups** (included both X'_2 and X'_3 as predictors

when imputed X_2 and X_3) and **FCS2group3** (excluded X'_2 as predictor when imputed X_2 and included X'_3 as predictor when imputed X_3). In all of the examined scenarios (tables 6.1 and 6.2), missing data were present in X_2 and X_3. In every case, the analysis model was the linear regression model ($Y \sim X_1 + X_2 + X_3$). We imputed continuous variable X_2 applying FCS, FCSgroup, FCSgroups with *predictive mean matching* (pmm) technique. We imputed categorical variable X_3 applying FCS, FCSgroup, and FCSgroups with *unordered data by polytomous regression* (polyreg) technique. We used 240621 as starting seed number to generate sequences of random numbers. Regardless of the imputation model, all complete datasets were analysed using the same multivariable linear regression model $Y \sim X_1 + X_{2,,} + X_{3,,}$ where $X_{2,,}$ and $X_{3,,}$ were the imputed data. The data were imputed either with m =5 or 10 times and it= 5 or 10 iterations of the chained equations algorithms. All imputation analyses have been performed with R package mice freely available on CRAN [98].

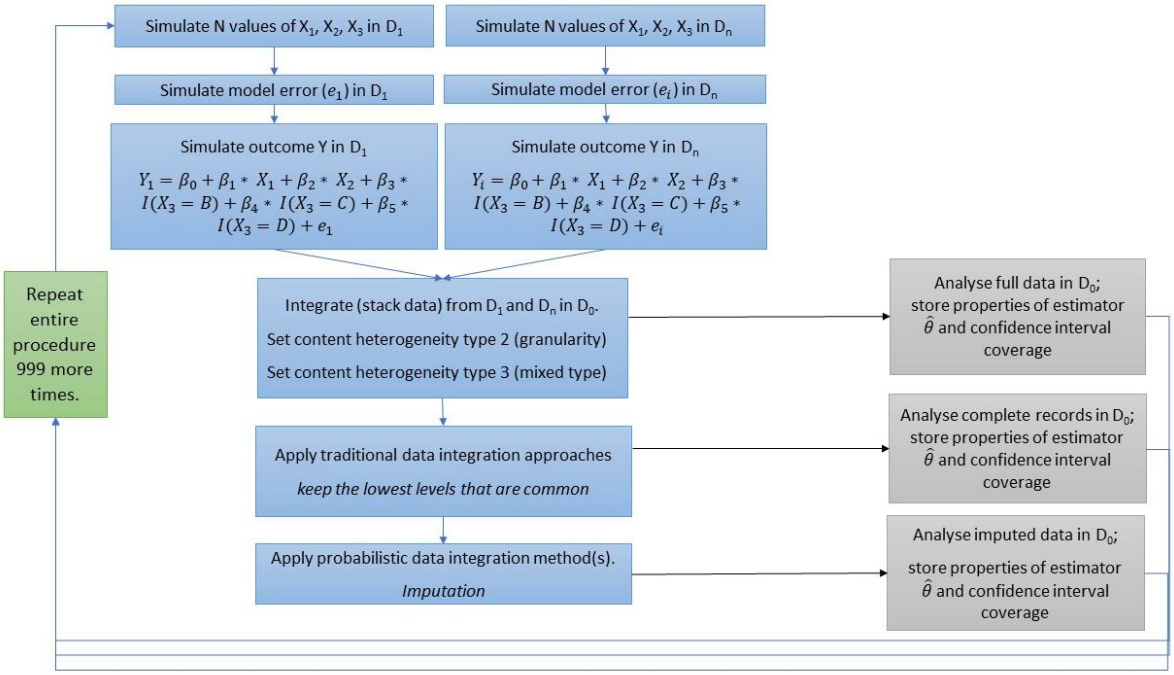


Figure 6.2. Combined content heterogeneity problems' simulation procedure: A pictorial representation of the simulation procedure for the combined content hterogeneoty problems.

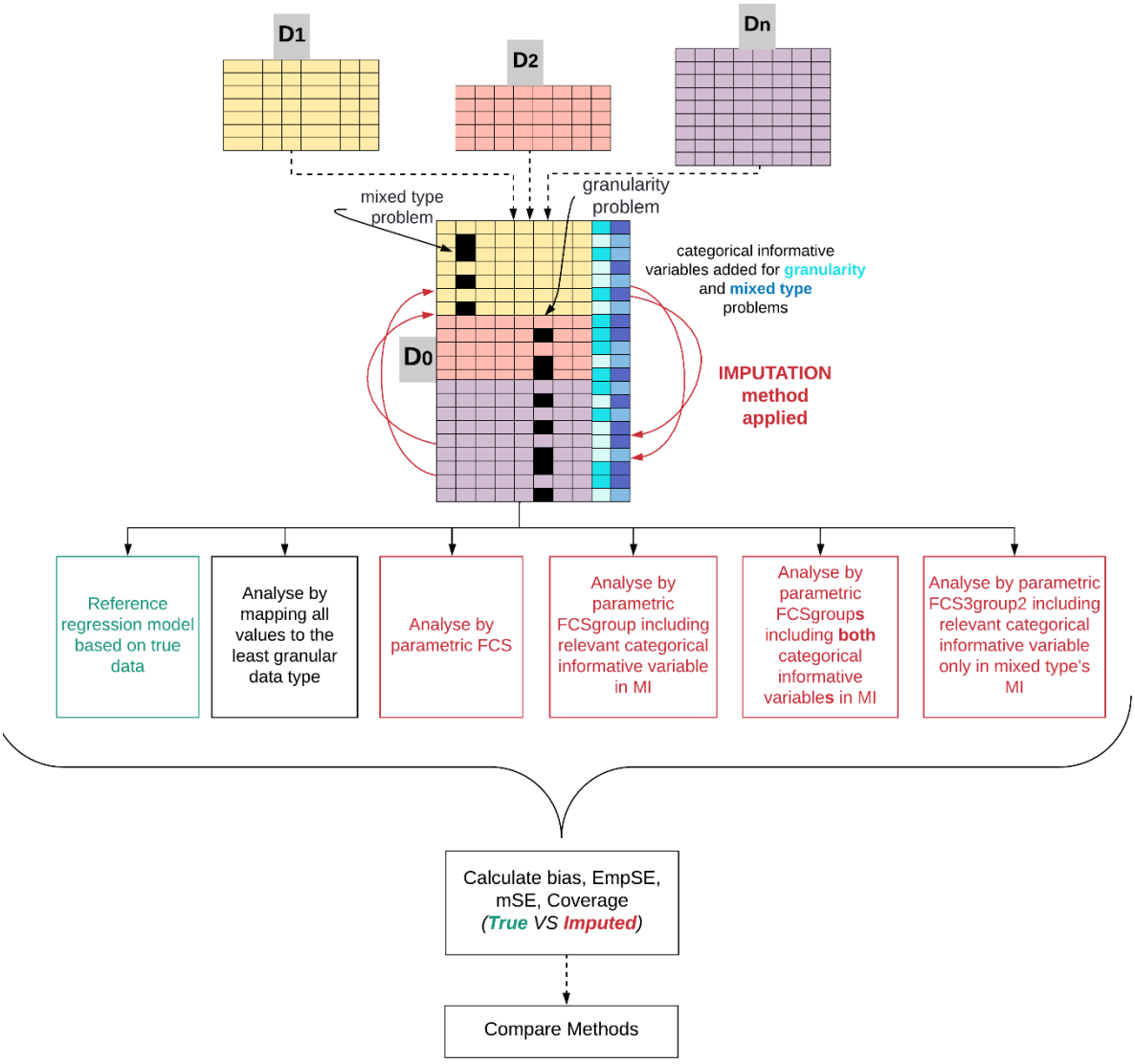


Figure 6.3. The flow diagram for the simulation process to solve granularity and mixed type problems after structured data integration in healthcare.

During different simulation scenarios (Tables 6.1 and 6.2), we investigated how the different level of model errors (small-medium-large) per study affects the parameter estimates. We also investigated how the number of imputed datasets, the number of iterations (when needed), number of individuals and different model error per study affected FCS, FCSgroup, FCSgroups and FCS2group3.

Table 6.1. Scenarios 1 - 5 used to generate data from Figure 6.2.

	Scenario 1	**Scenario 2**	**Scenario 3**	**Scenario 4**	**Scenario 5**
Number of individuals per study (N)	200	1000	200	1000	D_1: 200, D_2:150, D_3:50, D_4:75, D_5:100
Model error e_i: (same for each study)	N~(0,0.2) N~(0,2) N~(0,20)	N~(0,0.2) N~(0,2) N~(0,20)	N~(0,0.2) N~(0,2) N~(0,20)	N~(0,0.2) N~(0,2) N~(0,20)	N~(0,0.2) N~(0,2) N~(0,20)
Number of studies (D)	2	2	5	5	5
Imputations (m)	10	10	5	5	5
Iterations (it)	10	10	5	5	5
Missingness applied to:	X_2 from D_2 X_3 from D_1	X_2 from D_2 X_3 from D_1	X_2 from D_1, D_4 X_3 from D_2, D_5	X_2 from D_1, D_4 X_3 from D_2, D_5	X_2 from D_5 X_3 from D_4

Table 6.2. Scenarios 6 - 10 used to generate data from Figure 6.2.

	Scenario 6	**Scenario 7**	**Scenario 8**	**Scenario 9**	**Scenario 10**
Number of individuals per study (N)	D_1:800, D_2:150, D_3:50, D_4:75, D_5:350, D_6:200, D_7:150, D_8:500, D_9:750, D_{10}:100	100	200	500	1000
Model error e_i: (same for each study)	N~(0,0.2) N~(0,2) N~(0,20)	e_1: N~(0,1.2), e_2: N~(0,1.3)	e_1: N~(0,1.2), e_2: N~(0,1.3)	e_1: N~(0,1.2), e_2: N~(0,1.3)	e_1: N~(0,1.2), e_2: N~(0,1.3)
Number of studies (D)	10	2	2	2	2

Imputations (m)	5	10	10	10	10
Iterations (it)	5	10	10	10	10
Missingness applied to:	X_2 from D_3, D_9 X_3 from D_6, D_{10}	X_2 from D_1 X_3 from D_2	X_2 from D_1 X_3 from D_2	X_2 from D_1 X_3 from D_2	X_2 from D_1 X_3 from D_2

6.3.2 Results

In this section we present the results of the series of simulation studies designed to evaluate the use of different multiple imputation models. Similarly with the previous experimental chapters we performed scenarios, similar analyses, and results assessment. The data were analysed before content heterogeneities' application as a benchmark for the multiple imputation procedure, and after mapping to common levels using complete case analysis (Complete Records). After imputation, the estimates were combined using Rubin's rules. As in previous chapter, the performance of each method was assessed by computing the empirical mean of the parameter estimates, root mean square estimated standard error (mSE), empirical Monte Carlo standard error (EmpSE), and the coverage of nominal 95% confidence intervals (Cov).

All imputation models included most variables (mentioned in section 6.2) in the analysis model as predictors. Hence, the imputation methods were considered very compatible with the analysis model. Concerning graphical representation of results, we are aware that ideally the information included in figures 6.4 to 6.18 could also include regression estimates and performance measures for all X_3's categories. Therefore, in order to provide the reader with a more gentle introduction to the main findings of the study, we decided to include a graphical representation of the main results and focus on X_2 and $X_{3=B}$. Appendix D provides all the results obtained from simulation studies for all scenarios in tables, and the comparing method complete case analysis. They also include results for some additional cases of scenarios that had fewer number of imputations and iterations.

Simulations with studies of same sizes (Scenarios 1 - 4)

Scenario 1

In this scenario, we simulated data from two studies, **each with 200 individuals.** For each simulated dataset, we applied mixed type problem in X_2 (missingness in D_2) and granularity problem in X_3 (missingness in D_1). We present the simulation results in figure 6.4 with e_i:

(N ~ (0, 0.2)), (N ~ (0, 2)), (N ~ (0, 20)) respectively. We show results when multiple imputation models had 10 imputations and 10 iterations each. Analyses of datasets imputed gave very good results both in terms of bias (figure 6.4), precision (figures 6.10 and 6.11) and confidence interval coverage (figure 6.4). When simulating model error e_i: N ~ (0, 0.2), X_3 estimates were biased for FCS which resulted in low coverage. FCSgroups may have needed more imputations and iterations to achieve better convergence, when simulating model error e_i: N ~ (0, 20) in X_2. In general, FCSgroup and FCSgroups and FCS3group2 were almost compatible models with Full Data. Almost all imputation models offered really close to reality results and better estimates than Complete Records.

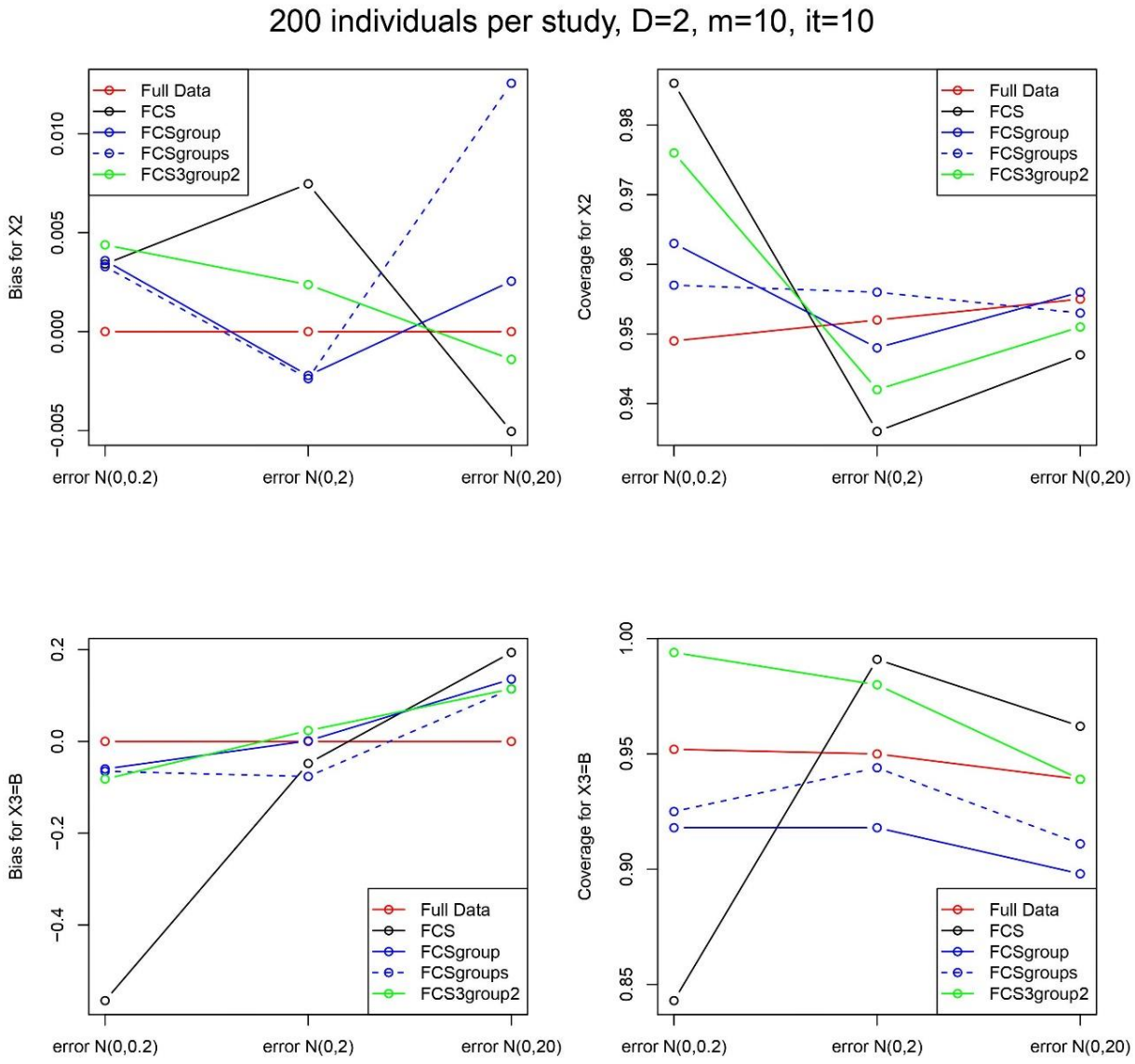


Figure 6.4. Main results from scenario 1's simulation study: Comparison of Bias and Coverage level, for $X_{3=B}$ and X_2 after 1000 simulations with true Full Data (red line), handling granularity and mixed type with FCS (black line), FCSgroup (blue line), FCSgroups (broken blue line) and FCS3group2 (green line), for three model errors.

FCS: both imputation models excluded categorical informative variables X'_2 and X'_3; *FCSgroup*: imputation model for X_2 included X'_2 and imputation model for X_3 included X'_3;

FCSgroups: both imputation models included X'$_2$ and X'$_3$; *FCS3group2:* X_2 was imputed with FCSgroup and X_3 was imputed with FCS; *D*: number of studies; *m:* imputed datasets; *it:* iterations.

Scenario 2

This scenario is like scenario 1 with a difference of **1000** individuals per study. We present the simulation results in figure 6.5 with e_i: (N ~ (0, 0.2)), (N ~ (0, 2)), (N ~ (0, 20)) respectively. Like scenario 1, FCS gives biased results when model error is e_i: N ~ (0, 0.2). e_i: N ~ (0, 2). It leads to biased X_3 estimates and low coverage. FCSgroup and FCSgroups provided the most accurate estimates (FCS3group2 followed) and should be chosen as the preferred methods.

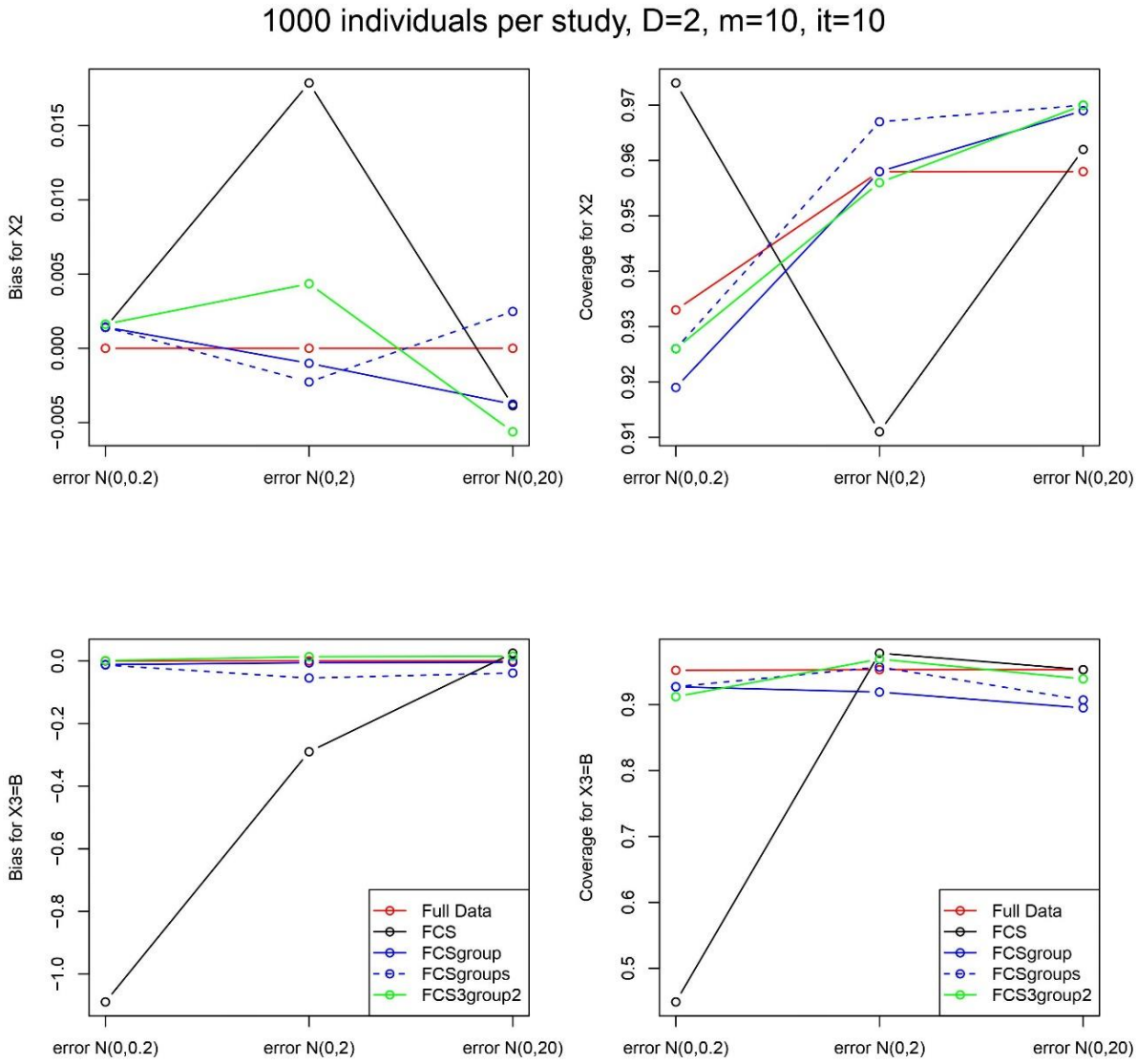


Figure 6.5. Main results from scenario 2's simulation study: Comparison of Bias and Coverage level, for $X_{3=B}$ and X_2 after 1000 simulations with true Full Data (red line),

handling granularity and mixed type with FCS (black line), FCSgroup (blue line), FCSgroups (broken blue line) and FCS3group2 (green line), for three model errors.

FCS: both imputation models excluded categorical informative variables X'_2 and X'_3; *FCSgroup*: imputation model for X_2 included X'_2 and imputation model for X_3 included X'_3; *FCSgroups*: both imputation models included X'_2 and X'_3; *FCS3group2:* X_2 was imputed with FCSgroup and X_3 was imputed with FCS; *D*: number of studies; *m:* imputed datasets; *it:* iterations.

Scenarios 3 – 4

In scenarios 3 - 4, we simulated data for **five studies**, each with the **same number of individuals** per study. In Scenario 3 each study had 200 individuals (so D_0 had 1000 individuals) and in scenario 4 each study had 1000 individuals (so D_0 had 5000 individuals). For each simulated dataset, we chose two random studies (D_4 and D_5 for scenario 3 and D_2 and D_5 for scenario 4) to apply the mixed type issue in X_2 in D_0. We present the simulation results in figures 6.6 and 6.7. In scenario 3, all imputation models gave very good results in terms of bias (figure 6.6), precision (figures 6.10, 6.11), and confidence interval coverage (figure 6). The integration of more same size studies showed that it helped imputation models gain more information and performed better. Similarly, in scenario 4, all imputation models performed very well so the mean estimates were very close to true Full Data. FCSgroup and FCSgroups gave the most accurate results in all model errors checked. Using FCS when imputing categorical X_3 showed worse estimates than using FCSgroup and FCSgroups. So, including 'group' variable(s) helped imputation model's accuracy.

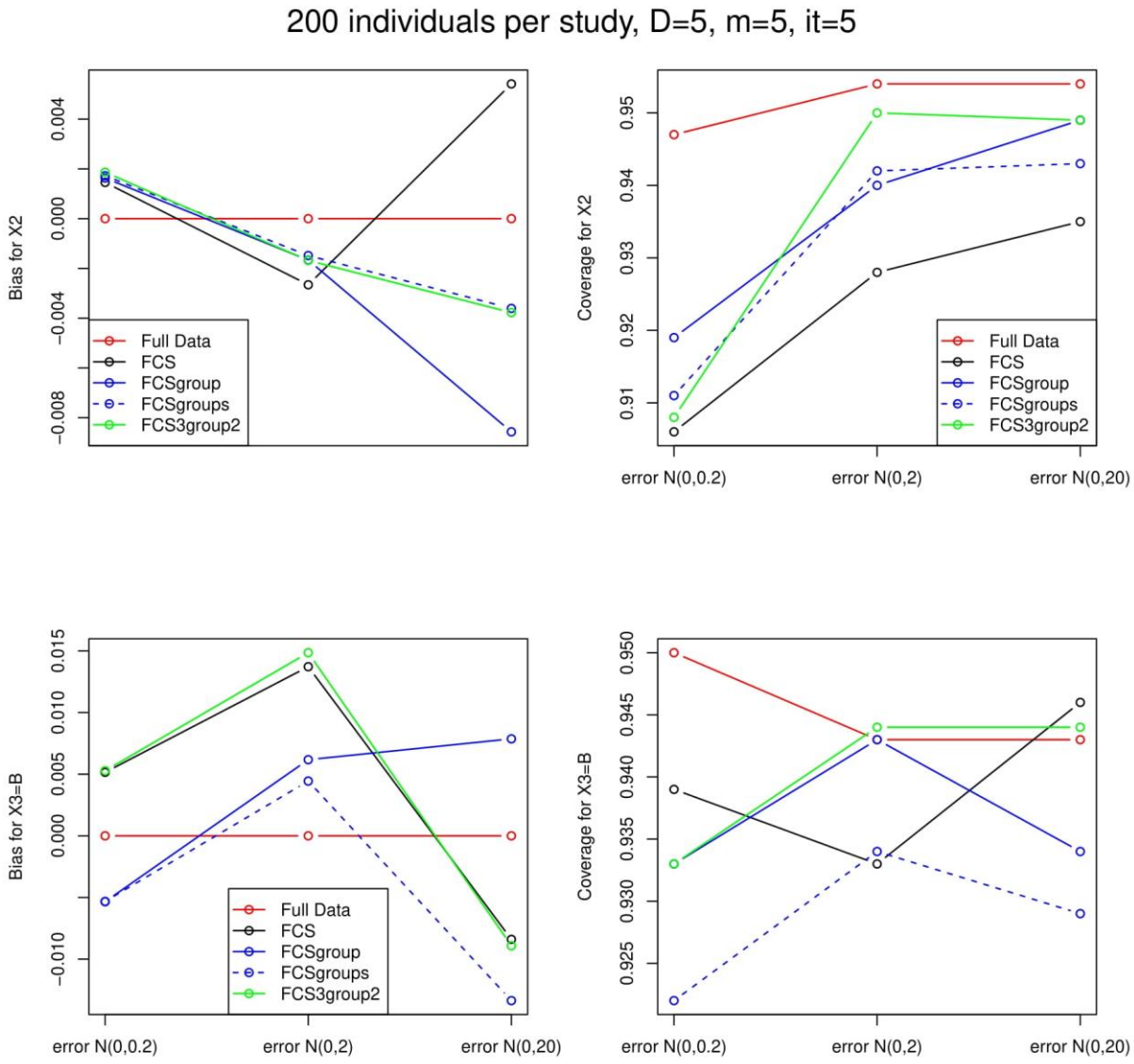


Figure 6.6. Main results from scenario 3's simulation study: Comparison of Bias and Coverage level, for $X_{3=B}$ and X_2 after 1000 simulations with true Full Data (red line), handling granularity and mixed type with FCS (black line), FCSgroup (blue line), FCSgroups (broken blue line) and FCS3group2 (green line), for three model errors.

FCS: both imputation models excluded categorical informative variables X'_2 and X'_3; *FCSgroup*: imputation model for X_2 included X'_2 and imputation model for X_3 included X'_3; *FCSgroups*: both imputation models included X'_2 and X'_3; *FCS3group2:* X_2 was imputed with FCSgroup and X_3 was imputed with FCS; *D*: number of studies; *m:* imputed datasets; *it:* iterations.

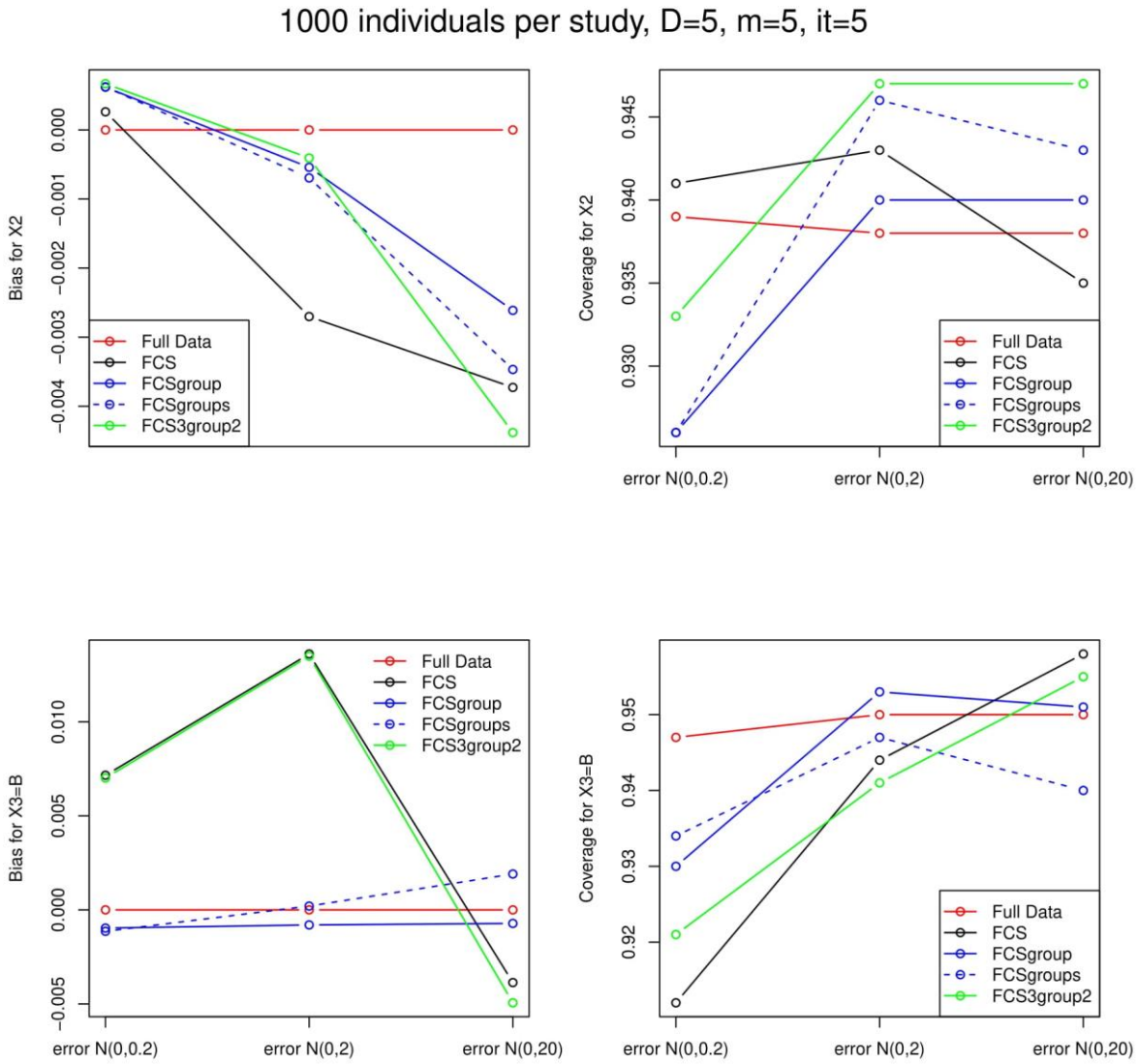


Figure 6.7. Main results from scenario 4's simulation study: Comparison of Bias and Coverage level, for $X_{3=B}$ and X_2 after 1000 simulations with true Full Data (red line), handling granularity and mixed type with FCS (black line), FCSgroup (blue line), FCSgroups (broken blue line) and FCS3group2 (green line), for three model errors.

FCS: both imputation models excluded categorical informative variables X'_2 and X'_3; *FCSgroup*: imputation model for X_2 included X'_2 and imputation model for X_3 included X'_3; *FCSgroups*: both imputation models included X'_2 and X'_3; *FCS3group2:* X_2 was imputed with FCSgroup and X_3 was imputed with FCS; *D*: number of studies; *m:* imputed datasets; *it:* iterations.

Overall, in all scenarios imputation methods produced accurate results (figure 6.8). As model error increased, mSE and EmpSE increased as well. mSE and EmpSE (figures 6.10 and 6.11) were almost identical per scenario and per imputation model, and they also decreased as the integrated dataset's size increased. In all four scenarios, when applying the Complete Records approach, the results were highly biased, resulting in large standard errors and under-coverage.

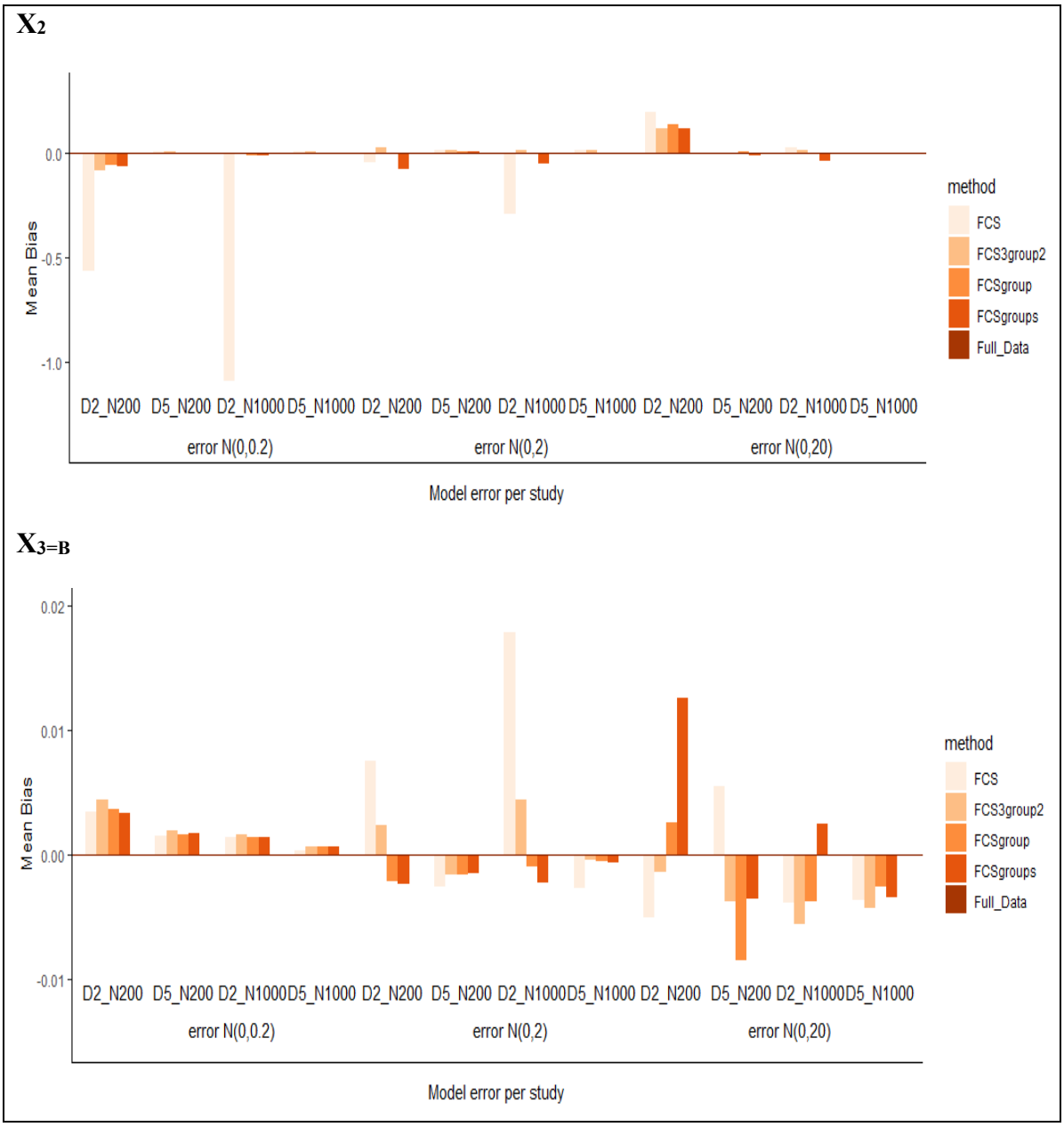


Figure 6.8. Bias for X_2 and $X_{3=B}$ for '2 datasets, N=200 per dataset' (D2_N200), '5 datasets, N=200 per dataset' (D5_N200), '2 datasets, N=1000 per dataset' (D2_N1000) and '5 datasets, N=1000 per dataset' (D5_N1000) for the three model errors.

FCS: both imputation models excluded categorical informative variables X'$_2$ and X'$_3$; *FCSgroup*: imputation model for X_2 included X'$_2$ and imputation model for X_3 included X'$_3$; *FCSgroups*: both imputation models included X'$_2$ and X'$_3$; *FCS3group2:* X_2 was imputed with FCSgroup and X_3 was imputed with FCS; *D*: number of studies; *m:* imputed datasets; *it:* iterations.

X2

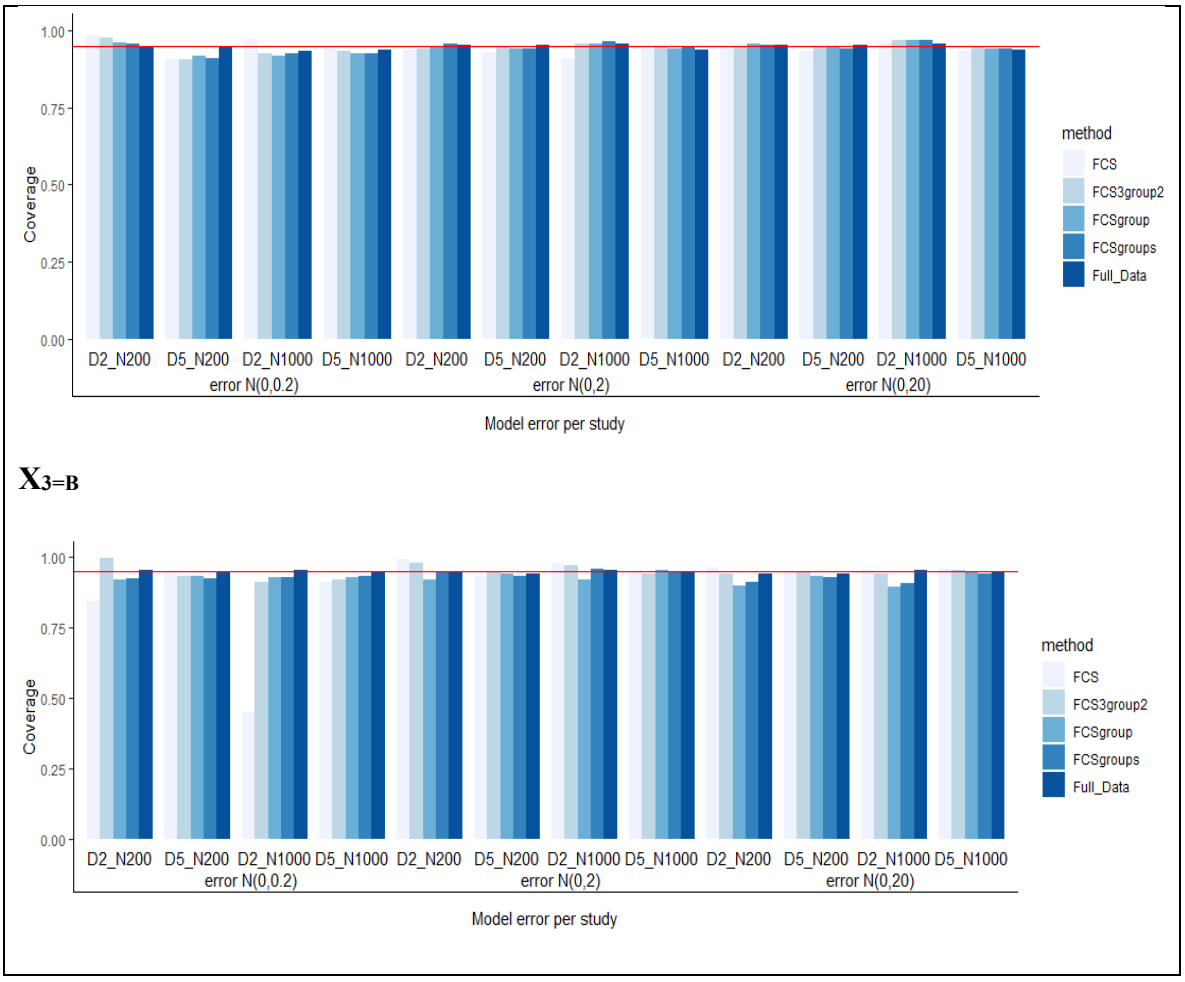
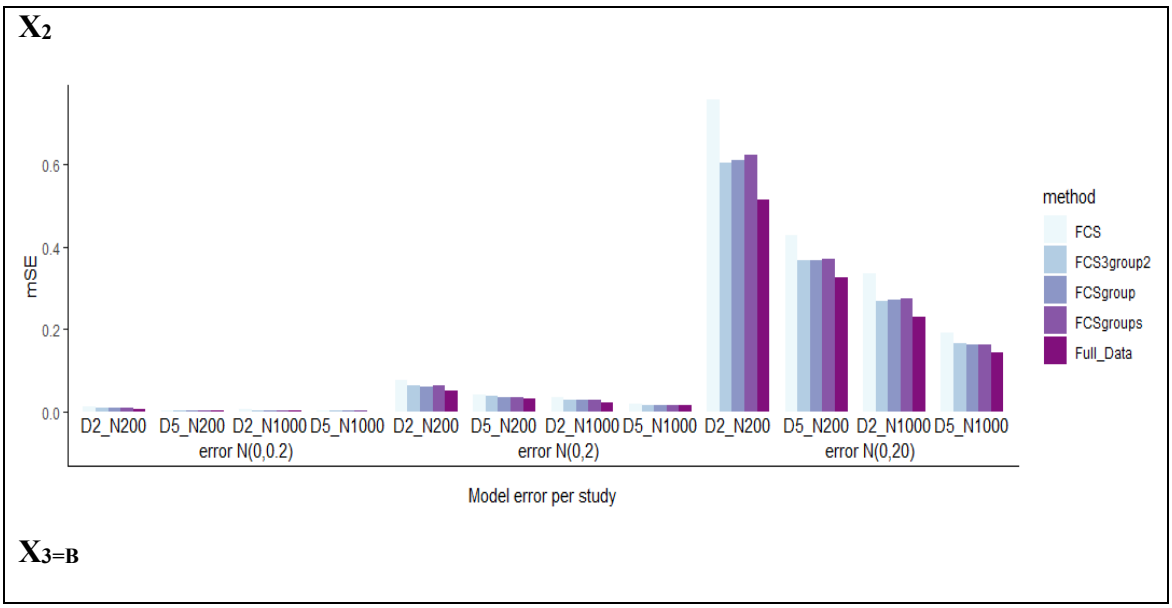


Figure 6.9. Coverage for X_2 and $X_{3=B}$ for '2 datasets, N=200 per dataset' (D2_N200), '5 datasets, N=200 per dataset' (D5_N200), '2 datasets, N=1000 per dataset' (D2_N1000) and '5 datasets, N=1000 per dataset' (D5_N1000) for the three model errors.

FCS: both imputation models excluded categorical informative variables X'_2 and X'_3; *FCSgroup*: imputation model for X_2 included X'_2 and imputation model for X_3 included X'_3; *FCSgroups*: both imputation models included X'_2 and X'_3; *FCS3group2:* X_2 was imputed with FCSgroup and X_3 was imputed with FCS; *D*: number of studies; *m:* imputed datasets; *it:* iterations.



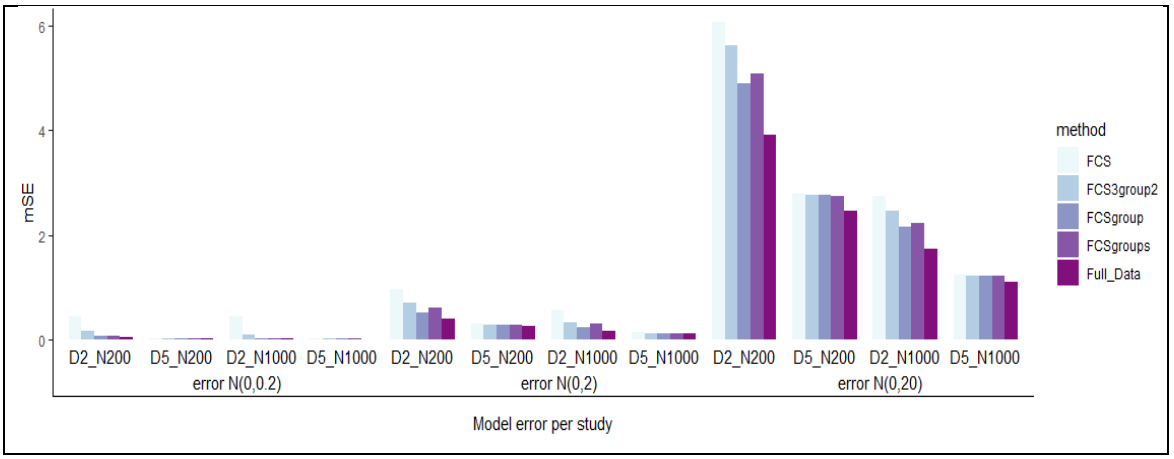


Figure 6.10. mSE for X_2 and $X_{3=B}$ for '2 datasets, N=200 per dataset' (D2_N200), '5 datasets, N=200 per dataset' (D5_N200), '2 datasets, N=1000 per dataset' (D2_N1000) and '5 datasets, N=1000 per dataset' (D5_N1000) for the three model errors.

FCS: both imputation models excluded categorical informative variables X'_2 and X'_3; *FCSgroup*: imputation model for X_2 included X'_2 and imputation model for X_3 included X'_3; *FCSgroups*: both imputation models included X'_2 and X'_3; *FCS3group2:* X_2 was imputed with FCSgroup and X_3 was imputed with FCS; *mSE:*mean model standard error.

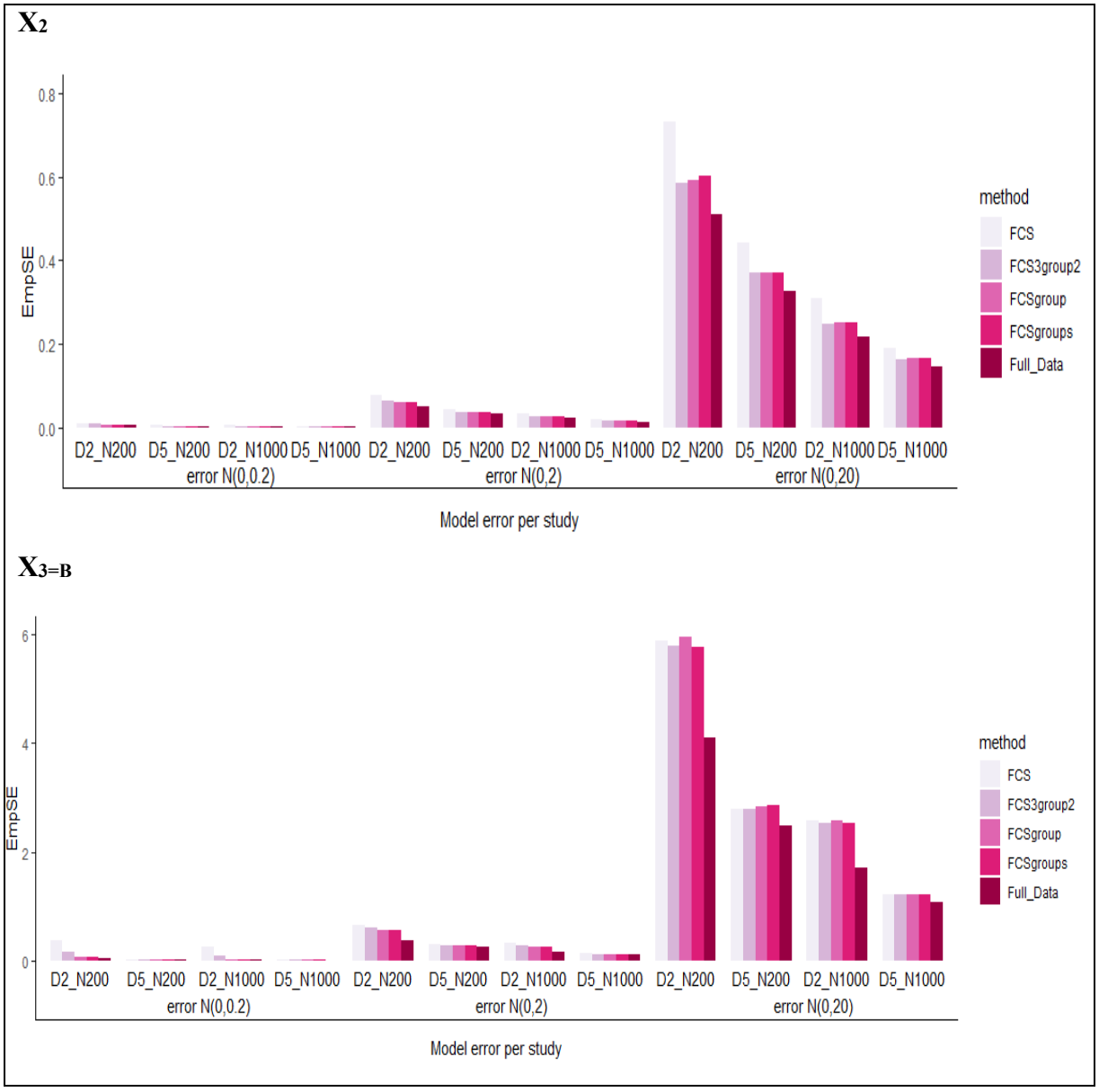


Figure 6.11. EmpSE for for X_2 and $X_{3=B}$ for '2 datasets, N=200 per dataset' (D2_N200), '5 datasets, N=200 per dataset' (D5_N200), '2 datasets, N=1000 per dataset' (D2_N1000) and '5 datasets, N=1000 per dataset' (D5_N1000) for the three model errors.

FCS: both imputation models excluded categorical informative variables X'_2 and X'_3; *FCSgroup*: imputation model for X_2 included X'_2 and imputation model for X_3 included X'_3; *FCSgroups*: both imputation models included X'_2 and X'_3; *FCS3group2:* X_2 was imputed with FCSgroup and X_3 was imputed with FCS; EmpSE: mean empirical standard error.

Simulations with studies of different size and same model error per study (Scenarios 5 - 6)

Scenario 5

We simulate data for **five studies**, each with **different number of individuals** (D_1:200, D_2:150, D_3:50, D_4:75, D_5:100). X_3 had missing values in study D_4 and X_2 had missing values in study D_5. See results in figure 6.12. All imputation models offered similar results when the model error was either e_i: N ~ (0, 0.2) or e_i: N ~ (0, 2). When the model error was e_i: N ~ (0, 20)), mSE and EmpSE increased but without introducing bias.

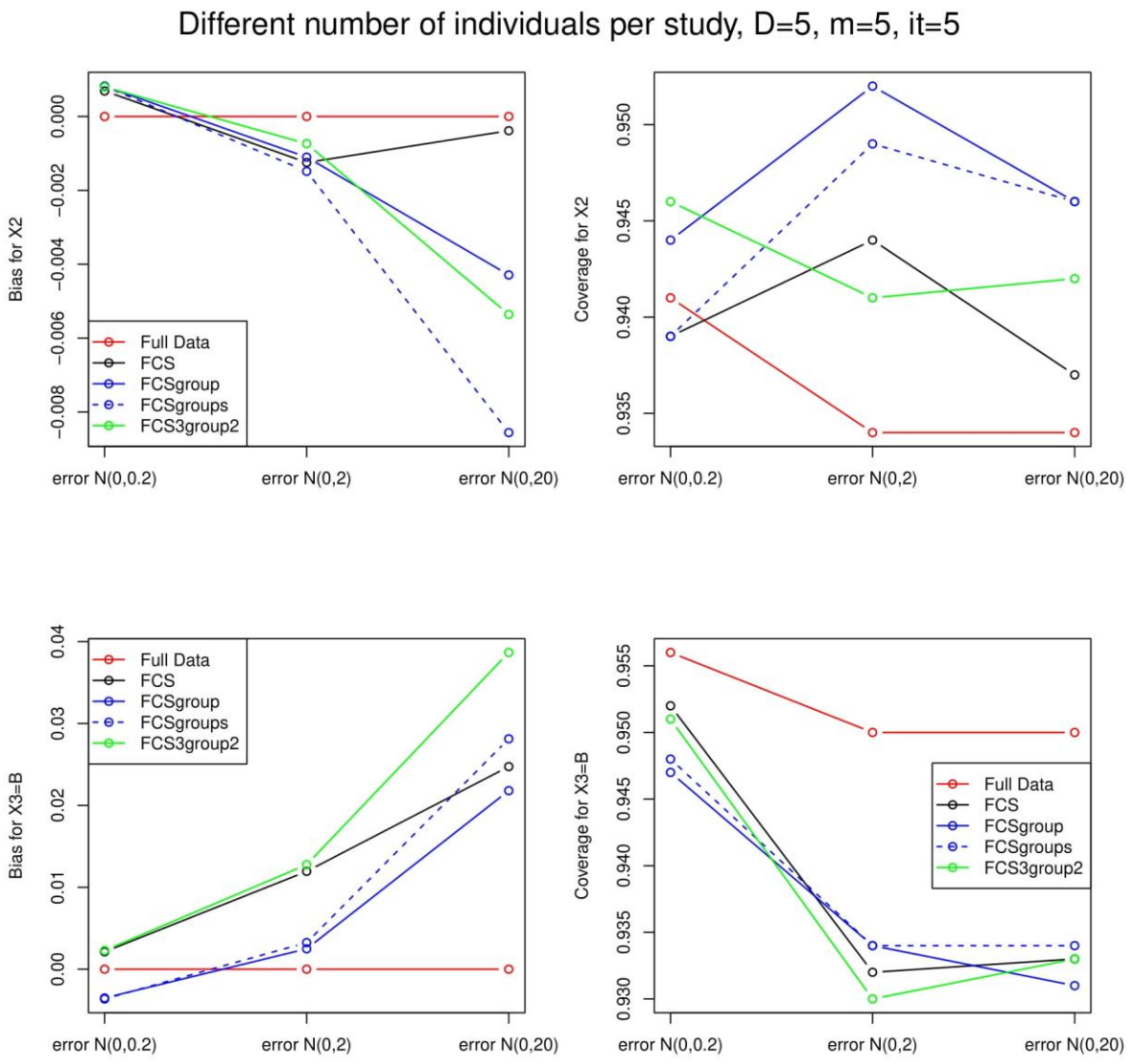


Figure 6.12. Main results from scenario 5's simulation study: Comparison of Bias and Coverage level, for $X_{3=B}$ and X_2 after 1000 simulations with true Full Data (red line), handling granularity and mixed type with FCS (black line), FCSgroup (blue line), FCSgroups (broken blue line) and FCS3group2 (green line), for three model errors.

FCS: both imputation models excluded categorical informative variables X'_2 and X'_3; *FCSgroup*: imputation model for X_2 included X'_2 and imputation model for X_3 included X'_3; *FCSgroups*: both imputation models included X'_2 and X'_3; *FCS3group2:* X_2 was imputed with FCSgroup and X_3 was imputed with FCS; *D*: number of studies; *m:* imputed datasets; *it:* iterations.

Scenario 6

We simulate data from **ten** studies, each **with different number of individuals** (D_1:800, D_2:150, D_3:50, D_4:75, D_5:350, D_6:200, D_7:150, D_8:500, D_9:750, D_{10}:100). There were missing data in X_3 in studies D_6 and D_{10}, due to granularity problem, and missing data in X_2 in studies D_3 and D_9 , due to mixed type problem. Results for this scenario are shown in figure 6.13. Compared to scenario 5, estimates from imputation models were closer to the true Full Data model and that led mSE and EmpSe be also smaller. Coverage levels were around 92-95%.

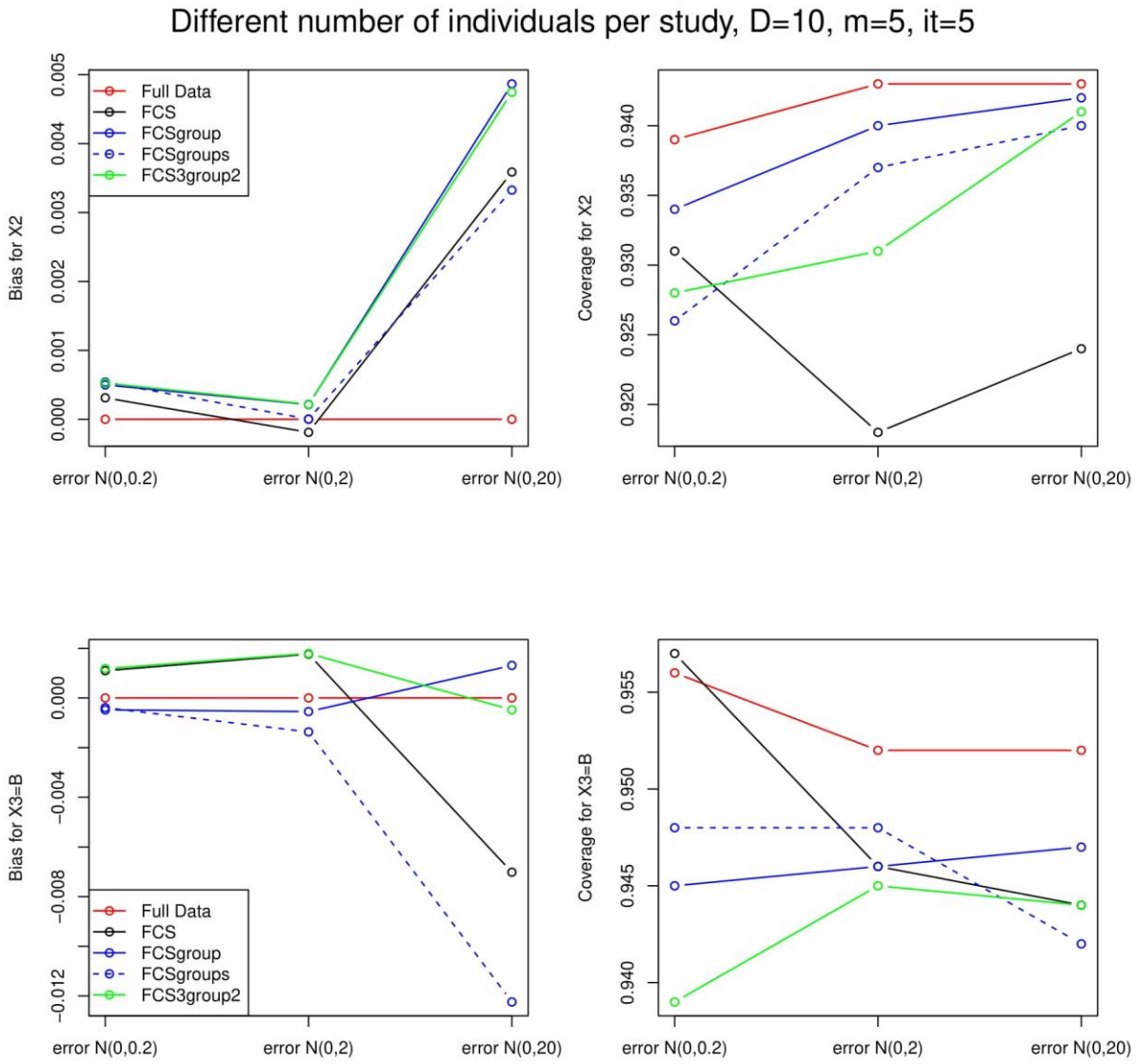


Figure 6.13. Main results from scenario 6's simulation study: Comparison of Bias and Coverage level, for $X_{3=B}$ and X_2 after 1000 simulations with true Full Data (red line), handling granularity and mixed type with FCS (black line), FCSgroup (blue line), FCSgroups (broken blue line) and FCS3group2 (green line), for three model errors.

FCS: both imputation models excluded categorical informative variables X'_2 and X'_3; *FCSgroup*: imputation model for X_2 included X'_2 and imputation model for X_3 included X'_3; *FCSgroups*: both imputation models included X'_2 and X'_3; *FCS3group2:* X_2 was imputed with FCSgroup and X_3 was imputed with FCS; *D*: number of studies; *m:* imputed datasets; *it:* iterations.

Overall, we can see a comparison of bias, coverage level, mSE and EmpSE for both scenarios for the three model errors in figures 6.14 to 6.17 respectively. Bias decreased when integrated dataset had more individuals - scenario 6 (figure 6.14). Model error affected estimations' accuracy. When outcome Y had a big range (e_i: N ~ (0, 20)), missing data's estimates were getting slightly worse. FCSgroup and FCSgroups had better estimates than FCS and FCS3group2 but again all imputation methods were unbiased. mSE and EmpSE

are similar per scenario, they decreased as the integrated dataset's size increased. In scenarios 5 and 6, all probabilistic models outperform Complete Records.

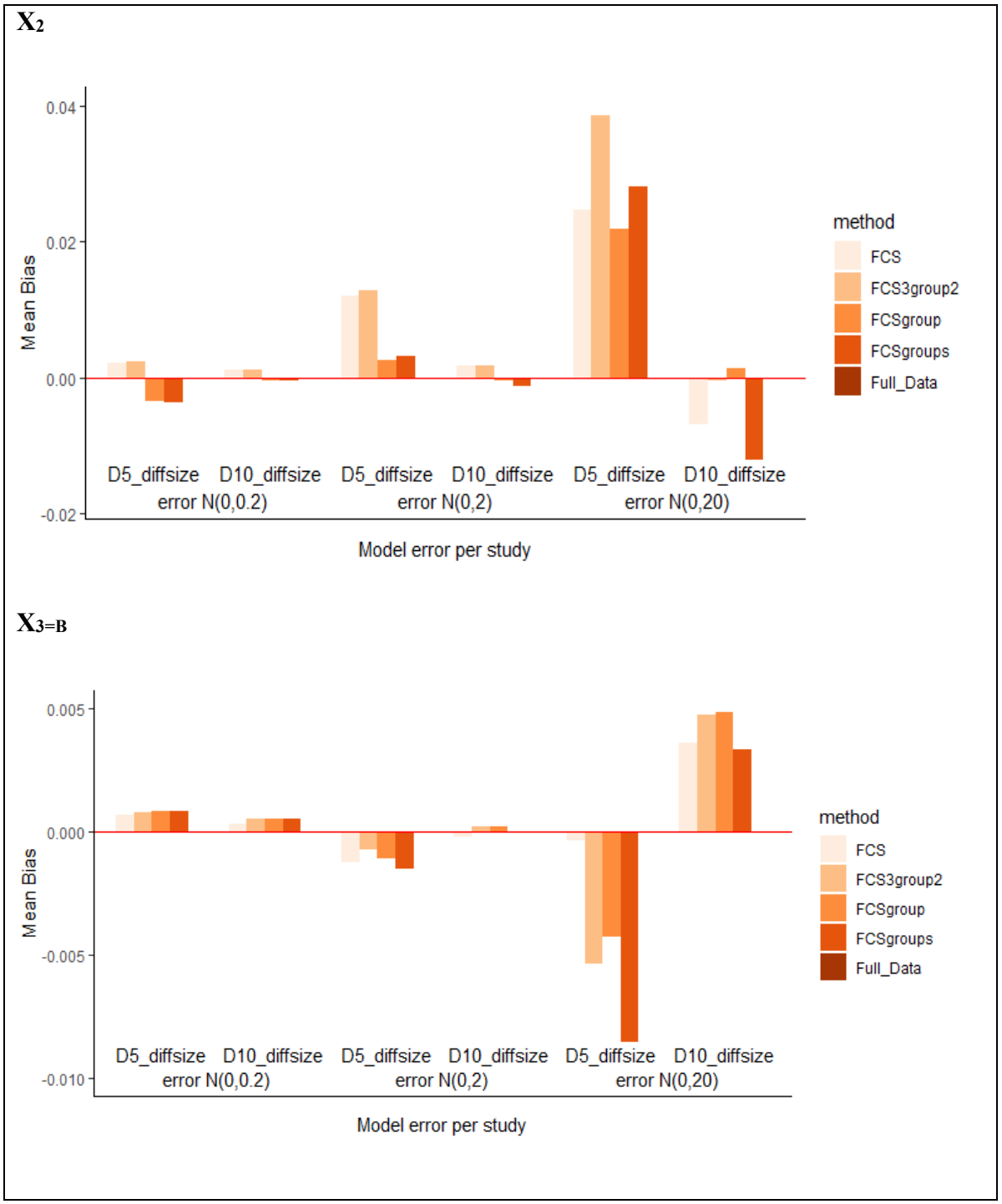


Figure 6.14. Bias for X_2 and $X_{3=B}$ for '5 datasets, N=different per dataset' (D5_diffsize), '10 datasets, N=different per dataset' (D10_diffsize) for the three model errors.

FCS: both imputation models excluded categorical informative variables X'_2 and X'_3; *FCSgroup*: imputation model for X_2 included X'_2 and imputation model for X_3 included X'_3; *FCSgroups*: both imputation models included X'_2 and X'_3; *FCS3group2:* X_2 was imputed with FCSgroup and X_3 was imputed with FCS.

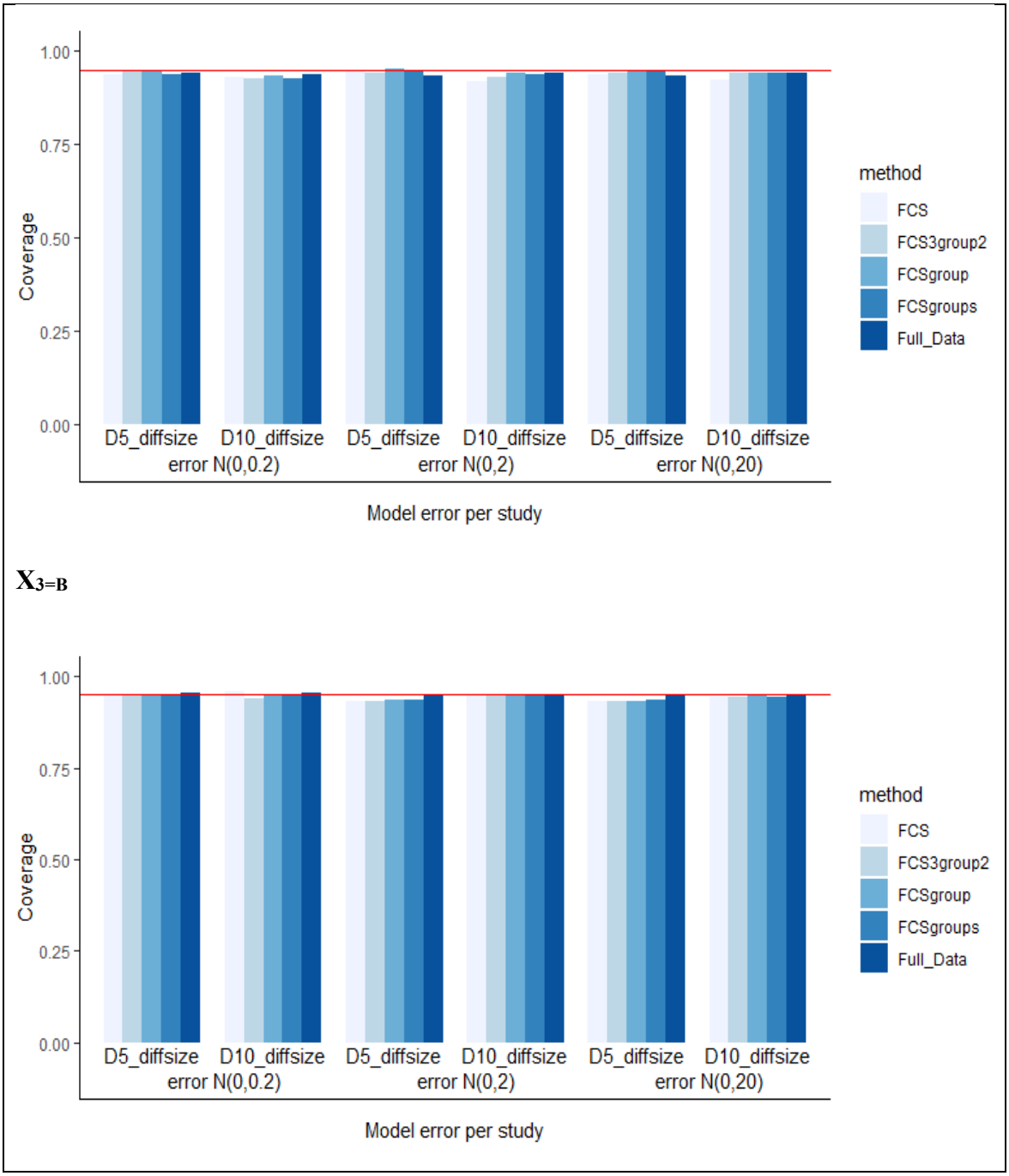
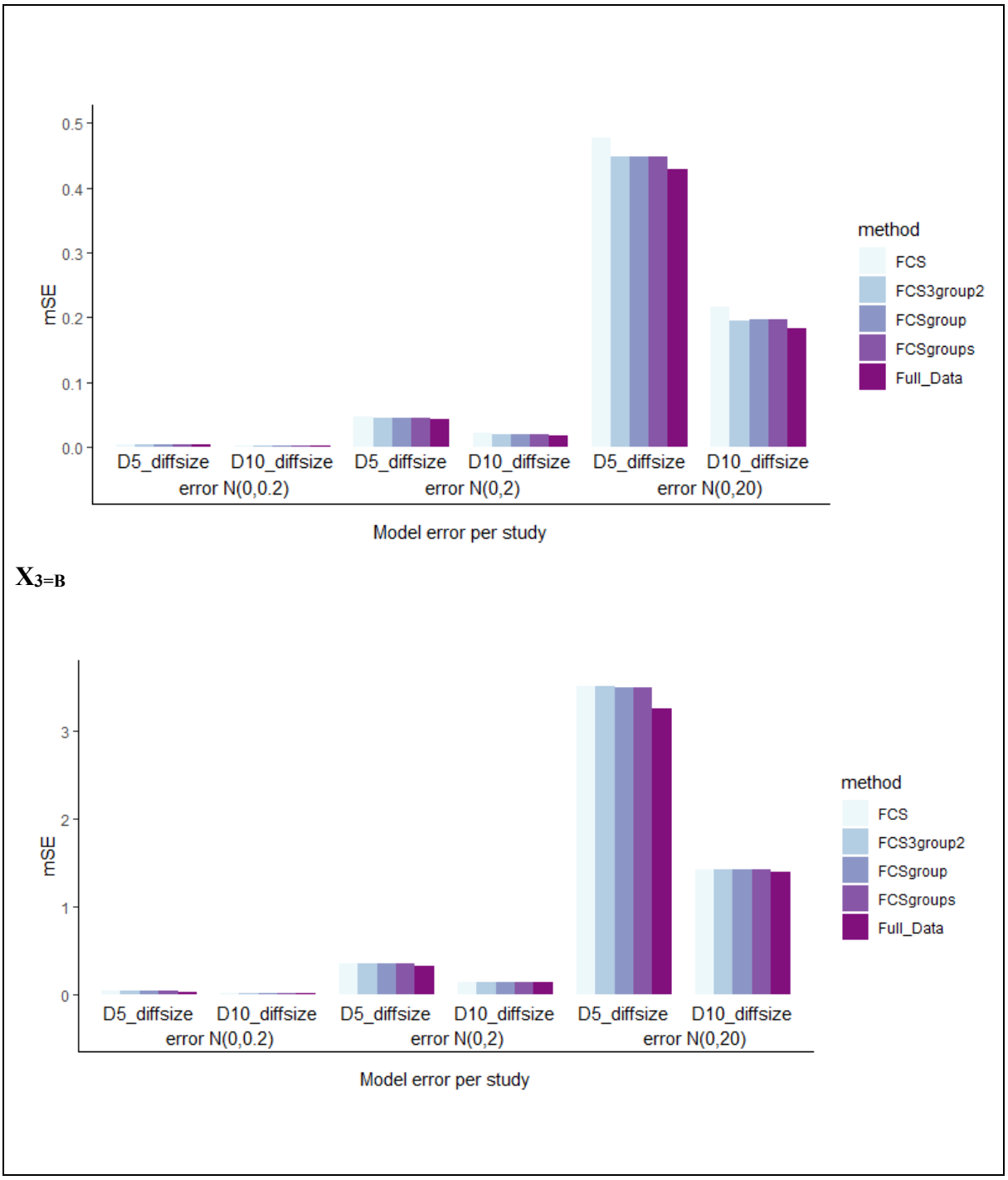


Figure 6.15. Coverage for X_2 and $X_{3=B}$ for '5 datasets, N=different per dataset' (D5_diffsize), '10 datasets, N=different per dataset' (D10_diffsize) for the three model errors.

FCS: both imputation models excluded categorical informative variables X'$_2$ and X'$_3$; *FCSgroup*: imputation model for X_2 included X'$_2$ and imputation model for X_3 included X'$_3$; *FCSgroups*: both imputation models included X'$_2$ and X'$_3$; *FCS3group2:* X_2 was imputed with FCSgroup and X_3 was imputed with FCS.

X_2



$X_{3=B}$

Figure 6.16. mSE for X_2 and $X_{3=B}$ for '5 datasets, N=different per dataset' (D5_diffsize), '10 datasets, N=different per dataset' (D10_diffsize) for the three model errors.

FCS: both imputation models excluded categorical informative variables X'$_2$ and X'$_3$; *FCSgroup*: imputation model for X_2 included X'$_2$ and imputation model for X_3 included X'$_3$; *FCSgroups*: both imputation models included X'$_2$ and X'$_3$; *FCS3group2:* X_2 was imputed with FCSgroup and X_3 was imputed with FCS; *mSE:* mean model standard error.

X_2

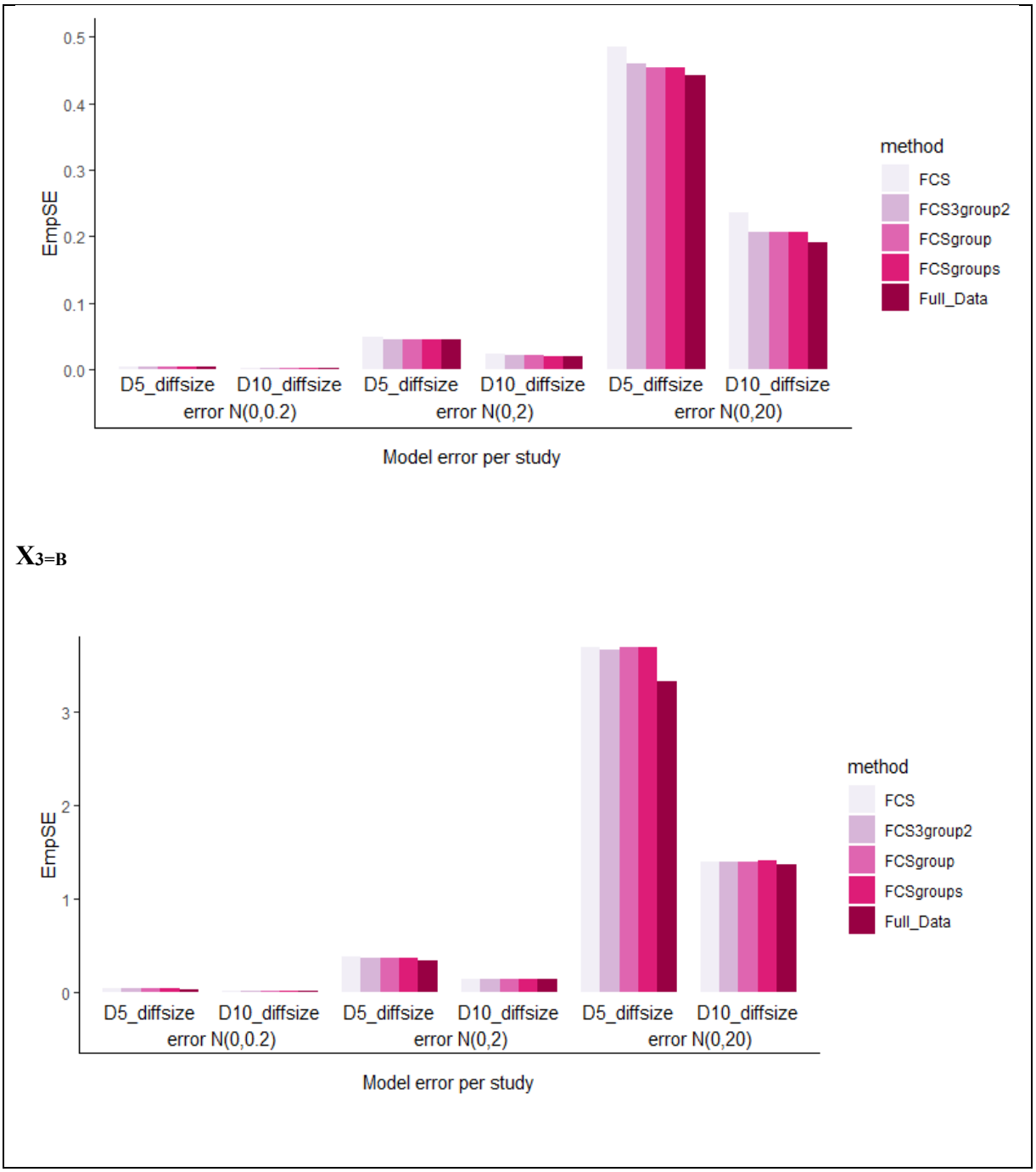


Figure 6.17. EmpSE for X_2 and $X_{3=B}$ for '5 datasets, N=different per dataset' (D5_diffsize), '10 datasets, N=different per dataset' (D10_diffsize) for the three model errors.

FCS: both imputation models excluded categorical informative variables X'$_2$ and X'$_3$; *FCSgroup*: imputation model for X_2 included X'$_2$ and imputation model for X_3 included X'$_3$; *FCSgroups*: both imputation models included X'$_2$ and X'$_3$; *FCS3group2:* X_2 was imputed with FCS and X_3 was imputed with FCS; *EmpSE:* mean empirical standard error.

Simulations with equal size studies, different model error per study (7 - 10)

Scenarios 7 – 10

We simulated data from two equal size studies but with different model error per study and different number of individuals per scenario. In Scenarios 7-10 each study had 100, 200, 500 and 1,000 individuals respectively. We present results for 10 imputations 10 iterations.

Simulations results improved a lot and achieved better convergence when we increased the number of imputations and iterations from 5 to 10 (check tables D.29-D.32 in Appendix D). For each simulated dataset, we applied the mixed type problem to X_2 (data missingness in D_2) and granularity problem to X_3 (data missingness in D_1). We present the simulation results in figure 5.18. Results showed that FCS produced biased estimates and under-coverage of the confidence interval. FCSgroup and FCSgroups showed no bias and very good coverage levels, almost identical with true model's, in all the examined size studies. FCSgroup, FCSgroups and FCS3group2 outperformed Complete Records which was not suggested as a data integration solution on this case either. FCS did not perform well and should not be chosen to solve both content heterogeneity problems as indicated scenarios 8 to 10.

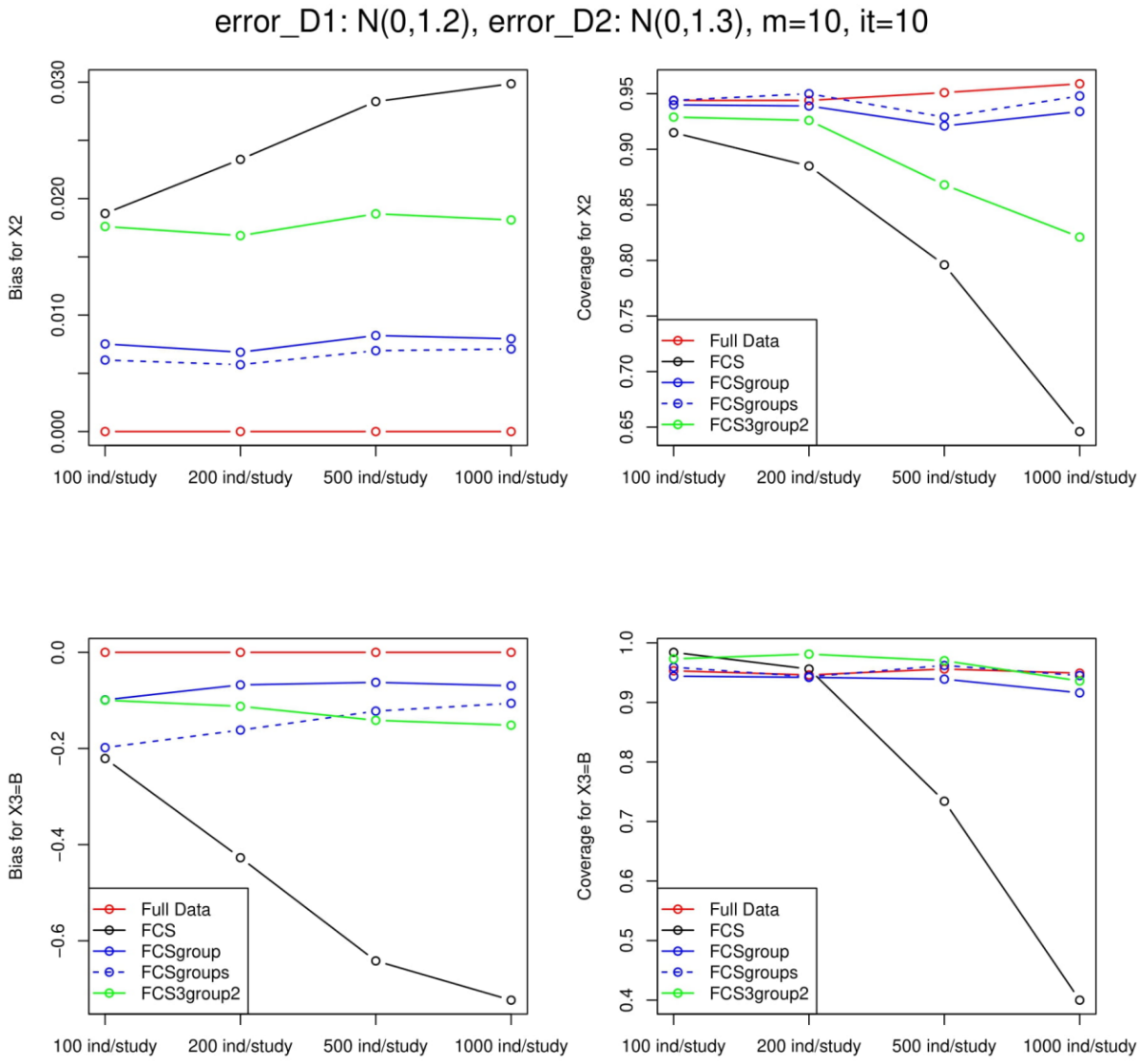


Figure 6.18. Main results from scenario 7-10's simulation study: Comparison of Bias and Coverage level, for $X_{3=B}$ and X_2 after 1000 simulations with true Full Data (red line), handling granularity and mixed type with FCS (black line), FCSgroup (blue line), FCSgroups (broken blue line) and FCS3group2 (green line), for three model errors.

FCS: both imputation models excluded categorical informative variables X'_2 and X'_3; *FCSgroup*: imputation model for X_2 included X'_2 and imputation model for X_3 included X'_3; *FCSgroups*: both imputation models included X'_2 and X'_3; *FCS3group2:* X_2 was imputed with FCSgroup and X_3 was imputed with FCS; *D*: number of studies; *m:* imputed datasets; *it:* iterations.

6.3.3 Summary of findings from simulation studies

In this section we summarise briefly the main points from simulation studies with combined heterogeneity types of problems.

Studies with the same model errors

We observe that FCSgroup and FCSgroups provided unbiased estimates, almost identical with true Full Data and they solved mixed type and granularity successfully. In all scenarios as the model error increased, mSE and EmpSE increased by remaining equal between them.

Studies with different model errors

Smaller datasets provide worse estimates probably because of smaller data sample. FCSgroup and FCSgroups had good coverage and standard errors were close to reference model's (Full Data). FCS should not be the preferred imputation model. We had to increase the number of imputations and iterations to 10 to explore if estimates would have been improved. Therefore, we may conclude that if the number of imputations and iterations is larger, it achieves better convergence.

Studies of same sizes

In all scenarios, FCSgroup and FCSgroups were compatible models with Full Data. FCS3group2's showed slightly biased estimates in X_3 which makses sense as X_3 was imputed with FCS. Imputation model FCS produced biased results in some cases was similar – probably due to data misclassification. However, for both scenarios 1 and 2 we had to increase the number of imputations and iterations from 5 to 10 and estimates improved significantly afterwards. In scenarios 2 and 4, where we had 5 studies integrated into one, regression estimates were unbiased and there was no need to increase the number of imputed datasetes. Therefore, when the integrated dataset consisted of two datasets, more than 5 imputations and iterations were needed to achieve unbiased results and good coverage. In general, FCSgroup, and FCSgroups' biases remained negligible, and coverage was very good and there was a lot more gain in information compared with a Complete Records analysis.

Studies of different sizes

In both scenarios, we have good coverage and there is either no bias or small (FCSgroup, D5_diffsize, large model error). When we have 10 datasets with different sizes both imputation models provided unbiased results and FCSgroup had slightly better estimates than FCS. However, in the 5 different size datasets example FCS shows better mean estimate than FCSgroup when model errors are small and large. Additionally, mSE and EmpSE decreased as the study size increased, the same happened to bias.

Size of the model error

The larger the model error per study, the larger mSE and EmpSE which made sense as all imptuation methods followed Full Data standard errors. When outcome Y had a big range (e_i: N ~ (0, 20)), missing data's estimates were getting slightly worse.

Size of studies

In cases with the same model error per study, we observed that the larger the dataset the better the estimates. On the contrary, in cases with different model error per study, estimates were getting worse with the increase of study size. A reason could be that data missingness was higher (only two datasets were integrated). mSE and EmpSE had same pattern in all scenarios, they decreased as the integrated dataset's size increased.

Number of imputations and iterations

In scenarios where we had two integrated study datasets (scenarios: 1, 2, 7-10), we increased the number of imputations to investigate if it would help the imputation algorithms perform better. They indeed improved a lot and provided very good results in most cases. In the FCS case and specifically in scenarios 7-10, the algorithm improved but not in the level to achieve unbiased and accepted results. Imputations for all scenarios with 5 imputations and 5 iterations are in Appendix D. All results for scenarios described in tables 6.1 and 6.2 are also available in Appendix D. Imputations and iterations' number should be larger when we have two datasets that each one has one content heterogeneity problem and therefore both studies have missing data.

Overall

In FCS, as the model error's variance increases there was underestimations for categorical X_3 and overestimation of refression coefficient for continuous X_2. We could say that large variances of the measured outcome Y would have large estimated values for the variance of the missing covariates. Our new suggestion about including categorical informative variable(s) in the imputation model(s) seems to help classification and it outperformed default FCS in many cases. Furthermore, as we discussed in chapter 5, FCS and FCSgroup

gave similar unbiased results with good coverage. We wanted to explore it more and for that reason, we introduced FCS3group2, to detect a difference between FCS and FCSgroup in mixed type problem when combined with granularity. Based on our parametric simulation studies we could conclude in FCSgroup's superiority above FCS in complex situations.

The probabilistic approaches FCSgroup and FCSgroups are suggested to solve coexisting content heterogeneity problems in an integrated dataset, i.e. granularity and mixed type problematic variables. Likewise, results show that using the available extra information may help having better estimations while solving both content hterogeneities.

6.4 Application – MASTERPLANS exemplar

As we did in the previous experimental chapters, in this section we illustrated the application of probabilistic approaches to solve content heterogeneity problems after integrating real-world data. We also compared traditional and probabilistic data integration in combined content heterogeneity problems.

6.4.1 Data characteristics and combined types of content heterogeneity

To illustrate the developed approaches (Figures 6.1 and 6.2), we integrated health data in order to answer a biomedical research question about SLE. For datasets D_1, D_2, D_3, we had datasets that contained data from ALMS, LUNAR, and EXPLORER respectively. The integrated dataset D_0 consisted of the following 17 variables: gender, age, ethnicity, height, weight, BMI, creatinine, current treatment, and various BILAG disease activity scores (Total, Cardiorespiratory, General, Mucocutaneous, Musculoskeletal, Neurological, Renal, Vasculitis, and Haematology). For the measure of response to treatment and creatinine levels, in order to limit to one per patient, we kept the value recorded at visit that had the least absolute difference from 90 days (3 months). The number of the patients were 597 after that filtering. Our aim was to construct a model predicting 'Renal response' using BMI and 4 other variables: age, ethnicity, creatinine, and gender (equation 6.1). For variables' summary and patients' baseline characteristics, we advise the reader to check Table 5.3, in section 5.5.

$$Renal\ BILAG\ Score = Age + Ethnicity + Creatinine + Gender + BMI \quad (6.1)$$

In order to ask the research question shown in equation 6.1, we introduced two content heterogeneity problems to illustrate the probabilistic and traditional data integration approaches (table 6.3).

About mixed type problem (table 6.4), we mutated 'age' variable from a continuous variable to a categorical one. In ALMS, patients' age was as '0-20', '21-40', '41-60', '>60' and in LUNAR and EXPLORER, patients age records were as integers. We also added the categorical informative 'age_{group}' variable to D_0 that categorised patients into four levels '0-20', '21-40', '41-60', '>60' based on their true age. We then removed patients' *age* records from ALMS study, so data missingness was introduced.

About granularity problem (table 6.4), 'ethnicity variable' in ALMS had 14 levels, i.e., 'Algerian', 'Asian', 'Black or African American', 'Cape Coloured', 'Caucasian', 'East Indian', 'Eritrean', 'Hispanic', 'Mexican Mestizo', 'Middle Eastern', 'Mixed', 'Moroccan', 'Native American', 'Peruvian'). However, in LUNAR and EXPLORER ethnicity's levels were 3 i.e., 'Caucasian' and 'Black or African American' and 'Other'. Therefore, in ALMS, ethnicity's granularity was very high. We added the categorical informative '$ethnicity_{group}$' variable to D_0 that categorised patients into three levels 'Caucasian' and 'Black or African American' and 'Other' based on their true ethnicity. We then removed patients' *ethnicity* records (only for level 'Other') from LUNAR and EXPLORER studies, so data missingness was introduced.

Table 6.3. Ethnicity's data characteristics after integrating lupus studies ALMS, LUNAR, EXPLORER.

Data characteristics	D_0, N=597 (%)	ALMS, N=248 (41.50)	LUNAR, N=138 (23.10)	EXPLORER, N=211 (35.40)
Ethnicity (%)				
Algerian	2 (0.34)	2 (0.81)		
Asian	81 (13.60)	81 (32.70)		
Black or African American	115 (19.30)	26 (10.50)	37 (26.80)	52 (24.60)
Cape Coloured	1 (0.16)	1 (0.40)		
Caucasian	273 (45.70)	108 (43.50)	44 (31.90)	121 (57.30)
East Indian	1 (0.16)	1 (0.40)		
Eritrean	2 (0.34)	2 (0.81)		
Hispanic	2 (0.34)	2 (0.81)		
Mexican Mestizo	19 (3.18)	19 (7.66)		
Middle Eastern	1 (0.16)	1 (0.40)		
Mixed	2 (0.34)	2 (0.81)		

Moroccan	1 (0.16)	1 (0.40)		
Native American	1 (0.16)	1 (0.40)		
Peruvian	1 (0.16)	1 (0.40)		
Other	95 (15.90)		57 (41.30)	38 (18.10)

Table 6.4. Mapping between ethnicity's levels, and age's levels. Traditional VS Probabilistic data integration.

Traditional data integration	**Probabilistic data integration**
Age – mixed type problem	
'0-20'	Integer
'21-40'	
'41-60'	
'>60'	
Ethnicity – granularity problem	
'Black or African American'	'Black or African American'
'Caucasian'	'Caucasian'
'Other'	'Algerian'
	'Asian'
	'Cape Coloured'
	'East Indian'
	'Eritrean'
	'Hispanic'
	'Mexican Mestizo'
	'Middle Eastern'
	'Mixed'
	'Moroccan'
	'Native American'
	'Peruvian'

6.4.2 Results

6.4.2.1 Traditional Data Integration – mapping all values present in all datasets

Traditionally, people solve these problems by mapping all datasets to a common data model. This model would need homogeneity in all similar variable types across studies. In the case of a mixed type problem for the Age variable (table 6.4), it was converted to a categorical

variable across all the three studies. So, we replaced patients' age records in LUNAR and EXPLORER as categories '0-20', '21-40', '41-60', '>60' (practically, we used the 'age_{group}' and '$ethnicity_{group}$' that we have already introduced in our dataset). This model would have also included only variables' levels that were present in all datasets, and, in case of granularity problem, levels that existed in all datasets i.e. 'Black or African American', 'Caucasian' and 'Other'. Table 6.5 shows the coefficients for the linear regression model that estimates the renal drug response based on age, ethnicity, creatinine, gender, and BMI.

Table 6.5. Coefficients (estimate, standard error, t statistic and p-values) for linear regression model from equation 6.1 after applying complete case analysis (Complete Records) in SLE data to solve combined content heterogeneity problems.

	estimate	**standard error**	**t statistic**	**p-value**
(Intercept)	8.7826	1.1361	7.7300	0.0000***
Age				
21-40	-2.0013	0.6879	-2.9090	0.0038**
41-60	-4.0000	0.7592	-5.2680	0.0000***
>60	-5.0391	1.7704	-2.8460	0.0046**
Ethnicity				
Caucasian	1.4780	0.5548	2.6640	0.0079**
Other	2.8542	0.5850	4.8790	0.0000***
Creatinine	1.8614	0.4609	4.0390	0.0001***
Gender				
Male	0.5598	0.6317	0.8860	0.3759
BMI	-0.1505	0.0319	-4.7200	0.0000***

*Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

6.4.2.2 Probabilistic Data Integration – multiple imputation

Our aim was to answer a research question (equation 6.1), so we applied the suggested probabilistic data integration approaches. *About mixed type problem*, we kept the largest number of levels in Age, as integer variable. Therefore, D_0 had missing data in ALMS for the Age variable. *About granularity problem*, we kept the highest number of levels in ethnicity which was 14 levels and had missing data in LUNAR and EXPLORER and imputed them based on some available information using FCS, FCSgroup, FCSgroup**s**, FCS3group2. In all imputation models the following variables were used as predictors: gender, age, ethnicity, BMI, creatinine, current treatment, BILAG disease activity scores

(Cardiorespiratory, General, Mucocutaneous, Musculoskeletal, Neurological, Renal, Vasculitis, and Haematology). BMI and total BILAG score variables in the MASTERPLANS dataset were not included in the imputation models because they were recoded versions and combinations of other data. We chose as imputation methods, *pmm* for Age and *polyreg* for Ethnicity and we also set imputations and iterations to 10. At the end, we fitted linear regression models to complete datasets that resulted from all multiple imputation methods, so we answered the research question (equation 6.1). Coefficients are shown in tables 6.6 to 6.9 for FCS, FCSgroup, FCSgroups, and FCS3group2 respectively. We set seed to 945783.

In general, the results of application in integrated SLE data show that multiple imputation approaches have a substantial impact on point estimates of the coefficients. We observe a gain in precision for all MI methods for complete variables (creatinine, BMI, gender) and for incomplete variables (age, ethnicity). Results from different imputation methods agree except for traditional data integration approach, whose coefficients and standard errors vary substantially.

Table 6.6. Coefficients (estimate, standard error, t statistic and p-values) for linear regression model from equation 6.1 after applying FCS in SLE data to solve combined content heterogeneity problems.

FCS	**estimate**	**standard error**	**t statistic**	**p-value**
(Intercept)	15.6709	3.4977	4.4804	0.0000***
Age	-0.1374	0.0303	-4.5395	0.0002***
Ethnicity				
Asian	-0.7194	3.3569	-0.2143	0.8306
Black or African American	-7.0475	3.5002	-2.0135	0.0466 .
Cape Coloured	-2.0421	5.5761	-0.3662	0.7151
Caucasian	-5.2502	3.5227	-1.4904	0.1394
East Indian	-1.0406	6.7519	-0.1541	0.8784
Eritrean	-7.6973	4.7869	-1.6080	0.1126
Hispanic	-0.9387	5.0923	-0.1843	0.8544
Mexican Mestizo	-1.1088	3.9314	-0.2820	0.7789
Middle Eastern	0.1794	5.0909	0.0352	0.9719
Mixed	-1.6365	4.9665	-0.3295	0.7426
Moroccan	0.5780	6.8938	0.0838	0.9336
Native American	0.0344	5.6953	0.0060	0.9952

Peruvian	-0.9578	6.5631	-0.1459	0.8846
Creatinine	1.9187	0.4471	4.2914	0.0000***
Gender				
Male	0.4283	0.5958	0.7188	0.4726
BMI	-0.0796	0.0321	-2.4812	0.0138*

*Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Table 6.7. Coefficients (estimate, standard error, t statistic and p-values) for linear regression model from equation 6.1 after applying FCSgroup in SLE data to solve combined content heterogeneity problems.

FCSgroup	**estimate**	**standard error**	**t statistic**	**p-value**
(Intercept)	15.8747	3.4962	4.5405	0.0000***
Age	-0.0985	0.0216	-4.5519	0.0000***
Ethnicity				
Asian	-2.6276	3.4009	-0.7726	0.4403
Black or African American	-6.5011	3.3889	-1.9184	0.0558
Cape Coloured	2.5737	5.9644	0.4315	0.6663
Caucasian	-4.9060	3.3736	-1.4542	0.1467
East Indian	-4.8222	4.6470	-1.0377	0.3003
Eritrean	-8.6377	3.5095	-2.4612	0.0142 .
Hispanic	0.2636	4.8983	0.0538	0.9572
Mexican Mestizo	-4.5726	3.4310	-1.3327	0.1834
Middle Eastern	2.5390	5.7418	0.4422	0.6587
Mixed	-2.2083	4.2658	-0.5177	0.6050
Moroccan	1.5953	5.6949	0.2801	0.7795
Native American	0.7154	4.5510	0.1572	0.8752
Peruvian	0.9829	5.7976	0.1695	0.8655
Creatinine	1.7781	0.4584	3.8789	0.0001***
Gender				
Male	0.3531	0.6299	0.5607	0.5753
BMI	-0.1325	0.0333	-3.9840	0.0001***

*Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Table 6.8. Coefficients (estimate, standard error, t statistic and p-values) for linear regression model from equation 6.1 after applying FCSgroups in SLE data to solve combined content heterogeneity problems.

FCSgroups	estimate	standard error	t statistic	p-value
(Intercept)	16.4658	3.5548	4.6320	0.0000***
Age	-0.1022	0.0258	-3.9582	0.0004***
Ethnicity				
Asian	-3.0366	3.4337	-0.8843	0.3769
Black or African American	-6.9395	3.4042	-2.0385	0.0420 .
Cape Coloured	1.3636	6.3171	0.2159	0.8294
Caucasian	-5.3452	3.3868	-1.5783	0.1151
East Indian	-4.8568	6.0353	-0.8047	0.4241
Eritrean	-8.8236	3.5534	-2.4831	0.0133 .
Hispanic	-0.0543	4.9299	-0.0110	0.9912
Mexican Mestizo	-5.0844	3.4913	-1.4563	0.1459
Middle Eastern	0.5645	5.8023	0.0973	0.9227
Mixed	-2.7119	4.5142	-0.6007	0.5488
Moroccan	0.4434	5.8237	0.0761	0.9393
Native American	0.6399	4.8482	0.1320	0.8951
Peruvian	-2.1492	6.3573	-0.3381	0.7367
Creatinine	1.7904	0.4665	3.8380	0.0001***
Gender				
Male	0.3332	0.6310	0.5280	0.5977
BMI	-0.1338	0.0337	-3.9748	0.0001***

*Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Table 6.9. Coefficients (estimate, standard error, t statistic and p-values) for linear regression model from equation 6.1 after applying FCS3group2 in SLE data to solve combined content heterogeneity problems.

FCS3group2	estimate	standard error	t statistic	p-value
(Intercept)	13.6279	3.2933	4.1381	0.0001***
Age	-0.1041	0.0200	-5.2043	0.0000***
Ethnicity				
Asian	0.7701	3.2518	0.2368	0.8132
Black or African American	-5.3824	3.2259	-1.6685	0.0974

Cape Coloured	0.3644	5.9764	0.0610	0.9517
Caucasian	-3.8681	3.1966	-1.2101	0.2282
East Indian	0.6654	5.5836	0.1192	0.9054
Eritrean	-5.8187	4.1777	-1.3928	0.1653
Hispanic	2.7048	4.4621	0.6062	0.5450
Mexican Mestizo	0.6828	3.3129	0.2061	0.8369
Middle Eastern	2.3303	6.2876	0.3706	0.7126
Mixed	-0.0277	4.5513	-0.0061	0.9952
Moroccan	1.9202	5.6577	0.3394	0.7349
Native American	2.6555	5.9393	0.4471	0.6556
Peruvian	3.3890	5.6728	0.5974	0.5507
Creatinine	1.8067	0.4449	4.0613	0.0001***
Gender				
Male	0.3935	0.5903	0.6666	0.5053
BMI	-0.0969	0.0306	-3.1627	0.0017**

*Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

6.5 Discussion

This work was motivated by a wish to extend multiple imputation for biomedical research to situations where content heterogeneity problems appear alongside each other when we stack multiple datasets and ask research questions. Therefore, we should be able to address combinations of content heterogeneity problems that co-exist within a single, stacked dataset. In this chapter, we addressed combinations of the varying granularity of a categorical variable, and a variable with mixed numeric and categorical data types. In this context, we proposed using full conditional specification approaches (multiple imputation) to solve combined content heterogeneities. The motivation is that this will allow appropriate borrowing of information across studies. We evaluated four imputation models (FCS, FCSgroup, FCSgroups, FCS3group2) through simulation studies and illustrated applications of the methods in SLE data.

In this section, we reflect on the utility of the methods to address a combination of content heterogeneity problems, their limitations, and discuss potential further research.

Summary of the main findings

Here, we described a new probabilistic approach for tackling granularity and mixed-type variable problems after data integration, from both a theoretical and a practical perspective.

Our initial assessment revealed very good results, and we see that our probabilistic method offered accurate results, and imputed data values were close to the real observed data. Based on this, we would argue that heterogeneous-type variables could be resolved to, without removing information; and multiple imputation could provide accurate results when answering research question(s).

Our probabilistic approaches differ in the way the categorical informative variable(s) used in imputation models. In the simulation studies, multiple imputations with chained equations with the inclusion of informative categorical variable(s) (FCSgroup and FCSgroups) produced estimates with no material bias for the linear models that we had data artificially introduced MCAR missingness. FCSgroup and FCSgroups produced good mSE and EmpSE, and confidence intervals close to 95% similar to reference true data. In very few cases - where we had different model errors between studies (scenarios 7-10) - the coverage probability was smaller than 93% suggesting than confidence intervals may have been conservative. FCS produced biased results in some scenarios which resulted FCS3group2 to have worse results than FCSgroup and FCSgroups. Our simulation studies showed broadly good performance for the FCSgroup, FCSgroups and FCS3group2 methods, a low impact on the different explored scenarios.

Overall, our results suggest that a probabilistic method such as multiple imputation by chained equations is useful for imputing complicated health data in which there is a combinations of content heterogeneity problems such as granularity and mixed type problems in datasets after integration. Specifically, in cases that our research question will be answered through linear regression models.

Imputation method including the extra informative variable(s) (FCSgroup and FCSgroups)

Traditional data integration approach could work in our advantage by introducing auxiliary variables in imputation model. Auxiliary variable is a variable that is used as a predictor in imputation but is not included as a predictor in regression analysis. The reason for their inclusion is because they may be correlated with the variables of interest or help retaining the missing process random [142]. The simulations results were very clear about our innovative idea to use the traditional approach, in the form of auxiliary variables that can help to make estimates on incomplete data, while they are not part of the main analysis to data integration as an additional step to algorithms' performance. The inclusion of the relevant informative categorical variable to the relevant imputation predictor (FCSgroup) and the inclusions of both categorical variables to both imputation predictors (FCSgroups) outperformed standard FCS. FCSgroup and FCSgroups methods performed equally well,

with negligible bias, similar empirical standard errors, appropriate model-based standard errors and appropriate coverage. In some cases, FCS3group produced worse estimates than FCSgroup and FCSgroup which showed that the default imputation method FCS, used in mixed type variable, resulted in underestimations. Hence, the usage of the extra informative variable(s) in the imputation model as predictors helped the imputations and improved imputed data classification. Including 'group' variable means in imputation models had serious impact on results.

Strengths

This study offers a great deal of strengths that should be mentioned. It offers a range of pragmatic alternative approaches to two common data integration problems in biomedical research.

To our knowledge, it has been the first study in a biomedical concept that describes the combinations of two content heterogeneity problems in detail, solved them with a theoretical solution, evaluated and compared traditional and probabilistic approaches using a series of simulation studies and illustrated the paradigm in real-world data. This work offers also an alternative to the work presented by Le Sueur et al. [129] where they showed traditional approach to solve different heterogeneities. In general, the advice is that imputation models should aim to include as many relevant variables as possible to make missingness assumption more plausible [143], [144].

This chapter also included evaluation of the general applicability of the method and compared results of the proposed integration techniques with gold standard methods through statistical simulation studies. In addition, the work has used real-world problems presented in MASTERPLANS project, tackled them and offered an alternative to current data integration approaches. Our research insists that we have gain in precision for all MI methods for both complete and incomplete variables. Results from most different imputation models agreed except for traditional approach whose coefficients and standard errors varied substantially.

It also offers an advancement in current MICE package where there is not the option of excluding chosen categorical levels in the imputation model. There is only the option of post processing of the imputations. The mice() function has an argument called post that takes a vector of strings of R commands. These commands are parsed and evaluated after the univariate imputation function returns, and thus provides a way of post-processing the imputed values while using the processed version in the imputation algorithm [145]. For example, we can squeeze the age integer conditionally on the age category. In the example

of our simulations, X_2 could be squeezed based on the categorical informative variable X'_2 levels: [min-mean], (mean-max]. When ranges limits difference is in decimals – so very small difference, postprocessing may not always be an option due to perfect prediction. For example, it may tend to impute values mainly in the common range limit - in our case mean 0. So, the imputation could be problematic because of perfect prediction [146] where automated procedures may give severely biased results [147]. Likelihood could tend to a limit as one or more regression parameters go to plus or minus infinity: loosely, these parameters have maximum likelihood estimate equal to plus or minus infinity [147].

Results show a clear usage of all the available information and superiority over complete case analysis. They gave larger point estimates than the newly introduced methods. The difference was not unexpected, since complete cases analysis made different assumptions from MI analysis. In the combined content heterogeneity problems, missing data have always occurred when information did not have same granularity across studies, so complete case analysis tended to exclude individuals under specific categories: this explains the differences seen.

Limitations

Similarly with other chapters, in applications, we need to remember that all of the techniques described so far assume that data are missing under MCAR assumption, and therefore, a proper sensitivity analysis to different assumptions should be considered.

Additionally, despite the exploration of different scenarios (number of imputations and iterations, number of studies, study sizes, mode error per study) that made our framework to be generalisable, we should consider the number of studies and individuals per study as a further problem. A large integrated dataset can make the computational time prohibitive. Similarly with previous chapter clustering could be included in multiple imputation especially in cases where there are many differences among integrated studies. However, in cases of only two studies there is not enough information available for random effect. Therefore, we cannot estimate higher-level coefficients that will be used in multilevel imputation.

In conclusion, the main reason for MI's growing popularity for handling missing data is its flexibility and practicality. In the case of imputing for solving content heterogeneity, we have described and evaluated a probabilistic approach to tackle granularity and mixed type problems. This extends current approaches, allowing multiple imputation when some variables have a substantial proportion of missing values in specific categories or are not collected in the same data type in all studies. We proposed four methods, but results showed

that only FCSgroup and FCSgroups easily take account of the potential heterogeneity in the imputation model, and they outperform the default FCS approach in most settings. FCS and FCS3group2 gave also good results but in complex cases (missingness in both studies – scenarios 1 and 2) may be struggling achieving good coverage. Therefore, including the informative variable(s) in the MI method appears to be a solid and efficient way. Our simulation evaluation of this approach and its application to real-world data showed promising results for implementation in health sciences.

Chapter 7: Discussion

This discussion offers an overview of the key findings of the thesis. It first reviews the overall rationale and objectives of the PhD, and then provides a summary of the four chapters with experimental results, with a focus on where the findings intersect. The discussion also highlights the contributions of our work to the integration of structured data in biomedical domain, implications for real-world translation**,** as well as the research strengths and limitations and implications for future research.

7.1 Introduction

The integration of data from different sources offers unique opportunities for health research; but poor standardisation across different datasets is often considered a major barrier to achieving this. The field of informatics has traditionally addressed this problem by defining universal information models and standards for coding (e.g., HL7v3 [24], Read [25], SNOMED [32]) aiming for perfect data-level standardisation when properly implemented, thereafter making data integration and analysis straightforward. To date, however, the successes of these efforts to pre-align data have been modest. For example, after ten years of development, HL7v3 was launched in 2006 but to date most healthcare messaging is still carried out with older version of HL7 that are semantically weak [31]. This strategy might result in waste of time, resources and costs of efforts to define coding standards and in integrating information models.

Existing methods for dataset integration rely on mapping to common data models, often resulting in a substantial loss of information that is present in the source datasets. For instance, in case of severe content heterogeneity, integrated datasets often cover just a small set of data items that are present in most source datasets and leave out the other items which impairs performance of the final model and reduces sample size and power [148]. But these will obviously diminish the potential to accurately answer research questions since it will ultimately reduce possibilities to harmonise cohort selection and confounder adjustment.

Ideally, shared data models would be implemented at the source, enabling uniform data collection at different sites and studies. But in reality, data integration is commonly, and likely to continue to be, imperfect. Our approach to address data integration and harmonisation embraces this imperfection rather than trying to extinguish it.

In this thesis we have argued for an alternative focus on the development of analytic methods that embrace the data variation and heterogeneity, rely less on standardised data items and have the capacity to process content heterogeneity in data. We believe that this is a viable

alternative − and much can probably be adapted from existing methods for multiple imputation. Our proposal is characterised by a probabilistic rather than a functional view of integration; by post-alignment rather than pre-alignment of data sources; and by a pragmatic, top-down approach to answering research questions rather than a laborious, bottom-up approach to data integration.

The overall scope of this thesis was to explore the potential of probabilistic methods, that are based on multiple imputation, to address content heterogeneity in biomedical dataset integration. We argued that the existing methodologies and data standards have significant limitations, and that health informatics should focus on developing methods for quantifying the uncertainty introduced by imperfect integration of existing sources, rather than striving for perfect, data-level integration. One potential and promising alternative relies on probabilistic methodologies.

This thesis has provided insights into probabilistic ways to overcome data integration issues due to systematically missing values, varying levels of granularity, similar variables with different data types, and combinations of those problems. Each content heterogeneity problem described in the previous chapters focused on a different aspect of the thesis' overall aim, and the findings were explained, discussed, and concluded in detail in each respective chapter. The current chapter will provide a general discussion about main findings of the studies, limitations and future steps. Data from real-world SLE studies were used to formulate concrete examples and illustrate our methods' applications. Simulated data were used to evaluate our methodologies that achieve successful data integration. This methodology and results presented in this thesis may be generalised to other data sources and types and are not limited to only biomedical sources. In respect of our overarching aim and objectives, this thesis has achieved to:

i. *Describe representational and content heterogeneity and identify potential issues that make structured data integration difficult*. In section 3.2 we have described the problem of systematically missing values, in section 4.2 the problem of varying levels of granularity, in section 5.2 the problem of mixed numeric and non-numeric data types, and in section 6.1 we explored combinations of granularity problems and mixed data types.
ii. *Build pragmatic probabilistic approaches to successful healthcare data integration*. Methodologies with theoretical solutions were presented in sections 3.3, 4.3, 5.3, and 6.2.
iii. *Evaluate the accuracy of statistical inference based on the developed probabilistic methodologies via simulation studies under complex scenarios*. We performed

simulation studies to study the empirical properties of the proposed solution, under ten scenarios per content heterogeneity problem. Simulation designs, results and discussions can be found in sections 3.4, 4.4, 5.4, and 6.3. We also evaluated our method to solve mixed type data with a resampling method of SLE data analysis in simulation studies in section 5.5.

i. *Illustrate methodologies' application to real world data's integration problems*. We demonstrated that it is feasible to apply our method to real-world data in sections 3.5, 4.5, 5.5, and 6.4, using the MASTERPLANS exemplar.

7.2 Summary of Results

Our primary aim was to develop an approach to creating an integrated dataset that included all the information from the constituent datasets, with a clear understanding that any integrated dataset would have many missing data due to content heterogeneity's different forms. In chapters 1 and 2 we introduced the problem of content heterogeneity of different kinds, and described in detail the gaps in literature, why this research work is important and how content heterogeneity is connected with missing data. Our main conclusion is that waiting for all data sources to be interoperable is hindering for opportunities to study human health; and the implementation of probabilistic approaches to integration of structured data can offer a great advancement in biomedical research.

In all simulation studies, we had a *data generating model*, a *reference model* that was estimated on the Full Data sample before introducing content heterogeneity, a *complete case analysis model* that was estimated on complete data after solving content heterogeneity with traditional approach, *probabilistic imputation model* that was estimated on imputed data after combined using Rubin's rules [78], solving content heterogeneity with a probabilistc appoach.

In Chapter 3 we presented the initial results of our attempt to solve the first content heterogeneity problem of systematically missing data. As an initial assessment of multiple imputation techniques, we applied, evaluated, and examined the utility of FCS with respect to simulated datasets and also applied it to the real-world lupus datasets. The conventional approach that uses post-alignment of dissimilar datasets was used as a comparison - excluding systematically missing variables from analysis. We showed that it led to biased estimates of regression coefficients, while FCS gave accurate coefficient estimates with very good coverage of the associated confidence interval (94-95.5%). We found that when

imputing with 500 and 1000 individuals per study, additional variability resulted in smaller standard errors (Scenario 9: 0.052, Scenario 10: 0.036) and mild under-coverage (Scenario 9: 88.4%, Scenario 10: 84.7%). When the model error is e_i: N ~ (0, 20), we observed that a relatively small part of the observed variation in outcomes was explained by the independent variables. In all scenarios, coverage rate was equal and slightly exceeded the nominal rate (94.5-95.2%) and the estimated regression coefficients were almost identical to those obtained from the original data in which there was no content heterogeneity. Overall, our results show that statistical inference on incomplete data, due to systematically missing values, that were imputed by regression imputation can produce the correct answer.

In Chapter 4, we introduced the problem of varying categorical levels of granularity after structured data integration. The traditional approach maps variables to a common data model removing granularity form the data. It was used as a comparison to our methods. We evaluated FCS using simulated datasets and afterwards we applied it to the real-world lupus datasets. We also introduced an additional imputation model, FCSgroup that the only difference with FCS is that it includes the group variable in the imputation model. Our motivation was that imputation model could include additional 'auxiliary' variables. In general, auxiliary variables are used in the imputation model but not in the regression model and they can improve imputations [108], [143]. In their paper, Hardt, Herke and Leonhart [142] concluded that when estimating a linear regression coefficient, the inclusion of auxiliary variables is most helpful when the correlations between the continuous correlated X's and outcome Y are high ($r \geq 50\%$). The only difference between FCS and FCSgroup is that the latter has one additional variable ('group') included in the imputation model. Our 'group' variable acts as an auxiliary variable and in some scenarios improved the FCS imputations. The simulation results were similar to chapter 3. Analyses of datasets imputed with FCS imputation model gave good results both in terms of bias, precision and confidence interval coverage.

We suggest that FCSgroup should be preferred as an imputation method as it gives at least as good results with FCS and it achieves better classification is some scenarios. The general suggestion is that the more information you use in imputation the more accurate the results will be. So, when there is some flexibility about computational time, FCSgroup should be the priority probabilistic method. Both probabilistic ways are preferred to solve granularity over complete case analysis/Complete Records.

In Chapter 5, we focused on the issue of mixed numeric and non-numeric data types. As in chapter 4, we also introduced the imputation model FCSgroup that uses a categorical variable that can help to make estimates on incomplete data, while they are not part of the

main regression analysis. Our probabilistic approaches FCS and FCSgroup had the same logic as in Chapter 4. FCSgroup included a categorical 'group' variable with the non-numeric data types of each individual. The simulation results provided useful insight to understanding if our probabilistic methods offer indeed accurate results in solving the issue of mixed data type. We designed two types of simulation studies to evaluate our designed methods to solve this type of content heterogeneity issue. Data were generated (i) by producing parametric draws from a known model (many times), and (ii) by repeated resampling with replacement from the MASTERPLANS data (where the true data-generating model was unknown) [112]. Parametric simulation studies used similar generating mechanisms with the other chapters. When we introduced different model errors, FCSgroup had good coverage (95%) and no bias (~0.001) for the four study sizes (100, 200, 500, 1000 individuals per study). FCS's coverage started with 90% when we had 100 individuals per study and slowly reduced as the study size increased. As we said before, we illustrated and evaluated our probabilistic approaches in the MASTERPLANS data. *Age* had no missing data, its type was integer and had complete data which allowed us to compare the imputed data against real raw data. Therefore, we applied content heterogeneity due to mixed type in *Age* variable. We concluded that FCSgroup was unbiased and identical to the reference model (Full Data sample). FCS results also showed small bias in estimates but good coverage for Age. It seems that in FCS, some values may have been imputed outside the range of possible values, this method had a low bias and estimated the within- and between- imputation variance adequately, resulting in generally good coverage across repeated sampling of the missingness. The increasing number of imputed datasets may have helped improving -already good- coverage but produced similar estimates.

We conclude that our parametric simulations did not show any difference between FCS and FCSgroup but in resampling simulation FCSgroup resulted in better estimates. Results show that the imputed data with FCS were close to true Full Data. Similarly, data were correctly imputed under FCSgroup imputation model. Thus, in terms of computational time, researchers may prefer to rely on FCS as FCSgroup takes to some degree more time when performing imputations and analyses. However, the inclusion of informative 'group' variable is suggested in cases that there are complex scenarios. In summary, we could argue that mixed type variable problem could be resolved through probabilistic ways to produce homogeneous type variable, without removing useful information.

In Chapter 6 we present the evaluation of probabilistic methods, and application when two content heterogeneity problems are present in healthcare data in two variables, i.e, the problem of granularity and mixed type variable. To examine our idea about the auxiliary variables, we introduced two categorical 'group' variables - one for each content

heterogeneous variable. Here, we examined four imputation models (FCS, FCSgroup, FCSgroup**s**, and FCS3group2) varying in inclusion/exclusion of 'group' variable(s). We observed that there was no bias in either of the two variables when imputing with FCSgroup, and FCSgroups. Both content heterogeneity problems achieved good coverage level when imputing with FCSgroup and FCSgroups. FCS3group2 was similar to FCS in some scenarios but it showed improvement which means that 'group' variable helped the algorithm perform better. FCS should not be preferred when we have 2 datasets with 200 (scenario 1) or 1000 (scenario 3) individuals per study. In those two scenarios, X_3's estimates were biased which resulted in low coverage ~ 84%. An explanation for FCS (scenario 1 and 3) is that there were two integrated studies therefore data missingness in both, so there was missing information that could help imputation's accuracy. As the proportion of missing data increases, FCS may need more imputations and iterations [149] to achieve accurate estimates. In scenarios 3-6, where we have at least 5 datasets with same and different study sizes, FCS and FCS3group2 could also be chosen as imputation models. In scenarios 7-10 In conclusion, FCSgroup, FCSgroups gave unbiased estimates and good coverage (95%) when imputing missing data in both variables X_2 and X_3. Results proved that including 'group' variable in imputation improved X_3's estimates (FCS3group2). All imputation models outperformed the use of Complete Records.

If we bring all the scenarios together for all the results chapters, some significant findings emerge.

When studies have the same size, FCS gives accurate results in the three types of content heterogeneity problems. When granularity and mixed type are combined then FCS should not be preferred as an imputation method for categorical granular variable when we have two datasets. FCSgroup showed accuracy in granularity and mixed type problem, and when these two are combined. FCSgroups also showed very good results, and outperformed FCSgroup when we had larger study sample, 1000 individuals per study. FCS3group2 performed very well in the mixed type but similar to FCS for granularity problem. In general, 'group' variable helps improving imputation model's accuracy, data classification and coverage level.

When studies are of different size, FCS gave accurate results in both scenarios in the three content heterogeneity problems. FCSgroup performed very good and similar to true Full Data. We observe that when we had 10 datasets with different sizes (scenario 6) the bias was smaller than when we had 5 datasets with different sizes (scenario 5). Coverage probability of confidence intervals was very good in scenarios 5 and 6. mSE and EmpSE was smaller in scenario 6 than in scenario 5.

With regards to the *size of the model error,* in all cases the larger the model error per study, the larger mSE and EmpSE which made sense as all imptuation methods followed Full Data standard errors. When outcome Y had a big range (e_i: N ~ (0, 20)), missing data's estimates were getting slightly worse in a few scenarios (mainly section 6.3.3).

When studies have different model errors, in all chapters we saw that as the number of individuals increased the mSE and EmpSE decreased. Bias was close to zero and coverage level is very good in most cases for FCS and FCSgroup (FCS was not always precise - combined types of heterogeneity chapter). Imputation models did not always achieve coverage level greater than 90% as the number of individuals went above 500 per study.

7.3 Novelty of this work

While some applied studies have applied multiple imputation to resolve content heterogeneity problems [97], [107], [121], [140], [150], [151], this thesis presents the first systematic assessment of the accuracy of this approach in different scenarios.

Variable and data integration across multiple datasets is important because it can provide more statistical power, more detailed models for general application, as well as maximising the study population size [152]. Our results showed that even when we have systematically missing data after integrating two studies, FCS's estimates perform better than classical methods.

Our probabilistic approach is not specific to data source or a specific disease or condition. It is flexible and automated enough so to benefit significantly other future research projects in health/biomedicine domain like bioinformatics [153], molecular epidemiology [154], and other research fields like environment [150], earth science [155], food sector [156]. Many of the steps carried out are already used to solve a very common problem - missing data - and researchers implement it already in numerous biomedical projects [85], [157]–[159]. Therefore, the suggested approaches are known to many researchers and do not need expertise and extensive training. Current pipelines could include our approach, so they answer research questions while achieving content homogeneity.

Another advantage of this work is that from the start, it focused on existing, real-world problems (using SLE data) with a clear impact. We followed a deep methodological investigation and designed simulation studies that were compared to the gold standard and widely accepted approach for integration. Consequently, the approaches that have presented are complete and could be implemented in real world data.

In addition, we explored many scenarios with variability in the size of the datasets/studies (i.e. 80 to 1000 number of individuals per study); varying model error per study (e_1: N ~ (0, 1.2) and e_2: N ~ (0, 1.3); varying model error per scenario (e_i: N ~ (0, 0.2), e_i: N ~ (0, 2) e_i: N ~ (0, 20)); different number of datasets (2, 5, 10). A sample of 1,000 might seem too large if compared against trial data, but it was a necessity if we were to investigate the existence of content heterogeneity problems in more than one study dataset and therefore high rates of missingness.

Another application of the presented methods is in prediction models and particularly their external evaluation. Prediction models should externally be validated before implementation in clinical practice. However, uncollected variables that missing from suitable validation cohorts are the main reason making external validation impossible [160]. Hence, results of this thesis may facilitate the use of cohort studies that do not include all predictors in a prediction model.

Additionally, our framework could facilitate EHR databases' integration and analyses. As we discussed in Chapter 1, most EHR databases are in OAV format. As we cannot analyse data in OAV format (e.g using regression methods), we first need to transform the OAV database into a conventional tabular format where each column is a different variable [161]. After the OAV database has been transformed into a tabular format, it will often emerge that there are missing values for some patients. Our probabilistic data integration method cannot be applied to databases in OAV format. But if we first transform the OAV databases to tabular format (which researchers already need to do) [162], then we can apply our method. This will be exactly the same as for other databases. For example, one EHR may have never recorded 'smoking status' while another EHR has recorded it. Or they both recorded it, but at different levels of granularity. So, all the aforementioned content heterogeneity problems may occur, and we can address them once the data are in tabular format.

When standards are not implemented at the source, in order to achieve integration and harmonisation, data are aligned by hand matching up equivalent or similar patient variables across the different studies as shown in the published article by Le Sueur et al. [129]. Harmonising data across studies could be complex and time and resource consuming. However, our probabilistic approach overcomes many challenges arising both form inherent complexity of the disease and data differences in terms of capture and representation across different studies. Therefore, our presented research highlights and addresses an important gap in the current data science toolkit, and we believe that our methods are very promising to improve efficiency and avoid analysts' bias when choosing how to deal with the common data harmonisation issues.

Overall, our approach is flexible, enabling the inclusion of additional new variables and new datasets. It means that the resulting output could be updated easily to answer any new research question. We also showed flexibility concerning study size, model errors, and number of studies using a small number of variables.

Addressing content heterogeneity problems using multiple imputation

Multiple imputation is considered as the most popular method for handling missing data in practice and is very easy to use [163]. Multiple imputation has recently been suggested as a method for individual patient data meta-analysis in cases of complete missing variables and heterogeneity among studies [97]. Methods to handle systematically missing data have been proposed for continuous data [122] and recently extended to binary and discrete data [121]. Resche-Rigon et al. [122] proposed a one-stage approach for imputing systematically missing continuous predictors in IPD meta-analysis. Their approach followed multilevel models with random intercept terms and random slopes to account for heterogeneity across IPD study datasets. When standard errors are used as as measure of uncertainty around between-study covariance parameters this may be demanding as uncertainty lean on being heavily skewed.

Jolani et al. [121] research paper was about a generalised approach using MICE imputation of systematically missing predictors in IPD meta-analysis. A generalised linear mixed model that allows for between-study heterogeneity. In terms of simulation setting they had similar setting to ours. Simulations had either 6 or 15 studies, 500 individuals per study leading to a total sample size of 3000 (N=6) and 6500 (N=13). However, they considered two main patterns i.e. a univariate pattern where one specific variable was missing (as we did in chapter 3), and a bivariate pattern where either variable, between two, was systematically missing. The total data missingness was either 20% or 50%. They compared four methods, complete case analysis, traditional multiple imputation (FCS), stratified muliple imputation and their method multilevel multiple imputation in MICE. Multilevel outperformed the other methods and its coverage rates were 90% whereas FCS's was 85%. Multilevel multiple imputation required, however, considerable more computation time to generate an imputed dataset as compared with FCS. Their approach needed significant computational power 100–150 times slower as compared with FCS.

Moreover, Quartagno and Carpenter [97] developed and evaluated a joint modelling approach to multiple imputation of IPD meta-analysis, with an across-study probability distribution for the study specific covariance matrices. Their analysis were through R package jomo [104]. They explored different scenarios and simulation results were very promising.

Resche-Rigon and White [107] proposed a MICE algorithm for multilevel data with arbitrary patterns of systematically and sporadically missing variables. They suggested two methods for imputing a single incomplete variable: an extension of an existing method and a new two-stage method which allows for heteroscedastic data. The total number of patients was fixed at 2000. The number of studies was 20, 100 individuals per study. Their simulation studies showed that with heteroscedastic models their two-stage methods outperformed the one-stage method in some scenarios. They also suggested for future work, an important extension of their imputation model would be to impute categorical variables.

In the context of IPD meta analysis, Burgess et al. [164] concluded in their research that in most cases the best approach is two-stage analysis (impute separately in each study, apply Rubin's rules to get a single summary for each study and perform meta analysis). Although, a one-stage analysis has the potential advantage of allowing us to borrow strength across studies, which may be important for estimation of covariate and subgroup effects. Our methodology allows to see the data as a whole and share information across the integrated dataset. An advancement is that we showed that is works in cases where we have only a few datasets integrated, even if they are just two.

It is clear that the issue of systematically missing values may be resolved to produce accurate results through multiple imputation methods. Researchers started exploring this issue recently [97], [107], [121], [165]. The application of multiple imputation to solve systematically misisng values is also demonstrated by the feasibility of our proposed framework. Based on results shown in chapters 3-6, we could argue that content heterogeneity could be resolved through multiple imputation and could provide 'close to reality' results that will enable us answer research question(s). Our plan was to make the models complex to allow for more flexibility to better answer any potential research questions based on the available information.

Content heterogeneity expressed as a granularity problem (varying categories among datasets) is one of the most common limitations in current frameworks as there is inevitable loss of information. Our study proposes multiple imputation techniques to solve granularity issues in structured data. A common problem that results from choosing the least granular (fine detailed) data is that we may choose specific groups of patients resulting in selection biases. An example (as shown in chapters 4 and 6) is about ethnicity. Previously published works generally agree that the more granular race, ethnicity, and demographic data in health sciences, the better the identification of at-risk populations [132]. However, the scientific community still struggles to organise granular data from multiple studies, subgroups, and diverse communities [166]. Moreover, data collection and data integration (i.e. EHRs and

genomic data) from different sources are essential as they enable the study of rare diseases like lymphoma due to adequate data granularity and volume [109].

To our knowledge, no published study to date has considered imputation as a method to overcome data heterogeneity among studies when a variable presented in mixed type i.e. categorical and integer. As we mentioned in chapter 5, Kalter et al [139] argue that a variable's type harmonisation should occur when age is as continuous in a study and as categorical in a new study. Then all the previously data from the other studies should be converted to categorical data in order to obtain homogeneous datasets for further analysis.

Data harmonisation extents the utility of existing datasets [167] and increases possibilities for multi-centre collaborations [12], [168], [169]. Research community recognises the need to explore relationship between disease's biomarkers and how disease progresses clinically [170]. These relationships are affected by different factors such as genetic, demographic, environmental, therefore data integration is essential. This need grows in cases where factors are uncounted in some studies, only present in small proportions of the population or have dissimilar levels. One way to achieve additional power from existing data would be to combine data from existing studies. In addition to increased statistical power, combining data across studies should also decrease bias due to sampling error, and improve the external validity of findings [171].

Another advantage is the computation time. The setting of our simulation studies allowed to eliminate computational time to hours (ranging from 1 to 12 hours). Moreover, simulations showed that our methods were unbiased under a small number of imputations and iterations (5 each). We also supported this claim with resampling simulation of SLE data (chapter 5) where we showed that there was no significant difference among 5,10,15,20 imputations. Estimates' coefficients were very similar, there was only some improvement in coverage level which was varying between 94-96% in all imputation models for all imputations.

Traditionally, representational heterogeneity in multi-dataset analyses has been resolved through post-hoc alignment of variables by mapping them to a common data model that is consistent with each of the individual study data models. Such common data models tend to achieve that consistency by removing granularity from the data. We have shown that this approach could lead to biased regression coefficients and should therefore better be avoided. However, we have also found that incorporating this approach in probabilistic data integration methods can be beneficial for the quality of the imputed values, and thus for the accuracy of estimated regression coefficients. We therefore recommend researchers to always apply probabilistic data integration, but also to incorporate the traditional common data model approach to optimise the quality if imputed values. This thesis shows that

sometimes a simple translation of a problem to another more manageable problem can enable the use established knowledge to solve it.

Implications for real-world translation

Methods that quantify the uncertainty could be more productive compared to the ones that strive for perfect data harmonisation. This thesis has tried to make it clear using real-world example from SLE cohort studies. MASTERPLANS data have helped us find examples of representational heterogeneity to devise our hypothesis, develop our approaches and construct our proposed frameworks for data integration and harmonisation while solving content heterogeneity problems. We later focused on utilising these data to test, examine and improve the different steps in our pipeline and to transform it into a generalisable methodology that will have applicability in many different and complex biomedical datasets.

A significant advantage to this research work was that it embedded real world data currently being utilised by other ongoing, large, projects – and as such, providing us with significant opportunities to work on, enhance and improve our approaches. Significant comparison of our work with the traditional data integration approach and how useful our alternative is, was presented in the paper by Le Sueur et al [129]. Our work offers solutions to some limitations of the traditional data integration framework approach presented by Le Sueur et al [129]. For example, they removed rows with missing data, so patients were excluded if they were missing key information. Additionally, their method led to inevitable loss of information either due to differences in granularity or data capture; and the final patient group may have suffered from selection biases. And finally, they excluded patients and rows of data based on missingness which reduced sample sizes and introduced potential biases.

The research community could argue that to overcome systematically missing variables, granularity and mixed type problems, traditional data integration approach could be preferred as it does not introduce important bias and has been the common technique for many years. However, our illustrations using SLE do not agree with that. The results indicate that there is a difference between real data and complete case analysis. In addition, results show a practically useful gain in information over Complete Records with multiple imputation.

In conclusion, our approach has a number of essential ingredients. First, it predicates upon the belief that data heterogeneity will persist in the foreseeable future despite standardisation efforts, and the approach therefore strives for post-alignment rather than pre-alignment of big datasets. Second, as post-alignment of heterogeneous data sources will often be

imperfect, a probabilistic rather than functional (deterministic) view of semantic equivalence is adopted in which semantic equivalence is a matter of uncertainty rather than a yes/no answer. In this perspective, it is not a problem that two data items are not semantically equivalent, as long as we can estimate the probability that they are. Third, the approach is extremely pragmatic in the sense that it will always provide an answer − although it might not be better than a flat prior in the very worst case where none of the source datasets is found to provide useful information to answer the research question. The other side of the same coin is that full, data-level standardisation is another special case (the very best case), in which the results should then be the same as those that would result from analysing an integrated dataset.

7.4 Limitations

We have identified six main limitations that apply to our work.

First, we have focused entirely on content heterogeneity and have assumed throughout that other types of representational heterogeneity had been resolved beforehand. In practice it might be possible, and more efficient, to resolve different types of representational heterogeneity in one step rather than in separate steps, but this was beyond the scope of this thesis.

Second (as mentioned briefly in chapter 5), based on our hypotheses, which were common for all the results chapters, we assumed that we had more than one study datasets and there was no need for record linkage. We also assumed that the observations within each study dataset and in the integrated dataset were statistically independent and identically distributed. We also made sure that this was the case in all simulation experiments, but one could argue that this assumption is not entirely realistic in real-world data, and that it is plausible that the observations within study single dataset are to some degree correlated, because they came from the same institution or the same geographical location. It would be possible to relax this assumption with multilevel regression models but this was beyond the scope of this thesis. In this scenario one could also consider using multilevel imputation. However, while multilevel imputation can be used for sporadically missing data, it does nor when there are systematically missing data: there is then not sufficient information to estimate higher-level regression coefficients. So, in theory this would be better, but because of the nature of the problem that we are trying to solve (systematically, not sporadically, missing data), we cannot use it. Further research is needed to address this methodological challenge.

Third, another addition to our current work would be investigating situations where patients do not have identical distributions. For example, let's say we integrate data from two different data sources: EHRs and trial data. We determine the expected benefit of using a drug in a patient, or set of patients, is certain ages, and under specific risk factors, with or without a specific biomarker. The observed effect of the drug is subject to confounding in the EHR data but not in the trial data. The two institutions that started using the treatment routinely likely gave it only to patients they believed would benefit from it, and not to others. As a result, the groups of treated and untreated patients are unbalanced in the EHR data, and comparing their health outcomes without correcting for this imbalance would give an unrealistic picture of the effect of the drug. In contrast, in the trial dataset, the groups of treated and untreated patients are perfectly balanced by design. To obtain an unbiased estimate of the treatment effect from the EHR data, the analysis has to adjust for potential confounding factors.

Additionally, computational time led us to select our largest simulated dataset to include 5,000 records. Unfortunately, this is not necessarily representative of a contemporary EHRs dataset which can hold hundreds of thousands or millions of cases. However, even that limited size is very different to the size of a clinical trial, on which multiple imputation methods have been routinely evaluated in the past. Therefore, we argue that we manage to provide an incomplete view on the relevance of these methods in larger datasets and that suggested probabilistic approaches may be less relevant to large health informatics databases than to randomised clinical trials or cohort studies.

Although realistic, our simulated scenarios cannot be extensive, and results may differ in other scenarios with different hypothesised associations between exposure, covariate and outcome and different distributions. Nevertheless, we would expect the methods to perform similarly, and our conclusions not to be affected.

In addition to that, both evaluations and illustrations did not explore situations where there is existing data missingness in some variables and data missingness in outcome. Our simulation scenarios explored content heterogeneity problems in exposure and covariates, but not in outcome. This would be an area where the current work could be extended to.

Unanswered questions and future research

Our work's imputation approaches were based on FCS, and we only compared with traditional data integration approach complete case analysis – mapping to common levels and omitting variables. Future steps may include comparing FCS with other missing data techniques regression imputation, mean substitution, last observation carried forward, joint

modelling, and machine learning techniques (random forest imputation [172], k nearest neighbour classification [173], support vector machine [174] etc).

An outstanding content heterogeneity problem that this thesis has not addressed is the *overlapping ranges* problem. Let's consider that in dataset 1, patients' age is represented as a categorical variable with ranges '0-20', '21-40', '41-60', '61+', and in dataset 2 age is represented as a categorical variable with ranges '0-15', '16-25', '26-40', '41-60' '61+'. Finding a common denominator between the same variables is not always a probable solution when integrating structured data. The integrated dataset should always be specified at the lowest level of granularity (e.g. '0-15', '16-21', '22-25', '26-40', '41-60' '61+'. Some patients will be included in at least one age range. For those that we are not sure in which range they belong, we leave age as missing. Based on our type of solution (i.e. the probabilistic data integration via multiple imputation techniques), we solve data missingness and fit the regression model on the dataset with imputed values.

As mentioned in limitations, patients from a single centre/study are more likely to be similar than patients from different centres/studies. Clustering is apparent from many multi-centre studies and meta-analyses. It is probably due to variation across centres/studies that arises from residual, unmeasured confounding. Further development to our methods is to account for clustering is using multi-level regression analysis [175] in which random effects are used to model centre-level and study-level variation.

One other possibility is the implementation of our methodology to larger real-world datasets containing biomedical data such as Asthma e-lab [176] or EHRs. Research has not been explored in situations like EHRs where there are potentially millions of records.

Additionally, our approach comprises a number of critical choices (such as the choice of predictive modelling algorithms [141] , number of study datasets, choice imputation models, number of imputed datasets, number of iterations etc) that require thorough methodological investigation. It also relies on assumptions (such as reasons for missingness) that can influence the end result and should therefore be further investigated through sensitivity analyses when it is applied in practice.

Another interesting step could be the creation of a flexible, user-friendly tool to achieve data integration using our methods to perform imputation and analysis simultaneously. It could impute any type of outcome and confounders, systematically and sporadically missing data, granularity, mixed type and combined problems. It could be a freely available tool that may be utilised easily with the development of novel statistical software (R package, Rshiny) that should not require knowledge of the statistical principles of multiple imputation [77].

Prospective alignment of legacy datasets without the availability of standards is labour intensive and very often impossible to achieve perfectly. In such cases, very little practical guidance is available and data integration frameworks about real-world data have not been easily shared to use as guidance. If researchers published healthcare dataset integration steps they followed during analyses, many problems would have been identified and therefore errors and biases would have been avoided. It would also be extremely beneficial for young professionals in health data science to have some guidance. If specific steps were taken during data cleaning/preparation/manipulation processes, and they were also shared then these traditional techniques would have been questioned, analysed and probably been improved much earlier.

7.5 Conclusion

This thesis aimed to identify issues that make structured data integration ambitious and overcome these by using promising alternative techniques. We defined three content heterogeneity problems i.e., systematically missing values, varying granularity, mixed type variables, and solved these theoretically with our novel approach. We designed a series of simulation studies, under a range of scenarios, to apply those problems (and their combinations), so to evaluate our methods.

We conclude that multiple imputation by chained equations is a valid approach to solve content heterogeneity problems. We found that in more complex scenarios the usage of a 'group' informative variable facilitated and improved imputations. Our approaches consistently outperformed traditional approaches - complete case analysis. Results of this thesis may facilitate structured data integration in biomedical research and pave the way for more widespread usage of probabilistic approaches to overcome other content heterogeneity problems. Translating content heterogeneity into a missing data problem and solving it using established methods is a very promising solution and could help leverage disparate datasets to gain knowledge and answer critical biomedical research questions.

References

[1] T. B. Murdoch and A. S. Detsky, “The inevitable application of big data to health care.,” *JAMA*, vol. 309, no. 13, pp. 1351–1352, Apr. 2013, doi: 10.1001/jama.2013.393.

[2] S. Schneeweiss, “Learning from big health care data.,” *N. Engl. J. Med.*, vol. 370, no. 23, pp. 2161–2163, Jun. 2014, doi: 10.1056/NEJMp1401111.

[3] R. Bellazzi, “Big Data and Biomedical Informatics: A Challenging Opportunity,” *IMIA Yearb.*, vol. 9, no. 1, pp. 8–13, 2014, doi: 10.15265/IY-2014-0024.

[4] F. Martin-Sanchez and K. Verspoor, “Big Data in Medicine Is Driving Big Changes,” *IMIA Yearb.*, vol. 9, no. 1, pp. 14–20, 2014, doi: 10.15265/IY-2014-0020.

[5] M. J. Khoury, “The case for a global human genome epidemiology initiative,” *Nature Genetics*, vol. 36, no. 10. Nat Genet, pp. 1027–1029, 2004, doi: 10.1038/ng1004-1027.

[6] M. Noale *et al.*, “Predictors of mortality: an international comparison of socio-demographic and health characteristics from six longitudinal studies on aging: the CLESA project,” *Exp. Gerontol.*, vol. 40, no. 1–2, pp. 89–99, Jan. 2005, doi: 10.1016/j.exger.2004.09.003.

[7] L. Serra-Majem *et al.*, “Comparative analysis of nutrition data from national, household, and individual levels: results from a WHO-CINDI collaborative project in Canada, Finland, Poland, and Spain*,” *J. Epidemiol. Community Health*, vol. 57, no. 1, p. 74, Jan. 2003, doi: 10.1136/JECH.57.1.74.

[8] P. A. Bath, D. Deeg, and J. Poppelaars, “The harmonisation of longitudinal data: A case study using data from cohort studies in the Netherlands and the United Kingdom,” *Ageing Soc.*, vol. 30, no. 8, pp. 1419–1437, Nov. 2010, doi: 10.1017/S0144686X1000070X.

[9] A. B. Holmes, A. Hawson, F. Liu, C. Friedman, H. Khiabanian, and R. Rabadan, “Discovering disease associations by integrating electronic clinical data and medical literature,” *PLoS One*, vol. 6, no. 6, 2011, doi: 10.1371/journal.pone.0021132.

[10] S. A. Sansone *et al.*, “Toward interoperable bioscience data,” *Nat. Genet.*, vol. 44, no. 2, p. 121, Feb. 2012, doi: 10.1038/NG.1054.

[11] P. A. Schad, L. R. Mobley, and C. M. Hamilton, “Building a Biomedical Cyberinfrastructure for Collaborative Research,” *Am. J. Prev. Med.*, vol. 40, no. 5, pp. S144–S150, May 2011, doi: 10.1016/J.AMEPRE.2011.01.018.

[12] D. Seminara *et al.*, “The Emergence of Networks in Human Genome Epidemiology: ‘Challenges and Opportunities,’” *Epidemiology*, vol. 18, no. 1, pp. 1–8, Mar. 2007, [Online]. Available: http://www.jstor.org/stable/20486309.

[13] J. Kaye, “From single biobanks to international networks: developing e-governance,” *Hum. Genet. 2011 1303*, vol. 130, no. 3, pp. 377–382, Jul. 2011, doi: 10.1007/S00439-011-1063-0.

[14] B. M. Knoppers *et al.*, “Towards a data sharing Code of Conduct for international genomic research,” *Genome Med.*, vol. 3, no. 7, pp. 1–4, Jul. 2011, doi: 10.1186/GM262/COMMENTS.

[15] D. Doiron *et al.*, “Data harmonization and federated analysis of population-based studies: the BioSHaRE project,” 2013. [Online]. Available: www.bioshare.eu.

[16] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, “FEDERATED OPTIMIZATION IN HETEROGENEOUS NETWORKS,” 2020.

[17] W. Sujansky, “Heterogeneous Database Integration in Biomedicine,” *J. Biomed. Inform.*, vol. 34, no. 4, pp. 285–298, 2001, doi: 10.1006/jbin.2001.1024.

[18] I. S. Kohane *et al.*, “Sharing electronic medical records across multiple heterogeneous and competing institutions.,” *Proc. AMIA Annu. Fall Symp.*, pp. 608–612, 1996, [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2233175/.

[19] P. E. Warming *et al.*, “Harmonization of the definition of sudden cardiac death in longitudinal cohorts of the European Sudden Cardiac Arrest network – towards Prevention, Education, and New Effective Treatments (ESCAPE-NET) consortium,” *Am. Heart J.*, vol. 245, pp. 117–125, Mar. 2022, doi: 10.1016/J.AHJ.2021.12.008.

[20] J. S. Ancker *et al.*, “How is the electronic health record being used? Use of EHR data to assess physician-level variability in technology use,” *J. Am. Med. Informatics Assoc.*, vol. 21, no. 6, pp. 1001–1008, 2014, doi: 10.1136/amiajnl-2013-002627.

[21] M. Allen and D. Cervo, *Multi-Domain Master Data Management: Advanced MDM*

and Data Governance in Practice, 1st ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2015.

[22] K. Holzer and W. Gall, "Utilizing IHE-based Electronic Health Record Systems for Secondary Use," *Methods Inf. Med.*, vol. 50, no. 4, pp. 319–325, 2011, doi: 10.3414/ME10-01-0060.

[23] G. H. Knibbs, "The International Classification of Disease and Causes of Death and its Revision.," *Med. J.*, vol. 1, pp. 2–12, 1929.

[24] W. V. Sujansky, J. M. Overhage, S. Chang, J. Frohlich, and S. A. Faus, "The Development of a Highly Constrained Health Level 7 Implementation Guide to Facilitate Electronic Laboratory Reporting to Ambulatory Electronic Health Record Systems," *J. Am. Med. Informatics Assoc.*, vol. 16, no. 3, pp. 285–290, 2009, doi: 10.1197/jamia.M2610.

[25] B. P. Bentley TE, Price C, "Structural and lexical features of successive versions of the Read Codes.," in *Proc Annual Conference of the Primary Health Care Specialist Group of the British Computer Society*, Teasdale S, Ed. UK, 1996, pp. 91–103.

[26] A. Rector, A. Gangemi, E. Galeazzi, A. J. Glowinski, and A. Rossi-Mori, "The {GALEN} {CORE} Model Schemata for Anatomy: Towards a re-usable application-independent model of medical concepts," *Proc.\12th Int.\Congress Eur. Fed. Med. Informatics*, pp. 229–233, 1994.

[27] R. A. Côte and S. Robboy, "Progress in medical information management: Systematized nomenclature of medicine (snomed)," *JAMA*, vol. 243, no. 8, pp. 756–762, Feb. 1980, [Online]. Available: http://dx.doi.org/10.1001/jama.1980.03300340032015.

[28] K. K. Kakazu, L. W. K. Cheung, and W. Lynne, "The Cancer Biomedical Informatics Grid (caBIG): pioneering an expansive network of information and tools for collaborative cancer research.," *Hawaii Med. J.*, vol. 63, no. 9, pp. 273–275, Sep. 2004.

[29] "ICNARC – Intensive Care National Audit & Research Centre." https://www.icnarc.org/ (accessed Mar. 28, 2022).

[30] D. A. B. Lindberg, B. L. Humphreys, and A. T. McCray, "The unified medical language system," *Methods Inf. Med.*, vol. 32, no. 4, pp. 281–291, Feb. 1993, doi: 10.1055/S-0038-1634945/ID/JR1634945-21.

[31] Corepoint Health, “The HL7 Evolution. Comparing HL7 Version 2 to Version 3, Including a History of Version 2,” 2009. [Online]. Available: http://www.corepointhealth.com/sites/default/files/whitepapers/hl7-history-v2-v3.pdf.

[32] R. Cornet and N. De Keizer, “Forty years of SNOMED: a literature review,” *BMC Med. Inform. Decis. Mak.*, vol. 8 Suppl 1, no. Suppl 1, 2008, doi: 10.1186/1472-6947-8-S1-S2.

[33] W. Maidhof and O. Hilas, “Lupus: an overview of the disease and management options.,” *P T*, vol. 37, no. 4, pp. 240–9, Apr. 2012, Accessed: Jul. 06, 2018. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22593636.

[34] “MASTERPLANS Maximising SLE Therapeutic Potential by Application of Novel and Systematic Approaches.” http://www.lupusmasterplans.org/home.html (accessed Mar. 19, 2018).

[35] M. A. Dooley *et al.*, “Mycophenolate versus Azathioprine as Maintenance Therapy for Lupus Nephritis,” *N. Engl. J. Med.*, vol. 365, no. 20, pp. 1886–1895, Nov. 2011, doi: 10.1056/NEJMoa1014460.

[36] G. B. Appel *et al.*, “Mycophenolate Mofetil versus Cyclophosphamide for Induction Treatment of Lupus Nephritis,” *J. Am. Soc. Nephrol.*, vol. 20, no. 5, pp. 1103–1112, May 2009, doi: 10.1681/ASN.2008101028.

[37] B. H. Rovin *et al.*, “Efficacy and safety of rituximab in patients with active proliferative lupus nephritis: The lupus nephritis assessment with rituximab study,” *Arthritis Rheum.*, vol. 64, no. 4, pp. 1215–1226, Apr. 2012, doi: 10.1002/art.34359.

[38] J. T. Merrill *et al.*, “Efficacy and safety of rituximab in moderately-to-severely active systemic lupus erythematosus: The randomized, double-blind, phase ii/iii systemic lupus erythematosus evaluation of rituximab trial,” *Arthritis Rheum.*, vol. 62, no. 1, pp. 222–233, Jan. 2010, doi: 10.1002/art.27233.

[39] “PubMed - NCBI,” 2017. https://www.ncbi.nlm.nih.gov/pubmed/ (accessed Dec. 07, 2017).

[40] B. Louie, P. Mork, F. Martin-Sanchez, A. Halevy, and P. Tarczy-Hornoch, “Data integration and genomic medicine,” *J. Biomed. Inform.*, vol. 40, no. 1, pp. 5–16, 2007, doi: 10.1016/j.jbi.2006.02.007.

[41] F. Freitas, S. Schulz, and E. Moraes, “Survey of current terminologies and

ontologies in biology and medicine," *R. Eletr. Com. Inf. Inov. Saúde. Rio Janeiro*, vol. 3, no. 1, pp. 8–20, 2009, doi: 10.3395/reciis.v3i1.239en.

[42] I. D. Dinov, "Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data," *Gigascience*, vol. 5, no. 1, p. 12, 2016, doi: 10.1186/s13742-016-0117-6.

[43] G. Nieva, "Integrating Heterogeneous Data." p. 104, 2016.

[44] M. Ivanović and Z. Budimac, "An overview of ontologies and data resources in medical domains," *Expert Syst. Appl.*, vol. 41, no. 11, pp. 5158–5166, 2014, doi: 10.1016/j.eswa.2014.02.045.

[45] A. W. Toga and I. D. Dinov, "Sharing big biomedical data," *J. Big Data*, vol. 2, no. 1, p. 7, 2015, doi: 10.1186/s40537-015-0016-1.

[46] M. Leventer-Roberts and R. Balicer, "Data Integration in Health Care," in *Handbook Integrated Care*, V. Amelung, V. Stein, N. Goodwin, R. Balicer, E. Nolte, and E. Suter, Eds. Cham: Springer International Publishing, 2017, pp. 121–129.

[47] J. Webster and R. T. Watson, "Analyzing the past to prepare for the future : Writing a literature review Reproduced with permission of the copyright owner . Further reproduction prohibited without permission .," *MIS Q.*, vol. 26, no. 2, pp. xiii–xxiii, 2002, doi: 10.2307/4132319.

[48] P. Ziegler and K. R. Dittrich, "Data Integration --- Problems, Approaches, and Perspectives," in *Conceptual Modelling in Information Systems Engineering*, J. Krogstie, A. L. Opdahl, and S. Brinkkemper, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 39–58.

[49] S. B. Dusetzina, S. Tyree, A.-M. Meyer, A. Meyer, L. Green, and W. R. Carpenter, "An overview of record linkage methods," *Link. Data Heal. Serv. Res. A Framew. Instr. Guid.*, pp. 29–49, 2014, doi: AHRQ No.14-EHC033.

[50] B. Smith *et al.*, "The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration," *Nat. Biotechnol.*, vol. 25, no. 11, pp. 1251–1255, 2007, doi: 10.1038/nbt1346.

[51] J. D. Tenenbaum, S.-A. Sansone, and M. Haendel, "A sea of standards for omics data: sink or swim?," *J. Am. Med. Informatics Assoc.*, vol. 21, no. 2, pp. 200–203, 2014, doi: 10.1136/amiajnl-2013-002066.

[52] N. Guarino, D. Oberle, and S. Staab, “Handbook on Ontologies,” pp. 1–17, 2009, doi: 10.1007/978-3-540-92673-3.

[53] S. Schulz, H. Stenzhorn, M. Boeker, and B. Smith, “Strengths and limitations of formal ontologies in the biomedical domain,” vol. 3, no. 1, pp. 31–45, 2009, doi: 10.3395/reciis.v3i1.241en.Strengths.

[54] B. Smith, W. Kusnierczyk, D. Schober, and W. Ceusters, “Towards a reference terminology for ontology research and development in the biomedical domain,” *CEUR Workshop Proc.*, vol. 222, pp. 57–65, 2006.

[55] “Gene Ontology Consortium,” 2017. http://www.geneontology.org/ (accessed Dec. 07, 2017).

[56] D. Robinson *et al.*, “Updating the Read Codes: user-interactive maintenance of a dynamic clinical vocabulary,” *J. Am. Med. Inform. Assoc.*, vol. 4, no. 6, pp. 465–472, 1997, doi: 10.1136/jamia.1997.0040465.

[57] K. Kuhn and et al., “The Cancer Biomedical Informatics Grid (caBIG): infrastructure and applications for a worldwide research community.,” *Stud. Health Technol. Inform.*, vol. 129, no. 1, pp. 330–4, 2007, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17911733.

[58] F. R. Elevitch, “SNOMED CT: electronic health record enhances anesthesia patient safety,” *AANA J.*, vol. 73, no. 5, pp. 361-366 6p, 2005, [Online]. Available: http://search.ebscohost.com/login.aspx?direct=true&db=rzh&AN=106546447&site=ehost-live.

[59] J. Millar, “The need for a global language-SNOMED CT introduction,” *Stud. Health Technol. Inform.*, vol. 225, pp. 683–685, 2016, doi: 10.3233/978-1-61499-658-3-683.

[60] M. F. Chiang, J. C. Hwang, A. C. Yu, D. S. Casper, J. J. Cimino, and J. B. Starren, “Reliability of SNOMED-CT coding by three physicians using two terminology browsers.,” *AMIA Annu. Symp. Proc.*, pp. 131–5, 2006, [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1839418&tool=pmcentrez&rendertype=abstract.

[61] S. De Lusignan, T. Chan, and S. Jones, “Large complex terminologies: More coding choice, but harder to find data - Reflections on introduction of SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms) as an NHS standard,” *Inform. Prim. Care*, vol. 19, no. 1, pp. 3–5, 2011, doi: 10.14236/jhi.v19i1.787.

[62] A. L. Rector, S. Brandt, and T. Schneider, “Getting the foot out of the pelvis: modeling problems affecting use of SNOMED CT hierarchies in practical applications,” *J. Am. Med. Informatics Assoc.*, vol. 18, no. 4, pp. 432–440, 2011, doi: 10.1136/amiajnl-2010-000045.

[63] “ICD-10-CM,” 2017. https://www.cdc.gov/nchs/icd/icd10cm.htm (accessed Dec. 12, 2017).

[64] “WHO | International Classification of Diseases,” *WHO*, 2017, Accessed: Dec. 07, 2017. [Online]. Available: http://www.who.int/classifications/icd/en/.

[65] I. Kurbasic *et al.*, “The Advantages and Limitations of International Classification of Diseases, Injuries and Causes of Death from Aspect of Existing Health Care System of B and H,” *Acta Inform. Medica*, vol. 16, no. 3, p. 159, 2008, doi: 10.5455/aim.2008.16.159-161.

[66] “OpenGALEN Mission Statement,” 2017. http://www.opengalen.org/ (accessed Dec. 07, 2017).

[67] “Foundational Model of Anatomy ontology,” 2017. http://sig.biostr.washington.edu/projects/fm/AboutFM.html (accessed Dec. 07, 2017).

[68] “Read Codes - NHS Digital,” *ReadCodes NHS Digital*, 2017. https://digital.nhs.uk/article/1104/Read-Codes (accessed Dec. 07, 2017).

[69] “SNOMED CT Worldwide.” https://www.snomed.org/snomed-ct/snomed-ct-worldwide (accessed Dec. 30, 2017).

[70] “Unified Medical Language System (UMLS),” Accessed: Dec. 07, 2017. [Online]. Available: https://www.nlm.nih.gov/research/umls/.

[71] “Fact Sheet - Unified Medical Language System®.” https://www.nlm.nih.gov/pubs/factsheets/umls.html (accessed Dec. 30, 2017).

[72] O. Bodenreider, “The Unified Medical Language System (UMLS): integrating biomedical terminology,” *Nucleic Acids Res.*, vol. 32, no. 90001, pp. 267D – 270, 2004, doi: 10.1093/nar/gkh061.

[73] P. R. Williamson *et al.*, “The COMET Handbook: version 1.0,” *Trials*, vol. 18, no. S3, p. 280, Jun. 2017, doi: 10.1186/s13063-017-1978-4.

[74] C. Huttenhower and O. G. Troyanskaya, “Bayesian Data Integration : a Functional Perspective,” *Gene*, pp. 341–351, 2006, doi: 10.1142/1860947573_0044.

[75] O. Gevaert, *a Bayesian Network Integration Framework for modeling biomedical data*, no. December. 2008.

[76] R. S. Savage, Z. Ghahramani, J. E. Griffin, B. J. de la Cruz, and D. L. Wild, "Discovering transcriptional modules by Bayesian data integration," *Bioinformatics*, vol. 26, no. 12, pp. 158–167, 2010, doi: 10.1093/bioinformatics/btq210.

[77] H. Kang, "The prevention and handling of the missing data.," *Korean J. Anesthesiol.*, vol. 64, no. 5, pp. 402–6, May 2013, doi: 10.4097/kjae.2013.64.5.402.

[78] D. B. Rubin, "Inference and Missing Data," *Biometrika*, vol. 63, no. 3, p. 581, Dec. 1976, doi: 10.2307/2335739.

[79] P. Allison, "Missing Data." Thousand Oaks, California, 2002, doi: 10.4135/9781412985079.

[80] A. Briggs, T. Clark, J. Wolstenholme, and P. Clarke, "Missing... presumed at random: cost-analysis of incomplete data.," *Health Econ.*, vol. 12, no. 5, pp. 377–392, May 2003, doi: 10.1002/hec.766.

[81] B. J. Wells, A. S. Nowacki, K. Chagin, and M. W. Kattan, "Strategies for Handling Missing Data in Electronic Health Record Derived Data," vol. 1, no. 3, 2013, doi: 10.13063/2327-9214.1035.

[82] R. H. H. Groenwold, I. R. White, A. R. T. Donders, J. R. Carpenter, D. G. Altman, and K. G. M. Moons, "Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis.," *CMAJ*, vol. 184, no. 11, pp. 1265–9, Aug. 2012, doi: 10.1503/cmaj.110977.

[83] Z. Zhang, "Missing data imputation: focusing on single imputation.," *Ann. Transl. Med.*, vol. 4, no. 1, p. 9, Jan. 2016, doi: 10.3978/j.issn.2305-5839.2015.12.38.

[84] G. J. M. G. van der Heijden, A. R. T. Donders, T. Stijnen, and K. G. M. Moons, "Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: A clinical example," *J. Clin. Epidemiol.*, vol. 59, no. 10, pp. 1102–1109, Oct. 2006, doi: 10.1016/j.jclinepi.2006.01.015.

[85] J. A. C. Sterne *et al.*, "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls," *BMJ*, vol. 338, no. 7713, pp. 157–160, Jun. 2009, doi: 10.1136/BMJ.B2393.

[86] J. C. Jakobsen, C. Gluud, J. Wetterslev, and P. Winkel, "When and how should

multiple imputation be used for handling missing data in randomised clinical trials - A practical guide with flowcharts," *BMC Med. Res. Methodol.*, vol. 17, no. 1, pp. 1–10, 2017, doi: 10.1186/s12874-017-0442-1.

[87] P. Hayati Rezvan, K. J. Lee, and J. A. Simpson, "The rise of multiple imputation: a review of the reporting and implementation of the method in medical research," *BMC Med. Res. Methodol.*, vol. 15, no. 1, p. 30, 2015, doi: 10.1186/s12874-015-0022-1.

[88] S. Sinharay, H. S. Stern, and D. Russell, "The use of multiple imputation for the analysis of missing data.," *Psychol. Methods*, vol. 6, no. 4, pp. 317–329, 2001, doi: 10.1037/1082-989X.6.4.317.

[89] S. van Buuren and K. Groothuis-Oudshoorn, "mice: 'Multivariate Imputation by Chained Equations in R,'" *J. Stat. Softw.*, vol. 45, no. 3, pp. 1–67, Dec. 2011, doi: 10.18637/jss.v045.i03.

[90] J. L. Schafer and M. K. Olsen, "Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective," *Multivariate Behav. Res.*, vol. 33, no. 4, pp. 545–571, Oct. 1998, doi: 10.1207/s15327906mbr3304_5.

[91] S. van Buuren, "Multiple imputation of discrete and continuous data by fully conditional specification," *Stat. Methods Med. Res.*, vol. 16, no. 3, pp. 219–242, Jun. 2007, doi: 10.1177/0962280206074463.

[92] "Markov Chain Monte Carlo and Gibbs Sampling," Accessed: Jul. 17, 2018. [Online]. Available: http://nitro.biosci.arizona.edu/courses/EEB519A-2007/pdfs/Gibbs.pdf.

[93] T. E. Raghunathan, J. M. Lepkowski, J. Van Hoewyk, P. Solenberger, and J. van Hoewyk, "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models Key Words: Item nonresponse; Missing at random; Multiple imputation; Nonignorable missing mechanism; Regression; Sampling properties and simulations," *Surv. Methodol.*, vol. 27, no. 1, 2001.

[94] Rubin Donald B, *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc., 1987.

[95] S. van Buuren and C. G. M. Oudshoorn, *Multivariate imputation by chained equations : MICE V1.0 users's manual*. Leiden: TNO Prevention and Health Public Health, 2000.

[96] A. M. G. Ali *et al.*, “Comparison of methods for handling missing data on immunohistochemical markers in survival analysis of breast cancer.,” *Br. J. Cancer*, vol. 104, no. 4, pp. 693–9, Feb. 2011, doi: 10.1038/sj.bjc.6606078.

[97] M. Quartagno and J. R. Carpenter, “Multiple imputation for IPD meta-analysis: allowing for heterogeneity and studies with missing covariates,” *Stat. Med.*, vol. 35, no. 17, pp. 2938–2954, Jul. 2016, doi: 10.1002/SIM.6837.

[98] S. van Buuren and K. Groothuis-Oudshoorn, “mice: Multivariate Imputation by Chained Equations in R,” *J. Stat. Softw.*, vol. 45, no. 3 SE-Articles, pp. 1–67, Dec. 2011, doi: 10.18637/jss.v045.i03.

[99] J. L. Schafer and R. M. Yucel, “Computational Strategies for Multivariate Linear Mixed-Effects Models With Missing Values,” *J. Comput. Graph. Stat.*, vol. 11, pp. 437–457, 2002.

[100] R. M. Yucel, “Random-covariances and mixed-effects models for imputing multivariate multilevel continuous data,” *Stat. Modelling*, vol. 11, no. 4, p. 351, Aug. 2011, doi: 10.1177/1471082X1001100404.

[101] J. L. S. Jing Hua Zhao, “pan: Multiple imputation for multivariate panel or clustered data.” R package, 2015.

[102] M. Quartagno and J. R. Carpenter, “Multiple imputation for discrete data: Evaluation of the joint latent normal model,” *Biometrical J.*, vol. 61, no. 4, pp. 1003–1019, Jul. 2019, doi: 10.1002/BIMJ.201800222.

[103] J. R. Carpenter, H. Goldstein, and M. G. Kenward, “REALCOM-IMPUTE Software for Multilevel Multiple Imputation with Mixed Response Types,” *J. Stat. Softw.*, vol. 45, no. 5, pp. 1–14, Dec. 2011, doi: 10.18637/JSS.V045.I05.

[104] Matteo Quartagno and James Carpenter, “jomo: Multilevel Joint Modelling Multiple Imputation,” *Multiple Imputation and its Application*. John Wiley and Sons Ltd, Feb. 08, 2022, doi: 10.1002/9781119942283.

[105] S. van Buuren, “Multiple Imputation of Multilevel Data,” in *Handbook of Advanced Multilevel Analysis*, Hox J and Roberts J, Eds. New York: Taylor & Francis, 2015, pp. 173–196.

[106] S. van Buuren, *Flexible Imputation of Missing Data*. Chapman and Hall/CRC, 2012.

[107] M. Resche-Rigon and I. R. White, “Multiple imputation by chained equations for systematically and sporadically missing multilevel data,” *Stat. Methods Med. Res.*,

vol. 27, no. 6, pp. 1634–1649, Jun. 2018, doi: 10.1177/0962280216666564.

[108] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, “Multiple imputation by chained equations: what is it and how does it work?,” *Int. J. Methods Psychiatr. Res.*, vol. 20, no. 1, pp. 40–49, Mar. 2011, doi: 10.1002/MPR.329.

[109] T. C. El-Galaly, C. Y. Cheah, and D. Villa, “Real world data as a key element in precision medicine for lymphoid malignancies: potentials and pitfalls,” *Br. J. Haematol.*, vol. 186, no. 3, pp. 409–419, Aug. 2019, doi: 10.1111/BJH.15965.

[110] D. B. Rubin, “Inference and missing data,” *Biometrika*, vol. 63, no. 3, pp. 581–592, Dec. 1976, doi: 10.1093/biomet/63.3.581.

[111] “Package ‘mice’ Title Multivariate Imputation by Chained Equations,” 2018, doi: 10.18637/jss.v045.i03>.

[112] T. P. Morris, I. R. White, and M. J. Crowther, “Using simulation studies to evaluate statistical methods,” *Stat. Med.*, vol. 38, no. 11, pp. 2074–2102, May 2019, doi: 10.1002/sim.8086.

[113] R. Bellazzi, “Big data and biomedical informatics: a challenging opportunity.,” *Yearb. Med. Inform.*, vol. 9, no. 1, pp. 8–13, May 2014, doi: 10.15265/IY-2014-0024.

[114] V. Audigier *et al.*, “Multiple Imputation for Multilevel Data with Continuous and Binary Variables,” vol. 33, no. 2, pp. 160–183, 2018, doi: 10.2307/26770989.

[115] R. R. Andridge, “Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials,” *Biometrical J.*, vol. 53, no. 1, pp. 57–74, Feb. 2011, doi: 10.1002/BIMJ.201000140.

[116] C. K. Enders, S. A. Mistler, and B. T. Keller, “Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation.,” *Psychol. Methods*, vol. 21, no. 2, pp. 222–240, Jun. 2016, doi: 10.1037/met0000063.

[117] S. A. Mistler, “A SAS ® Macro for Applying Multiple Imputation to Multilevel Data,” Accessed: Mar. 28, 2022. [Online]. Available: http://www.mistlerconsulting.com/software.html.

[118] M. Quartagno, J. R. Carpenter, and H. Goldstein, “Multiple Imputation with Survey Weights: A Multilevel Approach,” *J. Surv. Stat. Methodol.*, vol. 8, no. 5, pp. 965–989, Nov. 2020, doi: 10.1093/jssam/smz036.

[119] J. E. V Lloyd, J. Obradović, R. M. Carpiano, F. Motti-Stefanidi, and J. E. V,

"JMASM 32: Multiple Imputation of Missing Multilevel, Longitudinal Data: A Case When Practical Considerations Trump Best Practices?," *J. Mod. Appl. Stat. Methods*, vol. 12, no. 1, p. 29, 2013, doi: 10.22237/jmasm/1367382480.

[120] A. Hammon and S. Zinn, "Multiple imputation of binary multilevel missing not at random data," *J. R. Stat. Soc. Ser. C (Applied Stat.*, vol. 69, no. 3, pp. 547–564, Jun. 2020, doi: 10.1111/RSSC.12401.

[121] S. Jolani, T. P. A. Debray, H. Koffijberg, S. Van Buuren, and K. G. M. Moons, "Imputation of systematically missing predictors in an individual participant data meta-analysis : a generalized approach using MICE," no. January, 2015, doi: 10.1002/sim.6451.

[122] M. Resche-Rigon, I. R. White, J. W. Bartlett, S. A. E. Peters, and S. G. Thompson, "Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data," *Stat. Med.*, vol. 32, no. 28, pp. 4890–4905, Dec. 2013, doi: 10.1002/SIM.5894.

[123] J. L. Schafer, "Imputation of missing covariates under a multivariate linear mixed model," 1997. Accessed: Mar. 28, 2022. [Online]. Available: http://stat.psu.edu/research-old/technical-reports/archived-technical-reports.

[124] "CDISC | Clear Data. Clear Impact." https://www.cdisc.org/ (accessed Mar. 28, 2022).

[125] I. Fortier *et al.*, "Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies," *Int. J. Epidemiol.*, vol. 39, no. 5, pp. 1383–1393, Oct. 2010, doi: 10.1093/ije/dyq139.

[126] P. R. Burton, I. Fortier, and B. M. Knoppers, "The global emergence of epidemiological biobanks: Opportunities and challenges," *Hum. Genome Epidemiol. Build. Evid. Using Genet. Inf. to Improv. Heal. Prev. Dis. Second Ed.*, May 2010, doi: 10.1093/ACPROF:OSO/9780195398441.003.0005.

[127] M. J. Murtagh, I. Demir, J. R. Harris, and P. R. Burton, "Realizing the promise of population biobanks: A new model for translation," *Hum. Genet.*, vol. 130, no. 3, pp. 333–345, Sep. 2011, doi: 10.1007/s00439-011-1036-3.

[128] I. Fortier *et al.*, "Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies," *Int. J. Epidemiol.*, vol. 39, no. 5, pp. 1383–1393, Oct. 2010, doi: 10.1093/IJE/DYQ139.

[129] H. Le Sueur, I. N. Bruce, N. Geifman, and N. Geifman, “The challenges in data integration - Heterogeneity and complexity in clinical trials and patient registries of Systemic Lupus Erythematosus,” *BMC Med. Res. Methodol.*, vol. 20, no. 1, pp. 1–5, Jun. 2020, doi: 10.1186/S12874-020-01057-0/FIGURES/1.

[130] V. N. Hicken, S. N. Thornton, and R. A. Rocha, “Integration challenges of clinical information systems developed without a shared data dictionary.,” *Stud. Health Technol. Inform.*, vol. 107, no. Pt 2, pp. 1053–1057, 2004.

[131] N. Villarroel, E. Davidson, P. Pereyra-Zamora, A. Krasnik, and R. S. Bhopal, “Heterogeneity/granularity in ethnicity classifications project: the need for refining assessment of health status.,” Edinburgh, 2020. Accessed: Mar. 29, 2022. [Online]. Available: www.rwjf.org/.

[132] K. Wang, H. G. Nardini, L. Post, T. Edwards, M. Nunez-Smith, and C. Brandt, “Information Loss in Harmonizing Granular Race and Ethnicity Data: Descriptive Study of Standards,” *J Med Internet Res 2020;22(7)e14591 https//www.jmir.org/2020/7/e14591*, vol. 22, no. 7, p. e14591, Jul. 2020, doi: 10.2196/14591.

[133] A. B. Pedersen *et al.*, “Missing data and multiple imputation in clinical epidemiological research.,” *Clin. Epidemiol.*, vol. 9, pp. 157–166, 2017, doi: 10.2147/CLEP.S129785.

[134] “mice: Algorithmic convergence and inference pooling.” https://www.gerkovink.com/miceVignettes/Convergence_pooling/Convergence_and_pooling.html (accessed Mar. 30, 2022).

[135] S. Grund, O. Lüdtke, and A. Robitzsch, “Multiple imputation of missing data in multilevel models with the R package mdmb: a flexible sequential modeling approach,” *Behav. Res. Methods*, vol. 53, no. 6, pp. 2631–2649, 2021, doi: 10.3758/s13428-020-01530-0.

[136] P. Ranganathan, C. Pramesh, and R. Aggarwal, “Common pitfalls in statistical analysis: Logistic regression,” *Perspect. Clin. Res.*, vol. 8, no. 3, p. 148, Jul. 2017, doi: 10.4103/PICR.PICR_87_17.

[137] N. Geifman and E. Rubin, “Towards an Age-Phenome Knowledge-base,” *BMC Bioinformatics*, vol. 12, no. 1, pp. 1–9, Jun. 2011, doi: 10.1186/1471-2105-12-229/FIGURES/4.

[138] S. Van Buuren, “Derived variables.” https://stefvanbuuren.name/fimd/sec-

knowledge.html.

[139] J. Kalter, M. G. Sweegers, I. M. Verdonck-De Leeuw, J. Brug, and L. M. Buffart, "Development and use of a flexible data harmonization platform to facilitate the harmonization of individual patient data for meta-analyses," *BMC Res. Notes*, vol. 12, no. 1, p. 164, Mar. 2019, doi: 10.1186/s13104-019-4210-7.

[140] A. Sampri *et al.*, "Probabilistic Approaches to Overcome Content Heterogeneity in Data Integration: A Study Case in Systematic Lupus Erythematosus.," *Stud. Health Technol. Inform.*, vol. 270, pp. 387–391, Jun. 2020, doi: 10.3233/SHTI200188.

[141] N. Peek and A. Abu-Hanna, "Clinical prognostic methods: trends and developments," *Journal of biomedical informatics*, vol. 48. pp. 1–4, Apr. 01, 2014, doi: 10.1016/j.jbi.2014.02.016.

[142] J. Hardt, M. Herke, and R. Leonhart, "Auxiliary variables in multiple imputation in regression with missing X: A warning against including too many in small sample research," *BMC Med. Res. Methodol.*, vol. 12, no. 1, pp. 1–13, Dec. 2012, doi: 10.1186/1471-2288-12-184/COMMENTS.

[143] L. M. Collins, J. L. Schafer, and C. M. Kam, "A comparison of inclusive and restrictive strategies in modern missing data procedures.," *Psychol. Methods*, vol. 6, no. 4, pp. 330–351, Dec. 2001.

[144] X. H. Zhou, G. J. Eckert, and W. M. Tierney, "Multiple imputation in public health research.," *Stat. Med.*, vol. 20, no. 9–10, pp. 1541–1549, May 2001, doi: 10.1002/sim.689.

[145] "mice: Passive imputation and Post-processing." https://www.gerkovink.com/miceVignettes/Passive_Post_processing/Passive_imputation_post_processing.html (accessed Mar. 29, 2022).

[146] G. Heinze and M. Schemper, "A solution to the problem of separation in logistic regression," *Stat. Med.*, vol. 21, no. 16, pp. 2409–2419, Aug. 2002, doi: 10.1002/SIM.1047.

[147] I. R. White, R. Daniel, and P. Royston, "Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables," *Comput. Stat. Data Anal.*, vol. 54, no. 10, pp. 2267–2275, Oct. 2010, doi: 10.1016/J.CSDA.2010.04.005.

[148] F. G. Elfadaly *et al.*, "BIMAM—a tool for imputing variables missing across datasets using a Bayesian imputation and analysis model," *Int. J. Epidemiol.*, vol.

50, no. 5, pp. 1419–1425, Nov. 2021, doi: 10.1093/IJE/DYAB177.

[149] J. H. Lee and J. C. Huber Jr, “Evaluation of Multiple Imputation with Large Proportions of Missing Data: How Much Is Too Much?,” *Iran. J. Public Health*, vol. 50, no. 7, pp. 1372–1380, Jul. 2021, doi: 10.18502/ijph.v50i7.6626.

[150] E. Samoli *et al.*, “What is the impact of systematically missing exposure data on air pollution health effect estimates?,” *Air Qual. Atmos. Heal. 2014 74*, vol. 7, no. 4, pp. 415–420, Mar. 2014, doi: 10.1007/S11869-014-0250-2.

[151] M. H. Secrest, R. W. Platt, P. Reynier, C. R. Dormuth, A. Benedetti, and K. B. Filion, “Multiple imputation for systematically missing confounders within a distributed data drug safety network: A simulation study and real-world example,” *Pharmacoepidemiol. Drug Saf.*, vol. 29, no. S1, pp. 35–44, 2020, doi: 10.1002/pds.4876.

[152] K. A. Dahal *et al.*, “Harmonization of data from cohort studies– potential challenges and opportunities,” *Int. J. Popul. Data Sci.*, vol. 3, no. 4, Sep. 2018, doi: 10.23889/ijpds.v3i4.868.

[153] A. Burgun and O. Bodenreider, “Accessing and integrating data and knowledge for biomedical research.,” *Yearb. Med. Inform.*, pp. 91–101, 2008.

[154] M. Desai, J. Kubo, D. Esserman, and M. B. Terry, “The handling of missing data in molecular epidemiology studies,” *Cancer Epidemiol. Biomarkers Prev.*, vol. 20, no. 8, pp. 1571–1579, Aug. 2011, doi: 10.1158/1055-9965.EPI-10-1311/70869/P/THE-HANDLING-OF-MISSING-DATA-IN-MOLECULAR.

[155] G. Peng *et al.*, “Call to action for global access to and harmonization of quality information of individual earth science datasets,” *Data Sci. J.*, vol. 20, no. 1, May 2021, doi: 10.5334/DSJ-2021-019/METRICS/.

[156] A. Zeb, J.-P. Soininen, and N. Sozer, “Data harmonisation as a key to enable digitalisation of the food sector: A review,” *Food Bioprod. Process.*, vol. 127, pp. 360–370, May 2021, doi: 10.1016/J.FBP.2021.02.005.

[157] J. G. Ibrahim and G. Molenberghs, “Missing data methods in longitudinal studies: a review.,” *Test (Madr).*, vol. 18, no. 1, pp. 1–43, May 2009, doi: 10.1007/s11749-009-0138-x.

[158] R. J. Little *et al.*, “The Prevention and Treatment of Missing Data in Clinical Trials,” *N. Engl. J. Med.*, vol. 367, no. 14, pp. 1355–1360, Oct. 2012, doi:

10.1056/NEJMSR1203730/SUPPL_FILE/NEJMSR1203730_DISCLOSURES.PDF
.

[159] P. C. Austin, I. R. White, D. S. Lee, and S. van Buuren, “Missing Data in Clinical Research: A Tutorial on Multiple Imputation,” *Can. J. Cardiol.*, vol. 37, no. 9, pp. 1322–1331, Sep. 2021, doi: 10.1016/J.CJCA.2020.11.010/ATTACHMENT/61FA3316-533E-455E-947E-5A43D6B28DEC/MMC1.DOCX.

[160] U. Held *et al.*, “Methods for Handling Missing Variables in Risk Prediction Models,” *Am. J. Epidemiol.*, vol. 184, no. 7, pp. 545–551, Oct. 2016, doi: 10.1093/AJE/KWV346.

[161] V. Dinu and P. Nadkarni, “Guidelines for the effective use of entity-attribute-value modeling for biomedical databases,” *Int. J. Med. Inform.*, vol. 76, no. 11–12, pp. 769–779, Nov. 2007, doi: 10.1016/J.IJMEDINF.2006.09.023.

[162] C. A. Brandt *et al.*, “Metadata-driven creation of data marts from an EAV-modeled clinical research database,” *Int. J. Med. Inform.*, vol. 65, no. 3, pp. 225–241, Nov. 2002, doi: 10.1016/S1386-5056(02)00047-3.

[163] Y. Deng, C. Chang, M. S. Ido, and Q. Long, “Multiple Imputation for General Missing Data Patterns in the Presence of High-dimensional Data,” *Sci. Reports 2016 61*, vol. 6, no. 1, pp. 1–10, Feb. 2016, doi: 10.1038/srep21689.

[164] S. Burgess, I. R. White, M. Resche-Rigon, and A. M. Wood, “Combining multiple imputation and meta-analysis with individual participant data,” *Stat. Med.*, vol. 32, no. 26, pp. 4499–4514, 2013, doi: 10.1002/sim.5844.

[165] M. H. Secrest, R. W. Platt, P. Reynier, C. R. Dormuth, A. Benedetti, and K. B. Filion, “Multiple imputation for systematically missing confounders within a distributed data drug safety network: A simulation study and real-world example,” *Pharmacoepidemiol. Drug Saf.*, vol. 29, no. S1, pp. 35–44, Jan. 2020, doi: 10.1002/PDS.4876.

[166] P. L. Sankar and L. S. Parker, “The Precision Medicine Initiative’s All of Us Research Program: an agenda for research on its ethical, legal, and social issues,” *Genet. Med.*, vol. 19, no. 7, pp. 743–750, Jul. 2017, doi: 10.1038/GIM.2016.183.

[167] S. A. Sansone *et al.*, “Toward interoperable bioscience data,” *Nat. Genet.*, vol. 44, no. 2, pp. 121–126, Feb. 2012, doi: 10.1038/NG.1054.

[168] I. Budin-Ljøsne *et al.*, “Data sharing in large research consortia: experiences and recommendations from ENGAGE,” *Eur. J. Hum. Genet.*, vol. 22, no. 3, p. 317, Mar. 2014, doi: 10.1038/EJHG.2013.131.

[169] J. Bousquet *et al.*, “Pooling Birth Cohorts in Allergy and Asthma: European Union-Funded Initiatives – A MeDALL, CHICOS, ENRIECO, and GA2LEN Joint Paper,” *Int. Arch. Allergy Immunol.*, vol. 161, no. 1, pp. 1–10, 2013, doi: 10.1159/000343018.

[170] R. Shishegar *et al.*, “Using imputation to provide harmonized longitudinal measures of cognition across AIBL and ADNI,” *Sci. Reports 2021 111*, vol. 11, no. 1, pp. 1–11, Dec. 2021, doi: 10.1038/s41598-021-02827-6.

[171] L. Serra-Majem *et al.*, “Comparative analysis of nutrition data from national, household, and individual levels: results from a WHO-CINDI collaborative project in Canada, Finland, Poland, and Spain,” *J. Epidemiol. Community Health*, vol. 57, no. 1, pp. 74–80, Jan. 2003, doi: 10.1136/JECH.57.1.74.

[172] A. D. Shah, J. W. Bartlett, J. Carpenter, O. Nicholas, and H. Hemingway, “Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study,” *Am. J. Epidemiol.*, vol. 179, no. 6, pp. 764–774, 2014, doi: 10.1093/aje/kwt312.

[173] L. Beretta and A. Santaniello, “Nearest neighbor imputation algorithms: A critical evaluation,” *BMC Med. Inform. Decis. Mak.*, vol. 16, no. 3, pp. 197–208, Jul. 2016, doi: 10.1186/S12911-016-0318-Z/TABLES/5.

[174] K. Pelckmans, J. De Brabanter, J. A. K. Suykens, and B. De Moor, “Handling missing values in support vector machine classifiers,” *Neural Netw.*, vol. 18, no. 5–6, pp. 684–692, Jul. 2005, doi: 10.1016/J.NEUNET.2005.06.025.

[175] J. de Leeuw and E. Meijer, “Introduction to multilevel analysis.,” *Handbook of multilevel analysis.* Springer Science + Business Media, de Leeuw, Jan: Department of Statistics, University of California at Los Angeles, Los Angeles, CA, US, 90095-1554, deleeuw@stat.ucla.edu, pp. 1–75, 2008, doi: 10.1007/978-0-387-73186-5_1.

[176] A. Custovic *et al.*, “The study team for early life asthma research (STELAR) consortium ‘Asthma e-lab’: Team science bringing data, methods and investigators together,” *Thorax*, vol. 70, no. 8, pp. 799–801, 2015, doi: 10.1136/thoraxjnl-2015-206781.

Appendices

Appendix A: Systematically missing values

Here we give simulation results for 10 scenarios presented in Chapter 3. Full Data refers to complete true data, Complete Records to traditional data integration – complete case analysis where variable with systematically missing values is omitted from analysis model, FCS refers to fully conditional specification - multiple imputation, e_i refers to model error applied in outcome's data generating mechanism (equation 3.1).

Simulation with studies of same sizes

Scenario 1 (D=2, N=200 per study)

Table A. 1. Simulation results when one chosen study is completely missing X_1. Data generated under scenario 1, e_i: N ~ (0, 0.2), using equation 3.1.

e_i:N~(0,0.2)	**β₀**				**β₁**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	2.508	0.032	0.031	0.951	0.688	0.010	0.010	0.965
Complete Records	2.504	0.112	0.104	0.966	0.000	0.000	0.000	0.000
FCS	2.508	0.047	0.047	0.927	0.694	0.014	0.014	0.923
	β₂				**β₃**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-2.059	0.005	0.005	0.953	-0.788	0.039	0.038	0.952
Complete Records	-1.990	0.018	0.005	0.000	-0.782	0.137	0.134	0.960
FCS	-2.059	0.007	0.007	0.943	-0.787	0.057	0.059	0.931
	β₄				**β₅**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.624	0.034	0.034	0.954	-0.949	0.056	0.056	0.956
Complete Records	0.628	0.120	0.118	0.954	-0.938	0.196	0.194	0.959
FCS	0.624	0.050	0.052	0.933	-0.951	0.083	0.085	0.932

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table A. 2. Simulation results when one chosen study is completely missing X_1. Data generated under scenario 1, e_i: N ~ (0, 2), using equation 3.1.

e_i:N~(0,2)	**β₀**				**β₁**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	2.510	0.320	0.314	0.951	0.689	0.103	0.099	0.965
Complete Records	2.506	0.337	0.325	0.966	0.000	0.000	0.000	0.000
FCS	2.518	0.339	0.335	0.962	0.680	0.138	0.140	0.926
	β₂				**β₃**			

	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-2.058	0.051	0.051	0.953	-0.788	0.391	0.378	0.952
Complete Records	-1.989	0.053	0.049	0.768	-0.782	0.413	0.399	0.962
FCS	-2.055	0.055	0.051	0.959	-0.795	0.415	0.404	0.962
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.622	0.343	0.341	0.954	-0.947	0.559	0.557	0.956
Complete Records	0.626	0.362	0.355	0.963	-0.937	0.589	0.570	0.957
FCS	0.614	0.364	0.368	0.954	-0.963	0.594	0.591	0.948

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table A. 3. Simulation results when one chosen study is completely missing X_1. Data generated under scenario 1, e_i: N ~ (0, 20), using equation 3.1.

e_i:N~(0,20)	β_0				β_1			
	Mean	**mSE**	**EmpSE**	**Cov**	**Mean**	**mSE**	**EmpSE**	Cov
Full Data	2.530	3.196	3.140	0.951	0.704	1.027	0.988	0.965
Complete Records	2.527	3.194	3.132	0.952	0.000	0.000	0.000	0.000
FCS	2.528	3.218	3.158	0.947	0.691	1.479	1.429	0.940
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-2.046	0.514	0.505	0.953	-0.791	3.914	3.777	0.952
Complete Records	-1.976	0.503	0.491	0.951	-0.786	3.912	3.778	0.956
FCS	-2.042	0.525	0.514	0.961	-0.800	3.940	3.804	0.950
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.606	3.432	3.412	0.954	-0.927	5.592	5.572	0.956
Complete Records	0.609	3.429	3.403	0.956	-0.927	5.588	5.556	0.953
FCS	0.609	3.457	3.442	0.954	-0.933	5.632	5.617	0.954

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Scenario 2 (D=2, N=1000 per study)

Table A. 4. Simulation results when one chosen study is completely missing X_1. Data generated under scenario 2, e_i: N ~ (0, 0.2), using equation 3.1.

e_i:N~(0,0.2)	β_0				β_1			
	Mean	**mSE**	**EmpSE**	**Cov**	**Mean**	**mSE**	EmpSE	Cov
Full Data	2.508	0.014	0.014	0.954	0.687	0.005	0.005	0.957
Complete Records	2.508	0.050	0.046	0.972	0.000	0.000	0.000	0.000
FCS	2.508	0.020	0.021	0.927	0.689	0.006	0.006	0.933

	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-2.059	0.002	0.002	0.950	-0.787	0.017	0.017	0.950
Complete Records	-1.990	0.008	0.002	0.000	-0.789	0.061	0.059	0.955
FCS	-2.059	0.003	0.003	0.936	-0.787	0.025	0.026	0.931
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.624	0.015	0.015	0.952	-0.949	0.025	0.024	0.948
Complete Records	0.624	0.053	0.051	0.967	-0.949	0.086	0.086	0.952
FCS	0.624	0.022	0.022	0.932	-0.949	0.035	0.036	0.924

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table A. 5. Simulation results when one chosen study is completely missing X_1. Data generated under scenario 2, e_i: N ~ (0, 2), using equation 3.1.

e_i:N~(0,2)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	2.510	0.141	0.142	0.954	0.686	0.046	0.046	0.957
Complete Records	2.510	0.149	0.148	0.953	0.000	0.000	0.000	0.000
FCS	2.511	0.150	0.148	0.955	0.685	0.061	0.063	0.929
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-2.060	0.023	0.023	0.950	-0.786	0.173	0.174	0.950
Complete Records	-1.991	0.024	0.022	0.168	-0.789	0.183	0.184	0.952
FCS	-2.060	0.024	0.023	0.960	-0.786	0.184	0.182	0.950
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.623	0.152	0.153	0.952	-0.951	0.246	0.240	0.948
Complete Records	0.623	0.160	0.159	0.956	-0.951	0.259	0.255	0.952
FCS	0.620	0.161	0.160	0.947	-0.952	0.261	0.255	0.949

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table A. 6. Simulation results when one chosen study is completely missing X_1. Data generated under scenario 2, e_i: N ~ (0, 20), using equation 3.1.

e_i:N~(0,20)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	2.518	1.416	1.395	0.960	0.672	0.457	0.461	0.948
Complete Records	2.518	1.416	1.394	0.960	0.000	0.000	0.000	0.000
FCS	2.519	1.419	1.392	0.961	0.648	0.659	0.678	0.933
	β_2				β_3			

	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-2.060	0.228	0.224	0.952	-0.794	1.734	1.714	0.942
Complete Records	-1.993	0.224	0.220	0.943	-0.795	1.735	1.714	0.941
FCS	-2.057	0.234	0.231	0.952	-0.796	1.738	1.710	0.945
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.620	1.521	1.507	0.958	-0.997	2.460	2.446	0.956
Complete Records	0.621	1.521	1.506	0.960	-0.996	2.461	2.446	0.956
FCS	0.620	1.524	1.505	0.961	-0.997	2.465	2.453	0.958

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Scenario 3 (D=5, N=200 per study)

Table A. 7. Simulation results when two chosen studies are completely missing X_1. Data generated under scenario 3, e_i: N ~ (0, 0.2), using equation 3.1.

e_i:N~(0,0.2)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	2.508	0.020	0.019	0.960	0.687	0.006	0.006	0.955
Complete Records	2.505	0.070	0.069	0.950	0.000	0.000	0.000	0.000
FCS	2.508	0.026	0.026	0.942	0.689	0.008	0.008	0.934
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-2.059	0.003	0.003	0.958	-0.789	0.025	0.024	0.946
Complete Records	-1.990	0.011	0.003	0.000	-0.785	0.086	0.090	0.940
FCS	-2.059	0.004	0.004	0.934	-0.789	0.032	0.032	0.949
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.623	0.022	0.021	0.960	-0.949	0.035	0.035	0.949
Complete Records	0.626	0.076	0.077	0.945	-0.949	0.123	0.126	0.940
FCS	0.623	0.028	0.028	0.946	-0.948	0.046	0.047	0.942

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table A. 8. Simulation results when two chosen studies are completely missing X_1. Data generated under scenario 3, e_i: N ~ (0, 2), using equation 3.1.

e_i:N~(0,2)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	2.511	0.201	0.190	0.960	0.687	0.065	0.065	0.955
Complete Records	2.508	0.212	0.203	0.960	0.000	0.000	0.000	0.000
FCS	2.510	0.209	0.198	0.954	0.689	0.081	0.083	0.940

	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-2.058	0.032	0.031	0.958	-0.802	0.246	0.238	0.946
Complete Records	-1.989	0.033	0.031	0.438	-0.798	0.259	0.257	0.948
FCS	-2.058	0.034	0.032	0.963	-0.801	0.256	0.248	0.954
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.617	0.215	0.206	0.960	-0.949	0.349	0.351	0.949
Complete Records	0.620	0.227	0.220	0.960	-0.949	0.368	0.374	0.939
FCS	0.619	0.224	0.216	0.959	-0.949	0.363	0.364	0.948

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table A. 9. Simulation results when two chosen studies are completely missing X_1. Data generated under scenario 3, e_i: N ~ (0, 20), using equation 3.1.

e_i:N~(0,20)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	2.535	2.006	1.897	0.960	0.680	0.647	0.649	0.955
Complete Records	2.532	2.006	1.905	0.957	0.000	0.000	0.000	0.000
FCS	2.535	2.010	1.899	0.960	0.688	0.850	0.862	0.941
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-2.048	0.324	0.314	0.958	-0.931	2.457	2.385	0.946
Complete Records	-1.980	0.317	0.309	0.953	-0.927	2.457	2.395	0.943
FCS	-2.048	0.329	0.314	0.965	-0.931	2.462	2.388	0.946
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.556	2.155	2.058	0.960	-0.949	3.491	3.509	0.949
Complete Records	0.561	2.155	2.065	0.958	-0.950	3.491	3.514	0.948
FCS	0.558	2.159	2.062	0.962	-0.951	3.498	3.509	0.952

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Scenario 4 (D=5, N=1000 per study)

Table A. 10. Simulation results when two chosen studies are completely missing X_1. Data generated under scenario 4, e_i: N ~ (0, 0.2), using equation 3.1.

e_i:N~(0,0.2)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	2.508	0.009	0.009	0.949	0.687	0.003	0.003	0.942
Complete Records	2.508	0.031	0.031	0.951	0.000	0.000	0.000	0.000

FCS	2.507	0.012	0.012	0.930	0.688	0.004	0.004	0.945
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-2.059	0.001	0.001	0.943	-0.787	0.011	0.011	0.943
Complete Records	-1.990	0.005	0.001	0.000	-0.788	0.038	0.040	0.947
FCS	-2.059	0.002	0.002	0.935	-0.787	0.014	0.015	0.926
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.624	0.010	0.010	0.944	-0.949	0.016	0.016	0.954
Complete Records	0.624	0.034	0.035	0.940	-0.952	0.055	0.056	0.937
FCS	0.624	0.012	0.014	0.919	-0.949	0.020	0.020	0.942

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table A. 11. Simulation results when two chosen studies are completely missing X_1. Data generated under scenario 4, e_i: N ~ (0, 2), using equation 3.1.

e_i:N~(0,2)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	2.506	0.090	0.093	0.949	0.686	0.029	0.029	0.942
Complete Records	2.506	0.094	0.097	0.936	0.000	0.000	0.000	0.000
FCS	2.506	0.093	0.096	0.947	0.686	0.036	0.037	0.937
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-2.059	0.014	0.014	0.943	-0.786	0.110	0.114	0.943
Complete Records	-1.991	0.015	0.014	0.002	-0.787	0.116	0.120	0.937
FCS	-2.059	0.015	0.015	0.952	-0.785	0.114	0.118	0.935
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.627	0.096	0.099	0.944	-0.947	0.155	0.159	0.954
Complete Records	0.627	0.101	0.105	0.938	-0.950	0.164	0.170	0.952
FCS	0.627	0.100	0.103	0.945	-0.947	0.161	0.166	0.947

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table A. 12. Simulation results when two chosen studies are completely missing X_1. Data generated under scenario 4, e_i: N ~ (0, 20), using equation 3.1.

e_i:N~(0,20)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	2.490	0.895	0.925	0.949	0.673	0.289	0.295	0.942
Complete Records	2.490	0.896	0.926	0.947	0.000	0.000	0.000	0.000
FCS	2.489	0.896	0.926	0.946	0.667	0.381	0.383	0.948

	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-2.063	0.144	0.144	0.943	-0.774	1.096	1.136	0.943
Complete Records	-1.996	0.142	0.142	0.934	-0.775	1.097	1.136	0.942
FCS	-2.062	0.147	0.147	0.953	-0.773	1.097	1.135	0.945
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.651	0.962	0.993	0.944	-0.926	1.552	1.589	0.954
Complete Records	0.651	0.962	0.994	0.944	-0.929	1.553	1.592	0.953
FCS	0.652	0.963	0.994	0.943	-0.926	1.553	1.591	0.953

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Simulation with studies of different sizes

Scenario 5 (D=5, N: different per study)

Table A. 13. Simulation results when two chosen studies are completely missing X_1. Data generated under scenario 5, e_i: N ~ (0, 0.2), using equation 3.1.

e_i:N~(0,0.2)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	2.508	0.027	0.027	0.943	0.688	0.009	0.008	0.956
Complete Records	2.511	0.093	0.091	0.960	0.000	0.000	0.000	0.000
FCS	2.509	0.032	0.033	0.938	0.690	0.010	0.010	0.944
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-2.059	0.004	0.004	0.950	-0.788	0.032	0.033	0.942
Complete Records	-1.990	0.015	0.004	0.000	-0.791	0.114	0.115	0.949
FCS	-2.059	0.005	0.005	0.945	-0.790	0.040	0.041	0.944
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.624	0.028	0.029	0.943	-0.949	0.046	0.048	0.944
Complete Records	0.621	0.100	0.102	0.947	-0.958	0.163	0.164	0.948
FCS	0.623	0.035	0.035	0.941	-0.950	0.057	0.058	0.942

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table A. 14. Simulation results when two chosen studies are completely missing X_1. Data generated under scenario 5, e_i: N ~ (0, 2), using equation 3.1.

e_i:N~(0,2)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	2.510	0.265	0.272	0.943	0.690	0.086	0.085	0.956

Complete Records	2.513	0.279	0.286	0.945	0.000	0.000	0.000	0.000
FCS	2.512	0.272	0.278	0.937	0.687	0.100	0.104	0.943
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-2.060	0.043	0.042	0.950	-0.793	0.325	0.335	0.942
Complete Records	-1.991	0.044	0.042	0.678	-0.796	0.342	0.353	0.943
FCS	-2.059	0.044	0.043	0.946	-0.795	0.334	0.344	0.946
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.623	0.285	0.294	0.943	-0.952	0.464	0.484	0.944
Complete Records	0.621	0.300	0.310	0.941	-0.960	0.489	0.506	0.940
FCS	0.622	0.293	0.300	0.950	-0.948	0.476	0.495	0.946

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table A. 15. Simulation results when two chosen studies are completely missing X_1. Data generated under scenario 5, e_i: N ~ (0, 20), using equation 3.1.

e_i:N~(0,20)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	2.533	2.650	2.720	0.943	0.716	0.857	0.849	0.956
Complete Records	2.534	2.650	2.721	0.941	0.000	0.000	0.000	0.000
FCS	2.532	2.655	2.724	0.943	0.696	1.048	1.077	0.937
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-2.065	0.428	0.424	0.950	-0.840	3.248	3.346	0.942
Complete Records	-1.994	0.420	0.419	0.943	-0.843	3.247	3.344	0.941
FCS	-2.062	0.433	0.430	0.945	-0.838	3.254	3.350	0.942
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.618	2.847	2.938	0.943	-0.974	4.639	4.841	0.944
Complete Records	0.617	2.846	2.939	0.941	-0.977	4.638	4.822	0.945
FCS	0.618	2.852	2.943	0.942	-0.960	4.647	4.854	0.940

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Scenario 6 (D=10, N: different per study)

Table A. 16. Simulation results when four chosen studies are completely missing X_1. Data generated under scenario 6, e_i: N ~ (0, 0.2), using equation 3.1.

e_i:N~(0,0.2)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov

Full Data	2.508	0.011	0.012	0.955	0.687	0.004	0.004	0.963
Complete Records	2.509	0.040	0.038	0.958	0.000	0.000	0.000	0.000
FCS	2.508	0.014	0.015	0.934	0.688	0.004	0.004	0.941
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-2.059	0.002	0.002	0.949	-0.787	0.014	0.014	0.948
Complete Records	-1.990	0.006	0.002	0.000	-0.790	0.049	0.048	0.954
FCS	-2.059	0.002	0.002	0.949	-0.787	0.017	0.018	0.923
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.624	0.012	0.012	0.950	-0.949	0.020	0.019	0.953
Complete Records	0.622	0.043	0.043	0.939	-0.951	0.069	0.066	0.953
FCS	0.624	0.015	0.016	0.937	-0.949	0.025	0.026	0.925

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table A. 17. Simulation results when four chosen studies are completely missing X_1. Data generated under scenario 6, e_i: N ~ (0, 2), using equation 3.1.

e_i:N~(0,2)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	2.505	0.113	0.115	0.955	0.687	0.037	0.036	0.963
Complete Records	2.507	0.119	0.121	0.945	0.000	0.000	0.000	0.000
FCS	2.505	0.117	0.119	0.949	0.687	0.045	0.045	0.939
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-2.059	0.018	0.018	0.949	-0.785	0.139	0.139	0.948
Complete Records	-1.990	0.019	0.018	0.031	-0.788	0.146	0.148	0.950
FCS	-2.059	0.019	0.018	0.957	-0.785	0.143	0.146	0.943
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.625	0.122	0.122	0.950	-0.952	0.197	0.193	0.953
Complete Records	0.623	0.128	0.129	0.947	-0.953	0.207	0.202	0.953
FCS	0.625	0.126	0.126	0.952	-0.950	0.203	0.203	0.954

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table A. 18. Simulation results when four chosen studies are completely missing X_1. Data generated under scenario 6, e_i: N ~ (0, 20), using equation 3.1.

e_i:N~(0,20)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	2.481	1.132	1.152	0.955	0.685	0.366	0.360	0.963

Complete Records	2.484	1.133	1.152	0.954	0.000	0.000	0.000	0.000
FCS	2.481	1.133	1.152	0.951	0.679	0.464	0.462	0.942
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-2.056	0.183	0.178	0.949	-0.761	1.387	1.393	0.948
Complete Records	-1.987	0.179	0.175	0.936	-0.765	1.388	1.394	0.947
FCS	-2.055	0.185	0.180	0.950	-0.762	1.388	1.395	0.949
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.633	1.216	1.218	0.950	-0.979	1.967	1.933	0.953
Complete Records	0.630	1.217	1.219	0.951	-0.979	1.967	1.933	0.955
FCS	0.634	1.217	1.219	0.949	-0.978	1.968	1.936	0.953

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Scenario 7 (D=2, N=100 per study, e_1: N ~ (0, 1.2), e_2: N ~ (0, 1.3))

Table A. 19. Simulation results when one chosen study is completely missing X_1. Data generated under scenario 7, e_1: N ~ (0, 1.2), e_2: N ~ (0, 1.3), using equation 3.1.

e_1: N~(0,1.2), e_2: N~(0,1.3)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	2.512	0.286	0.287	0.953	0.690	0.091	0.089	0.964
Complete Records	2.509	0.324	0.314	0.960	0.000	0.000	0.000	0.000
FCS	2.507	0.324	0.328	0.948	0.655	0.117	0.126	0.927
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-2.056	0.046	0.045	0.947	-0.800	0.350	0.360	0.944
Complete Records	-1.987	0.051	0.044	0.745	-0.799	0.397	0.401	0.945
FCS	-2.050	0.051	0.049	0.956	-0.793	0.395	0.415	0.937
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.623	0.307	0.312	0.943	-0.977	0.504	0.502	0.955
Complete Records	0.627	0.348	0.343	0.955	-0.968	0.572	0.561	0.951
FCS	0.629	0.347	0.357	0.935	-0.971	0.569	0.567	0.948

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Scenario 8 (D=2, N=200 per study, e_1: N ~ (0, 1.2), e_2: N ~ (0, 1.3))

Table A. 20. Simulation results when one chosen study is completely missing X_1. Data generated under scenario 8, e_1: N ~ (0, 1.2), e_2: N ~ (0, 1.3), using equation 3.1.

e_1: N~(0,1.2), e_2: N~(0,1.3)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	2.509	0.200	0.197	0.950	0.688	0.064	0.062	0.962
Complete Records	2.505	0.227	0.216	0.967	0.000	0.000	0.000	0.000
FCS	2.516	0.226	0.225	0.949	0.654	0.082	0.083	0.925
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-2.058	0.032	0.032	0.954	-0.787	0.245	0.237	0.952
Complete Records	-1.989	0.036	0.031	0.517	-0.781	0.278	0.270	0.957
FCS	-2.053	0.036	0.033	0.959	-0.791	0.276	0.273	0.955
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.623	0.215	0.214	0.955	-0.948	0.350	0.349	0.954
Complete Records	0.627	0.244	0.238	0.958	-0.938	0.397	0.380	0.955
FCS	0.617	0.242	0.247	0.949	-0.961	0.396	0.399	0.946

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Scenario 9 (D=2, N=500 per study, e_1: N ~ (0, 1.2), e_2: N ~ (0, 1.3))

Table A. 21. Simulation results when one chosen study is completely missing X_1. Data generated under scenario 9, e_1: N ~ (0, 1.2), e_2: N ~ (0, 1.3), using equation 3.1.

e_1: N~(0,1.2), e_2: N~(0,1.3)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	2.503	0.126	0.126	0.951	0.686	0.040	0.040	0.952
Complete Records	2.501	0.143	0.137	0.963	0.000	0.000	0.000	0.000
FCS	2.507	0.142	0.141	0.948	0.654	0.052	0.055	0.884
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-2.058	0.020	0.020	0.948	-0.783	0.154	0.158	0.949
Complete Records	-1.989	0.023	0.020	0.093	-0.780	0.175	0.173	0.954
FCS	-2.054	0.023	0.021	0.963	-0.789	0.174	0.178	0.946
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.629	0.135	0.133	0.946	-0.947	0.218	0.223	0.947
Complete Records	0.630	0.153	0.146	0.955	-0.942	0.248	0.249	0.954
FCS	0.625	0.152	0.150	0.945	-0.951	0.246	0.252	0.949

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Scenario 10 (D=2, N=1000 per study, e_1: N ~ (0, 1.2), e_2: N ~ (0, 1.3))

Table A. 22. Simulation results when one chosen study is completely missing X_1. Data generated under scenario 10, e_1: N ~ (0, 1.2), e_2: N ~ (0, 1.3), using equation 3.1.

e_1: N~(0,1.2), e_2: N~(0,1.3)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	2.509	0.088	0.089	0.952	0.687	0.029	0.028	0.957
Complete Records	2.509	0.100	0.099	0.955	0.000	0.000	0.000	0.000
FCS	2.510	0.100	0.099	0.952	0.654	0.036	0.037	0.847
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-2.060	0.014	0.014	0.951	-0.787	0.108	0.109	0.951
Complete Records	-1.991	0.016	0.014	0.004	-0.789	0.123	0.123	0.954
FCS	-2.056	0.016	0.015	0.955	-0.786	0.122	0.122	0.947
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.623	0.095	0.096	0.951	-0.950	0.154	0.150	0.948
Complete Records	0.624	0.108	0.106	0.956	-0.950	0.174	0.172	0.947
FCS	0.622	0.107	0.107	0.949	-0.950	0.173	0.169	0.952

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Appendix B: Varying granularity of categorical variables

In this section, we give simulation results for 10 scenarios presented in Chapter 4. Full Data refers to complete true data, Complete Records refers to traditional data integration – keep the lowest level that is common in analysis model. FCS refers to fully conditional specification - multiple imputation, FCSgroup: imputation model includes informative 'group' variable. e_i refers to model error applied in outcome's data generating mechanism (equation 5.1).

Simulations with studies of same sizes

Scenario 1 (D=2, N=200 per study)

Table B. 1. Simulation results when data missingness (due to granularity problem) is applied to X_3 from D_1. Data generated under scenario 1, e_i: N ~ (0, 0.2), using equation 4.1.

e_i:N~(0,0.2)	**β₀**				**β₁**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.493	0.032	0.031	0.952	-1.749	0.010	0.011	0.942
Complete Records	2.466	0.039	0.065	0.000	-1.748	0.022	0.022	0.957
FCS	1.493	0.043	0.036	0.974	-1.748	0.013	0.012	0.967
FCSgroup	1.510	0.044	0.034	0.974	-1.748	0.012	0.011	0.972
	β₂				**β₃**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.426	0.005	0.005	0.958	1.460	0.039	0.039	0.956
Complete Records	0.425	0.011	0.011	0.941	0.000	0.000	0.000	0.000
FCS	0.425	0.006	0.005	0.978	1.464	0.050	0.042	0.982
FCSgroup	0.425	0.006	0.005	0.975	1.431	0.051	0.042	0.957
	β₄				**β₅**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.954	0.034	0.033	0.950	0.485	0.056	0.056	0.950
Complete Records	-1.927	0.047	0.067	0.000	-0.488	0.105	0.081	0.000
FCS	-0.946	0.045	0.037	0.980	0.455	0.079	0.068	0.943
FCSgroup	-0.971	0.046	0.037	0.974	0.470	0.067	0.058	0.964

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table B. 2. Simulation results when data missingness (due to granularity problem) is applied to X_3 from D_1. Data generated under scenario 1, e_i: N ~ (0, 2), using equation 4.1.

e_i:N~(0,2)	**β₀**				**β₁**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.487	0.319	0.309	0.952	-1.750	0.103	0.106	0.942
Complete Records	2.469	0.186	0.190	0.001	-1.750	0.104	0.108	0.940

FCS	1.678	0.464	0.482	0.916	-1.750	0.107	0.111	0.944
FCSgroup	1.474	0.371	0.422	0.916	-1.751	0.104	0.108	0.937
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.427	0.051	0.050	0.958	1.474	0.391	0.388	0.956
Complete Records	0.426	0.052	0.050	0.956	0.000	0.000	0.000	0.000
FCS	0.427	0.053	0.052	0.950	1.527	0.562	0.583	0.932
FCSgroup	0.427	0.052	0.051	0.954	1.491	0.473	0.549	0.896
	β_4				**β_5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.943	0.343	0.333	0.950	0.492	0.558	0.561	0.950
Complete Records	-1.925	0.226	0.232	0.011	-0.490	0.501	0.499	0.499
FCS	-0.967	0.484	0.507	0.940	0.514	0.659	0.680	0.936
FCSgroup	-0.930	0.392	0.442	0.926	0.506	0.590	0.623	0.938

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table B. 3. Simulation results when data missingness (due to granularity problem) is applied to X_3 from D_1. Data generated under scenario 1, e_i: N ~ (0, 20), using equation 4.1.

e_i:N~(0,20)	**β_0**				**β_1**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.429	3.191	3.093	0.952	-1.770	1.027	1.063	0.942
Complete Records	2.500	1.831	1.793	0.921	-1.771	1.026	1.062	0.944
FCS	1.554	4.519	4.664	0.942	-1.769	1.029	1.068	0.941
FCSgroup	1.312	3.773	4.408	0.905	-1.776	1.030	1.072	0.940
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.437	0.513	0.495	0.958	1.610	3.908	3.879	0.956
Complete Records	0.435	0.513	0.494	0.959	0.000	0.000	0.000	0.000
FCS	0.436	0.515	0.497	0.959	1.703	5.557	5.693	0.940
FCSgroup	0.437	0.515	0.498	0.958	1.754	4.883	5.879	0.881
	β_4				**β_5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.836	3.427	3.328	0.950	0.561	5.578	5.606	0.950
Complete Records	-1.907	2.216	2.212	0.922	-0.511	4.919	4.932	0.948
FCS	-0.738	4.705	4.902	0.943	0.603	6.442	6.589	0.939
FCSgroup	-0.719	3.977	4.622	0.911	0.677	5.939	6.341	0.924

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Scenario 2 (D=2, N=1000 per study)

Table B. 4. Simulation results when data missingness (due to granularity problem) is applied to X_3 from D_1. Data generated under scenario 2, e_i: N ~ (0, 0.2), using equation 4.1.

e_i:N~(0,0.2)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.494	0.014	0.014	0.955	-1.748	0.005	0.005	0.944
Complete Records	2.467	0.017	0.029	0.000	-1.748	0.010	0.010	0.949
FCS	1.485	0.016	0.015	0.943	-1.748	0.005	0.005	0.958
FCSgroup	1.498	0.016	0.014	0.957	-1.748	0.005	0.005	0.953
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.425	0.002	0.002	0.954	1.458	0.017	0.016	0.959
Complete Records	0.425	0.005	0.005	0.959	0.000	0.000	0.000	0.000
FCS	0.425	0.002	0.002	0.957	1.471	0.019	0.017	0.931
FCSgroup	0.425	0.002	0.002	0.961	1.452	0.018	0.017	0.956
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.956	0.015	0.015	0.957	0.483	0.025	0.024	0.953
Complete Records	-1.928	0.021	0.029	0.000	-0.489	0.046	0.036	0.000
FCS	-0.943	0.017	0.015	0.917	0.466	0.028	0.027	0.906
FCSgroup	-0.959	0.017	0.015	0.966	0.480	0.026	0.024	0.961

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table B. 5. Simulation results when data missingness (due to granularity problem) is applied to X_3 from D_1. Data generated under scenario 2, e_i: N ~ (0, 2), using equation 4.1.

e_i:N~(0,2)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.500	0.142	0.140	0.954	-1.749	0.046	0.046	0.944
Complete Records	2.469	0.083	0.085	0.000	-1.748	0.046	0.047	0.946
FCS	1.706	0.205	0.205	0.813	-1.749	0.047	0.048	0.949
FCSgroup	1.499	0.164	0.176	0.925	-1.749	0.046	0.047	0.940
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.426	0.023	0.023	0.954	1.453	0.174	0.164	0.958
Complete Records	0.425	0.023	0.023	0.954	0.000	0.000	0.000	0.000
FCS	0.426	0.024	0.024	0.954	1.498	0.249	0.243	0.947
FCSgroup	0.426	0.023	0.023	0.950	1.455	0.211	0.225	0.927
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.961	0.152	0.150	0.956	0.470	0.246	0.238	0.953
Complete Records	-1.930	0.101	0.098	0.000	-0.499	0.221	0.211	0.008
FCS	-0.995	0.214	0.212	0.944	0.490	0.291	0.290	0.947
FCSgroup	-0.959	0.174	0.184	0.932	0.471	0.260	0.265	0.942

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table B. 6. Simulation results when data missingness (due to granularity problem) is applied to X_3 from D_1. Data generated under scenario 2, e_i: N ~ (0, 20), using equation 4.1.

e_i:N~(0,20)	**β_0**				**β_1**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.555	1.417	1.400	0.954	-1.753	0.457	0.463	0.944
Complete Records	2.491	0.817	0.806	0.786	-1.753	0.457	0.464	0.944
FCS	1.742	1.997	1.971	0.947	-1.753	0.458	0.463	0.942
FCSgroup	1.547	1.669	1.839	0.915	-1.753	0.457	0.463	0.943
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.427	0.229	0.229	0.954	1.403	1.736	1.639	0.958
Complete Records	0.426	0.229	0.229	0.954	0.000	0.000	0.000	0.000
FCS	0.427	0.229	0.229	0.954	1.401	2.457	2.379	0.956
FCSgroup	0.427	0.229	0.229	0.950	1.415	2.176	2.424	0.918
	β_4				**β_5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-1.008	1.522	1.504	0.956	0.338	2.459	2.382	0.953
Complete Records	-1.944	0.988	0.954	0.835	-0.598	2.169	2.086	0.924
FCS	-0.959	2.082	2.042	0.951	0.382	2.843	2.784	0.950
FCSgroup	-1.001	1.760	1.908	0.924	0.346	2.615	2.697	0.939

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Scenario 3 (D=5, N =200 per study)

Table B. 7. Simulation results when data missingness (due to granularity problem) is applied to X_3 from D_4 and D_5. Data generated under scenario 3, e_i: N ~ (0, 0.2), using equation 4.1.

e_i:N~(0,0.2)	**β_0**				**β_1**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.494	0.020	0.020	0.939	-1.749	0.006	0.006	0.952
Complete Records	2.467	0.025	0.042	0.000	-1.749	0.014	0.013	0.956
FCS	1.489	0.023	0.021	0.960	-1.748	0.007	0.007	0.962
FCSgroup	1.499	0.023	0.021	0.953	-1.748	0.007	0.006	0.957
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.425	0.003	0.003	0.959	1.459	0.025	0.024	0.956
Complete Records	0.426	0.007	0.007	0.950	0.000	0.000	0.000	0.000
FCS	0.425	0.004	0.003	0.962	1.467	0.027	0.025	0.964
FCSgroup	0.425	0.003	0.003	0.968	1.451	0.027	0.025	0.963

	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.956	0.022	0.022	0.947	0.486	0.035	0.035	0.951
Complete Records	-1.928	0.030	0.042	0.000	-0.487	0.065	0.049	0.000
FCS	-0.947	0.024	0.022	0.955	0.470	0.041	0.037	0.951
FCSgroup	-0.960	0.024	0.023	0.954	0.482	0.037	0.035	0.958

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table B. 8. Simulation results when data missingness (due to granularity problem) is applied to X_3 from D_4 and D_5. Data generated under scenario 3, e_i: N ~ (0, 2), using equation 4.1.

e_i:N~(0,2)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.496	0.201	0.203	0.939	-1.751	0.065	0.064	0.952
Complete Records	2.470	0.118	0.122	0.000	-1.751	0.066	0.065	0.948
FCS	1.665	0.265	0.272	0.907	-1.750	0.067	0.066	0.956
FCSgroup	1.502	0.227	0.244	0.926	-1.751	0.065	0.065	0.947
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.426	0.032	0.031	0.959	1.461	0.246	0.242	0.956
Complete Records	0.426	0.033	0.032	0.961	0.000	0.000	0.000	0.000
FCS	0.426	0.033	0.032	0.957	1.480	0.322	0.318	0.947
FCSgroup	0.426	0.033	0.032	0.953	1.453	0.289	0.303	0.934
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.957	0.216	0.218	0.947	0.498	0.349	0.345	0.951
Complete Records	-1.931	0.142	0.146	0.000	-0.476	0.313	0.303	0.133
FCS	-0.997	0.278	0.285	0.931	0.496	0.394	0.396	0.942
FCSgroup	-0.963	0.241	0.254	0.932	0.492	0.365	0.370	0.946

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table B. 9. Simulation results when data missingness (due to granularity problem) is applied to X_3 from D_4 and D_5. Data generated under scenario 3, e_i: N ~ (0, 20), using equation 4.1.

e_i:N~(0,20)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.518	2.012	2.030	0.939	-1.771	0.649	0.639	0.952
Complete Records	2.504	1.158	1.150	0.858	-1.772	0.648	0.639	0.949
FCS	1.809	2.600	2.654	0.932	-1.772	0.649	0.640	0.950
FCSgroup	1.602	2.302	2.473	0.926	-1.772	0.649	0.639	0.952
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov

Full Data	0.431	0.324	0.313	0.959	1.481	2.463	2.416	0.956
Complete Records	0.431	0.324	0.313	0.965	0.000	0.000	0.000	0.000
FCS	0.430	0.325	0.313	0.963	1.360	3.195	3.133	0.954
FCSgroup	0.431	0.325	0.313	0.962	1.362	2.970	3.155	0.932
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.974	2.160	2.182	0.947	0.617	3.490	3.451	0.951
Complete Records	-1.960	1.400	1.415	0.900	-0.369	3.076	3.016	0.940
FCS	-1.097	2.727	2.777	0.942	0.480	3.872	3.873	0.941
FCSgroup	-1.058	2.434	2.587	0.924	0.533	3.668	3.732	0.947

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Scenario 4 (D=5, N=1000 per study)

Table B. 10. Simulation results when data missingness (due to granularity problem) is applied to X_3 from D_2 and D_5. Data generated under scenario 4, e_i: N ~ (0, 0.2), using equation 4.1.

e_i:N~(0,0.2)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.494	0.009	0.009	0.934	-1.748	0.003	0.003	0.947
Complete Records	2.466	0.011	0.018	0.000	-1.748	0.006	0.006	0.948
FCS	1.485	0.010	0.010	0.863	-1.748	0.003	0.003	0.951
FCSgroup	1.495	0.009	0.009	0.937	-1.748	0.003	0.003	0.951
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.425	0.001	0.001	0.949	1.459	0.011	0.011	0.940
Complete Records	0.425	0.003	0.003	0.945	0.000	0.000	0.000	0.000
FCS	0.425	0.002	0.001	0.956	1.469	0.012	0.012	0.849
FCSgroup	0.425	0.001	0.001	0.951	1.457	0.011	0.012	0.941
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.955	0.010	0.010	0.938	0.485	0.016	0.016	0.948
Complete Records	-1.927	0.013	0.018	0.000	-0.487	0.029	0.022	0.000
FCS	-0.945	0.010	0.010	0.831	0.474	0.016	0.017	0.898
FCSgroup	-0.956	0.010	0.010	0.934	0.484	0.016	0.016	0.953

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table B. 11. Simulation results when data missingness (due to granularity problem) is applied to X_3 from D_2 and D_5. Data generated under scenario 4, e_i: N ~ (0, 2), using equation 4.1.

e_i:N~(0,2)	**β_0**				**β_1**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.492	0.089	0.093	0.934	-1.750	0.029	0.029	0.947
Complete Records	2.465	0.053	0.053	0.000	-1.750	0.029	0.030	0.953
FCS	1.649	0.117	0.126	0.722	-1.750	0.030	0.030	0.950
FCSgroup	1.492	0.101	0.112	0.923	-1.750	0.029	0.030	0.949
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.425	0.014	0.014	0.949	1.460	0.110	0.115	0.940
Complete Records	0.425	0.015	0.015	0.950	0.000	0.000	0.000	0.000
FCS	0.425	0.015	0.015	0.956	1.493	0.143	0.152	0.923
FCSgroup	0.425	0.015	0.015	0.949	1.461	0.129	0.144	0.915
	β_4				**β_5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.952	0.096	0.100	0.938	0.489	0.155	0.156	0.948
Complete Records	-1.925	0.064	0.066	0.000	-0.484	0.139	0.135	0.000
FCS	-0.979	0.123	0.133	0.921	0.502	0.175	0.182	0.941
FCSgroup	-0.951	0.107	0.117	0.927	0.490	0.162	0.167	0.943

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table B. 12. Simulation results when data missingness (due to granularity problem) is applied to X_3 from D_2 and D_5. Data generated under scenario 4, e_i: N ~ (0, 20), using equation 4.1.

e_i:N~(0,20)	**β_0**				**β_1**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.480	0.894	0.929	0.934	-1.762	0.289	0.292	0.947
Complete Records	2.460	0.517	0.508	0.537	-1.762	0.289	0.292	0.946
FCS	1.643	1.152	1.225	0.925	-1.762	0.289	0.292	0.950
FCSgroup	1.483	1.030	1.128	0.929	-1.762	0.289	0.292	0.951
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.423	0.144	0.143	0.949	1.472	1.096	1.150	0.940
Complete Records	0.423	0.145	0.143	0.950	0.000	0.000	0.000	0.000
FCS	0.423	0.145	0.143	0.949	1.484	1.417	1.491	0.941
FCSgroup	0.423	0.145	0.143	0.949	1.467	1.334	1.487	0.924
	β_4				**β_5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.925	0.961	1.004	0.938	0.528	1.553	1.564	0.948
Complete Records	-1.905	0.625	0.636	0.683	-0.453	1.371	1.339	0.901
FCS	-0.915	1.209	1.292	0.932	0.536	1.718	1.762	0.946

FCSgroup	-0.927	1.089	1.186	0.941	0.525	1.636	1.683	0.946

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Simulations with studies of different sizes

Scenario 5 (D=5, N: different per study)

Table B. 13. Simulation results when data missingness (due to granularity problem) is applied to X_3 from D_4 and D_5. Data generated under scenario 5, e_i: N ~ (0, 0.2), using equation 4.1.

e_i:N~(0,0.2)	**β_0**				**β_1**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.494	0.027	0.027	0.958	-1.748	0.009	0.009	0.933
Complete Records	2.467	0.032	0.056	0.000	-1.749	0.018	0.018	0.948
FCS	1.491	0.030	0.027	0.971	-1.748	0.009	0.009	0.959
FCSgroup	1.500	0.031	0.027	0.970	-1.748	0.009	0.009	0.957
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.425	0.004	0.004	0.957	1.457	0.033	0.032	0.959
Complete Records	0.425	0.009	0.009	0.961	0.000	0.000	0.000	0.000
FCS	0.425	0.005	0.004	0.965	1.461	0.036	0.032	0.975
FCSgroup	0.425	0.005	0.004	0.962	1.447	0.036	0.033	0.964
	β_4				**β_5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.956	0.029	0.028	0.958	0.486	0.046	0.048	0.944
Complete Records	-1.928	0.039	0.057	0.000	-0.487	0.086	0.069	0.000
FCS	-0.950	0.032	0.028	0.971	0.472	0.055	0.051	0.955
FCSgroup	-0.961	0.032	0.029	0.970	0.481	0.049	0.048	0.962

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table B. 14. Simulation results when data missingness (due to granularity problem) is applied to X_3 from D_4 and D_5. Data generated under scenario 5, e_i: N ~ (0, 2), using equation 4.1.

e_i:N~(0,2)	**β_0**				**β_1**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.497	0.266	0.266	0.958	-1.748	0.086	0.089	0.933
Complete Records	2.455	0.155	0.165	0.000	-1.749	0.087	0.091	0.938
FCS	1.606	0.323	0.323	0.949	-1.748	0.087	0.091	0.937
FCSgroup	1.491	0.292	0.302	0.942	-1.748	0.086	0.090	0.938
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.425	0.043	0.042	0.957	1.437	0.325	0.316	0.959

Complete Records	0.425	0.043	0.042	0.958	0.000	0.000	0.000	0.000
FCS	0.426	0.044	0.042	0.958	1.461	0.394	0.393	0.945
FCSgroup	0.426	0.043	0.042	0.958	1.447	0.368	0.387	0.931
	β_4				**β_5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.957	0.285	0.281	0.958	0.497	0.461	0.481	0.944
Complete Records	-1.915	0.187	0.193	0.000	-0.462	0.413	0.438	0.365
FCS	-0.973	0.341	0.337	0.958	0.509	0.502	0.513	0.944
FCSgroup	-0.951	0.310	0.316	0.951	0.502	0.477	0.500	0.943

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table B. 15. Simulation results when data missingness (due to granularity problem) is applied to X_3 from D_4 and D_5. Data generated under scenario 5, e_i: N ~ (0, 20), using equation 4.1.

e_i:N~(0,20)	**β_0**				**β_1**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.525	2.657	2.659	0.958	-1.746	0.855	0.895	0.933
Complete Records	2.344	1.525	1.536	0.922	-1.747	0.855	0.894	0.931
FCS	1.607	3.188	3.168	0.954	-1.747	0.856	0.893	0.933
FCSgroup	1.458	2.953	3.089	0.932	-1.747	0.856	0.893	0.933
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.426	0.428	0.417	0.957	1.238	3.251	3.164	0.959
Complete Records	0.426	0.427	0.416	0.957	0.000	0.000	0.000	0.000
FCS	0.426	0.428	0.416	0.954	1.288	3.917	3.893	0.950
FCSgroup	0.426	0.428	0.416	0.956	1.344	3.771	3.994	0.929
	β_4				**β_5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.967	2.853	2.812	0.958	0.606	4.608	4.811	0.944
Complete Records	-1.786	1.843	1.827	0.928	-0.213	4.057	4.355	0.928
FCS	-0.934	3.365	3.315	0.956	0.652	4.946	5.043	0.944
FCSgroup	-0.900	3.131	3.223	0.943	0.673	4.786	5.013	0.950

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Scenario 6 (D=10, N: different per study)

Table B. 16. Simulation results when data missingness (due to granularity problem) is applied to X_3 from D_3 , D_6 , D_9 and D_{10}. Data generated under scenario 6, e_i: N ~ (0, 0.2), using equation 4.1.

e_i:N~(0,0.2)	**β_0**				**β_1**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov

Full Data	1.494	0.011	0.011	0.953	-1.748	0.004	0.004	0.938
Complete Records	2.467	0.014	0.024	0.000	-1.748	0.008	0.008	0.944
FCS	1.487	0.012	0.011	0.932	-1.748	0.004	0.004	0.946
FCSgroup	1.495	0.012	0.011	0.961	-1.748	0.004	0.004	0.944
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.425	0.002	0.002	0.945	1.459	0.014	0.014	0.957
Complete Records	0.425	0.004	0.004	0.939	0.000	0.000	0.000	0.000
FCS	0.425	0.002	0.002	0.952	1.468	0.015	0.014	0.930
FCSgroup	0.425	0.002	0.002	0.946	1.457	0.014	0.014	0.957
	β_4				**β_5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.955	0.012	0.012	0.953	0.484	0.020	0.020	0.950
Complete Records	-1.928	0.017	0.024	0.000	-0.488	0.037	0.028	0.000
FCS	-0.947	0.013	0.012	0.920	0.475	0.021	0.021	0.926
FCSgroup	-0.956	0.013	0.012	0.958	0.483	0.020	0.020	0.950

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table B. 17. Simulation results when data missingness (due to granularity problem) is applied to X_3 from D_3 , D_6 , D_9 and D_{10}. Data generated under scenario 6, e_i: N ~ (0, 2), using equation 4.1.

e_i:N~(0,2)	**β_0**				**β_1**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.493	0.113	0.111	0.953	-1.747	0.037	0.037	0.938
Complete Records	2.467	0.066	0.069	0.000	-1.747	0.037	0.038	0.941
FCS	1.627	0.143	0.141	0.847	-1.747	0.037	0.038	0.941
FCSgroup	1.493	0.126	0.130	0.946	-1.747	0.037	0.038	0.940
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.424	0.018	0.018	0.945	1.461	0.139	0.135	0.957
Complete Records	0.424	0.019	0.019	0.944	0.000	0.000	0.000	0.000
FCS	0.425	0.019	0.019	0.951	1.488	0.174	0.171	0.948
FCSgroup	0.425	0.018	0.019	0.941	1.462	0.160	0.165	0.936
	β_4				**β_5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.954	0.122	0.120	0.953	0.485	0.196	0.198	0.950
Complete Records	-1.928	0.080	0.083	0.000	-0.489	0.176	0.178	0.000
FCS	-0.976	0.150	0.149	0.947	0.498	0.217	0.219	0.949
FCSgroup	-0.954	0.133	0.137	0.939	0.485	0.204	0.208	0.946

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table B. 18. Simulation results when data missingness (due to granularity problem) is applied to X_3 from D_3 , D_6 , D_9 and D_{10}. Data generated under scenario 6, e_i: N ~ (0, 20), using equation 4.1.

e_i:N~(0,20)	**β₀**				**β₁**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.484	1.132	1.112	0.953	-1.737	0.366	0.373	0.938
Complete Records	2.473	0.653	0.648	0.677	-1.737	0.366	0.373	0.941
FCS	1.621	1.404	1.377	0.962	-1.737	0.366	0.373	0.939
FCSgroup	1.486	1.270	1.311	0.942	-1.737	0.366	0.373	0.940
	β₂				**β₃**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.416	0.183	0.184	0.945	1.482	1.386	1.351	0.957
Complete Records	0.416	0.183	0.184	0.945	0.000	0.000	0.000	0.000
FCS	0.416	0.183	0.184	0.944	1.494	1.730	1.686	0.950
FCSgroup	0.416	0.183	0.184	0.944	1.477	1.630	1.707	0.937
	β₄				**β₅**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.941	1.216	1.197	0.953	0.494	1.962	1.977	0.950
Complete Records	-1.930	0.790	0.793	0.768	-0.495	1.731	1.771	0.906
FCS	-0.928	1.478	1.452	0.955	0.517	2.140	2.155	0.957
FCSgroup	-0.944	1.345	1.385	0.939	0.492	2.046	2.091	0.947

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Studies with different model errors

Scenario 7 (D=2, N=100 per study, e_1: N ~ (0, 1.2), e_2: N ~ (0, 1.3))

Table B. 19. Simulation results when data missingness (due to granularity problem) is applied to X_3 from D_1. Data generated under scenario 7, e_1: N ~ (0, 1.2), e_2: N ~ (0, 1.3), using equation 4.1.

e_1: N~(0,1.2), e_2: N~(0,1.3)	**β₀**				**β₁**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.501	0.286	0.291	0.944	-1.743	0.091	0.092	0.935
Complete Records	2.471	0.170	0.186	0.001	-1.743	0.095	0.095	0.948
FCS	1.668	0.423	0.464	0.900	-1.743	0.098	0.099	0.939
FCSgroup	1.531	0.334	0.397	0.914	-1.743	0.093	0.097	0.936
	β₂				**β₃**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.420	0.046	0.046	0.941	1.455	0.350	0.367	0.939
Complete Records	0.420	0.048	0.048	0.942	0.000	0.000	0.000	0.000
FCS	0.420	0.049	0.048	0.944	1.472	0.502	0.555	0.934
FCSgroup	0.420	0.047	0.048	0.939	1.402	0.416	0.505	0.901
	β₄				**β₅**			

	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.963	0.307	0.313	0.948	0.479	0.507	0.490	0.964
Complete Records	-1.932	0.206	0.219	0.005	-0.490	0.467	0.445	0.435
FCS	-1.012	0.441	0.480	0.932	0.495	0.600	0.621	0.933
FCSgroup	-0.992	0.352	0.413	0.913	0.452	0.538	0.562	0.949

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Scenario 8 (D=2, N=200 per study, e_1: N ~ (0, 1.2), e_2: N ~ (0, 1.3))

Table B. 20. Simulation results when data missingness (due to granularity problem) is applied to X_3 from D_1. Data generated under scenario 8, e_1: N ~ (0, 1.2), e_2: N ~ (0, 1.3), using equation 4.1.

e_1: N~(0,1.2), e_2: N~(0,1.3)	**β_0**				**β_1**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.490	0.200	0.194	0.952	-1.750	0.064	0.067	0.941
Complete Records	2.468	0.120	0.129	0.000	-1.750	0.067	0.069	0.939
FCS	1.661	0.294	0.315	0.883	-1.750	0.069	0.072	0.942
FCSgroup	1.523	0.231	0.258	0.919	-1.750	0.066	0.069	0.932
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.426	0.032	0.031	0.960	1.469	0.244	0.243	0.953
Complete Records	0.426	0.034	0.032	0.957	0.000	0.000	0.000	0.000
FCS	0.426	0.034	0.034	0.954	1.478	0.347	0.366	0.934
FCSgroup	0.426	0.033	0.032	0.952	1.411	0.289	0.322	0.909
	β_4				**β_5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.948	0.214	0.209	0.952	0.489	0.349	0.350	0.947
Complete Records	-1.926	0.145	0.154	0.000	-0.489	0.322	0.316	0.143
FCS	-0.997	0.306	0.331	0.925	0.508	0.414	0.434	0.934
FCSgroup	-0.981	0.244	0.270	0.928	0.457	0.370	0.386	0.946

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Scenario 9 (D=2, N=500 per study, e_1: N ~ (0, 1.2), e_2: N ~ (0, 1.3))

Table B. 21. Simulation results when data missingness (due to granularity problem) is applied to X_3 from D_1. Data generated under scenario 9, e_1: N ~ (0, 1.2), e_2: N ~ (0, 1.3), using equation 4.1.

e_1: N~(0,1.2), e_2: N~(0,1.3)	**β_0**				**β_1**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.496	0.126	0.124	0.960	-1.750	0.040	0.040	0.946

Complete Records	2.468	0.076	0.083	0.000	-1.750	0.042	0.042	0.952
FCS	1.678	0.183	0.201	0.808	-1.751	0.043	0.043	0.946
FCSgroup	1.536	0.145	0.164	0.911	-1.750	0.041	0.041	0.947
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.425	0.020	0.021	0.956	1.458	0.154	0.154	0.956
Complete Records	0.425	0.021	0.022	0.949	0.000	0.000	0.000	0.000
FCS	0.425	0.022	0.022	0.937	1.465	0.217	0.237	0.922
FCSgroup	0.425	0.021	0.022	0.941	1.395	0.182	0.209	0.894
	β_4				**β_5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.955	0.135	0.135	0.959	0.479	0.219	0.214	0.955
Complete Records	-1.927	0.091	0.098	0.000	-0.493	0.201	0.193	0.000
FCS	-1.013	0.191	0.211	0.903	0.498	0.258	0.279	0.932
FCSgroup	-0.995	0.153	0.172	0.911	0.439	0.231	0.244	0.930

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Scenario 10 (D=2, N=1000 per study, e_1: N ~ (0, 1.2), e_2: N ~ (0, 1.3))

Table B. 22. Simulation results when data missingness (due to granularity problem) is applied to X_3 from D_1. Data generated under scenario 10, e_1: N ~ (0, 1.2), e_2: N ~ (0, 1.3), using equation 4.1.

e_1: N~(0,1.2), e_2: N~(0,1.3)	**β_0**				**β_1**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.497	0.089	0.087	0.953	-1.749	0.029	0.029	0.942
Complete Records	2.468	0.053	0.057	0.000	-1.748	0.030	0.030	0.950
FCS	1.681	0.129	0.135	0.675	-1.749	0.031	0.031	0.948
FCSgroup	1.538	0.102	0.109	0.904	-1.749	0.029	0.030	0.937
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.425	0.014	0.014	0.956	1.455	0.109	0.103	0.959
Complete Records	0.425	0.015	0.015	0.952	0.000	0.000	0.000	0.000
FCS	0.426	0.015	0.015	0.955	1.457	0.153	0.154	0.940
FCSgroup	0.426	0.015	0.015	0.948	1.389	0.128	0.134	0.899
	β_4				**β_5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.959	0.095	0.094	0.958	0.475	0.154	0.149	0.953
Complete Records	-1.929	0.065	0.065	0.000	-0.495	0.142	0.134	0.000
FCS	-1.018	0.134	0.139	0.919	0.495	0.182	0.185	0.943
FCSgroup	-1.000	0.108	0.114	0.918	0.434	0.163	0.164	0.936

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Appendix C: Mixed numeric and non-numeric data types

In this appendix, we present simulation results for 10 scenarios presented in Chapter 5. Full Data refers to complete true data, Complete Records refers to traditional data integration – mapping to common levels - where mixed type problematic variable is converted to categorical in analysis model. FCS refers to fully conditional specification - multiple imputation, FCSgroup: FCS including informative 'group' variable. e_i refers to model error applied in outcome's data generating mechanism (equation 5.1).

Simulations with studies of same sizes

Scenario 1 (D=2, N=200 per study)

Table C. 1. Simulation results when data missingness (due to mixed type problem) is applied to X_2 from D_2. Data generated under scenario 1, e_i: N ~ (0, 0.2), using equation 5.1.

e_i:N~(0,0.2)	**β₀**				**β₁**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.493	0.032	0.032	0.953	-1.749	0.010	0.010	0.949
Complete Records	0.824	0.092	0.086	0.000	-1.686	0.028	0.023	0.353
FCS	1.492	0.048	0.049	0.933	-1.748	0.015	0.015	0.928
FCSgroup	1.492	0.046	0.048	0.939	-1.748	0.015	0.015	0.939
	β₂				**β₃**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.426	0.005	0.005	0.950	1.460	0.039	0.038	0.956
Complete Records	1.338	0.056	0.028	0.000	1.461	0.107	0.106	0.956
FCS	0.430	0.007	0.007	0.891	1.460	0.059	0.059	0.930
FCSgroup	0.430	0.007	0.007	0.892	1.461	0.056	0.058	0.927
	β₄				**β₅**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.954	0.034	0.033	0.956	0.484	0.056	0.057	0.952
Complete Records	-0.954	0.094	0.092	0.955	0.491	0.153	0.151	0.955
FCS	-0.954	0.051	0.051	0.940	0.486	0.084	0.087	0.935
FCSgroup	-0.953	0.049	0.051	0.936	0.486	0.081	0.084	0.932

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table C. 2. Simulation results when data missingness (due to mixed type problem) is applied to X_2 from D_2. Data generated under scenario 1, e_i: N ~ (0, 2), using equation 5.1.

e_i:N~(0,2)	**β₀**				**β₁**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.482	0.321	0.322	0.953	-1.751	0.103	0.103	0.949
Complete Records	0.811	0.348	0.346	0.490	-1.689	0.105	0.104	0.915
FCS	1.479	0.348	0.348	0.955	-1.746	0.112	0.107	0.957

FCSgroup	1.483	0.333	0.333	0.952	-1.747	0.107	0.106	0.949
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.427	0.051	0.051	0.950	1.473	0.393	0.385	0.956
Complete Records	1.342	0.210	0.195	0.004	1.473	0.405	0.400	0.955
FCS	0.423	0.067	0.073	0.912	1.477	0.427	0.417	0.955
FCSgroup	0.420	0.061	0.063	0.934	1.473	0.407	0.399	0.954
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.942	0.344	0.335	0.956	0.481	0.559	0.569	0.952
Complete Records	-0.942	0.355	0.346	0.957	0.488	0.577	0.592	0.950
FCS	-0.938	0.374	0.367	0.961	0.485	0.606	0.626	0.942
FCSgroup	-0.943	0.357	0.347	0.953	0.478	0.579	0.591	0.949

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table C. 3. Simulation results when data missingness (due to mixed type problem) is applied to X_2 from D_2. Data generated under scenario 1, e_i: N ~ (0, 20), using equation 5.1.

e_i:N~(0,20)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.375	3.207	3.219	0.953	-1.775	1.028	1.026	0.949
Complete Records	0.682	3.369	3.358	0.951	-1.713	1.022	1.020	0.947
FCS	1.382	3.232	3.244	0.950	-1.766	1.052	1.040	0.952
FCSgroup	1.382	3.213	3.229	0.953	-1.767	1.036	1.028	0.955
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.436	0.514	0.512	0.950	1.600	3.927	3.847	0.956
Complete Records	1.389	2.039	1.935	0.928	1.599	3.929	3.847	0.954
FCS	0.430	0.748	0.760	0.927	1.583	3.960	3.876	0.953
FCSgroup	0.423	0.606	0.605	0.946	1.591	3.935	3.857	0.953
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.817	3.442	3.348	0.956	0.449	5.588	5.694	0.952
Complete Records	-0.821	3.443	3.343	0.956	0.451	5.591	5.701	0.954
FCS	-0.821	3.469	3.374	0.953	0.447	5.629	5.761	0.954
FCSgroup	-0.824	3.448	3.359	0.959	0.437	5.597	5.715	0.951

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Scenario 2 (D=2, N=1000 per study)

Table C. 4. Simulation results when data missingness (due to mixed type problem) is applied to X_2 from D_4 and D_5. Data generated under scenario 2, e_i: N ~ (0, 0.2), using equation 5.1.

e_i:N~(0,0.2)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.493	0.014	0.015	0.936	-1.748	0.005	0.005	0.947
Complete Records	0.823	0.041	0.038	0.000	-1.685	0.012	0.010	0.000
FCS	1.494	0.020	0.022	0.923	-1.748	0.007	0.007	0.944
FCSgroup	1.493	0.020	0.022	0.914	-1.749	0.006	0.006	0.940
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.425	0.002	0.002	0.954	1.459	0.017	0.018	0.942
Complete Records	1.338	0.025	0.013	0.000	1.461	0.047	0.046	0.952
FCS	0.427	0.003	0.003	0.925	1.459	0.025	0.026	0.932
FCSgroup	0.427	0.003	0.003	0.899	1.460	0.024	0.026	0.926
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.955	0.015	0.016	0.940	0.485	0.025	0.026	0.953
Complete Records	-0.953	0.042	0.041	0.956	0.482	0.067	0.067	0.953
FCS	-0.955	0.022	0.023	0.926	0.484	0.035	0.037	0.929
FCSgroup	-0.955	0.021	0.023	0.909	0.485	0.035	0.037	0.921

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table C. 5. Simulation results when data missingness (due to mixed type problem) is applied to X_2 from D_4 and D_5. Data generated under scenario 2, e_i: N ~ (0, 2), using equation 5.1.

e_i:N~(0,2)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.489	0.142	0.146	0.936	-1.747	0.046	0.045	0.947
Complete Records	0.817	0.154	0.159	0.014	-1.684	0.047	0.046	0.728
FCS	1.489	0.153	0.159	0.935	-1.746	0.050	0.047	0.962
FCSgroup	1.488	0.147	0.153	0.946	-1.746	0.048	0.046	0.956
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.425	0.023	0.023	0.954	1.463	0.174	0.178	0.942
Complete Records	1.341	0.094	0.093	0.000	1.464	0.179	0.181	0.945
FCS	0.425	0.030	0.031	0.934	1.464	0.188	0.193	0.940
FCSgroup	0.424	0.027	0.028	0.943	1.464	0.180	0.185	0.941
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.951	0.152	0.156	0.940	0.487	0.246	0.255	0.953
Complete Records	-0.950	0.157	0.160	0.947	0.485	0.253	0.263	0.942
FCS	-0.952	0.165	0.169	0.936	0.487	0.267	0.276	0.949
FCSgroup	-0.950	0.158	0.163	0.941	0.487	0.254	0.263	0.945

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table C. 6. Simulation results when data missingness (due to mixed type problem) is applied to X_2 from D_4 and D_5. Data generated under scenario 2, e_i: N ~ (0, 20), using equation 5.1.

e_i:N~(0,20)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.441	1.418	1.464	0.936	-1.736	0.457	0.452	0.947
Complete Records	0.757	1.489	1.562	0.909	-1.676	0.454	0.450	0.949
FCS	1.440	1.421	1.465	0.936	-1.734	0.468	0.457	0.956
FCSgroup	1.440	1.419	1.464	0.936	-1.735	0.461	0.453	0.951
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.425	0.228	0.231	0.954	1.501	1.736	1.776	0.942
Complete Records	1.368	0.907	0.922	0.822	1.500	1.737	1.776	0.943
FCS	0.423	0.333	0.329	0.939	1.503	1.741	1.778	0.943
FCSgroup	0.421	0.272	0.278	0.936	1.501	1.738	1.777	0.942
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.913	1.523	1.561	0.940	0.511	2.456	2.554	0.953
Complete Records	-0.913	1.523	1.560	0.940	0.506	2.456	2.556	0.955
FCS	-0.913	1.527	1.562	0.939	0.510	2.462	2.552	0.950
FCSgroup	-0.912	1.524	1.561	0.940	0.512	2.457	2.551	0.952

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Scenario 3 (D=5, N =200 per study)

Table C. 7. Simulation results when data missingness (due to mixed type problem) is applied to X_2 from D_2 and D_5. Data generated under scenario 3, e_i: N ~ (0, 0.2), using equation 5.1.

e_i:N~(0,0.2)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.492	0.020	0.020	0.946	-1.748	0.006	0.006	0.957
Complete Records	0.824	0.057	0.054	0.000	-1.685	0.018	0.014	0.016
FCS	1.492	0.026	0.027	0.932	-1.749	0.009	0.009	0.941
FCSgroup	1.493	0.025	0.026	0.936	-1.749	0.008	0.008	0.945
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.425	0.003	0.003	0.954	1.460	0.025	0.024	0.952
Complete Records	1.337	0.035	0.018	0.000	1.463	0.067	0.064	0.954
FCS	0.427	0.004	0.004	0.934	1.460	0.032	0.033	0.931
FCSgroup	0.427	0.004	0.004	0.918	1.460	0.031	0.032	0.934

	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.954	0.022	0.022	0.948	0.486	0.035	0.035	0.946
Complete Records	-0.955	0.059	0.058	0.951	0.487	0.095	0.092	0.955
FCS	-0.954	0.028	0.029	0.926	0.486	0.046	0.045	0.939
FCSgroup	-0.954	0.027	0.028	0.938	0.485	0.045	0.044	0.949

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table C. 8. Simulation results when data missingness (due to mixed type problem) is applied to X_2 from D_2 and D_5. Data generated under scenario 3, e_i: N ~ (0, 2), using equation 5.1.

e_i:N~(0,2)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.478	0.201	0.201	0.946	-1.749	0.065	0.063	0.957
Complete Records	0.809	0.217	0.218	0.114	-1.685	0.066	0.064	0.850
FCS	1.478	0.213	0.212	0.950	-1.748	0.069	0.065	0.960
FCSgroup	1.476	0.205	0.205	0.953	-1.748	0.067	0.064	0.959
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.426	0.032	0.033	0.954	1.474	0.246	0.242	0.952
Complete Records	1.337	0.132	0.131	0.000	1.477	0.254	0.247	0.954
FCS	0.424	0.039	0.041	0.937	1.474	0.260	0.258	0.952
FCSgroup	0.424	0.036	0.037	0.945	1.475	0.251	0.247	0.947
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.941	0.215	0.216	0.948	0.504	0.349	0.349	0.946
Complete Records	-0.942	0.222	0.221	0.953	0.504	0.360	0.357	0.948
FCS	-0.942	0.228	0.227	0.952	0.502	0.370	0.365	0.954
FCSgroup	-0.939	0.221	0.218	0.955	0.506	0.357	0.357	0.951

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table C. 9. Simulation results when data missingness (due to mixed type problem) is applied to X_2 from D_4 and D_5. Data generated under scenario 3, e_i: N ~ (0, 20), using equation 5.1.

e_i:N~(0,20)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.332	2.006	2.007	0.946	-1.752	0.648	0.630	0.957
Complete Records	0.663	2.107	2.128	0.935	-1.688	0.644	0.626	0.959
FCS	1.329	2.010	2.011	0.945	-1.747	0.659	0.641	0.956
FCSgroup	1.332	2.007	2.010	0.944	-1.751	0.652	0.636	0.960
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov

Full Data	0.427	0.324	0.326	0.954	1.611	2.458	2.424	0.952
Complete Records	1.341	1.284	1.299	0.890	1.613	2.459	2.417	0.951
FCS	0.418	0.431	0.435	0.940	1.613	2.464	2.428	0.954
FCSgroup	0.428	0.368	0.371	0.946	1.611	2.460	2.427	0.954
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.814	2.155	2.156	0.948	0.675	3.493	3.487	0.946
Complete Records	-0.816	2.156	2.153	0.950	0.677	3.494	3.487	0.945
FCS	-0.810	2.160	2.158	0.952	0.681	3.501	3.498	0.952
FCSgroup	-0.813	2.156	2.157	0.948	0.678	3.495	3.492	0.946

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Scenario 4 (D=5, N=1000 per study)

Table C. 10. Simulation results when data missingness (due to mixed type problem) is applied to X_2 from D_2 and D_5. Data generated under scenario 4, e_i: N ~ (0, 0.2), using equation 5.1.

e_i:N~(0,0.2)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.495	0.027	0.026	0.945	-1.748	0.009	0.009	0.950
Complete Records	0.825	0.076	0.071	0.000	-1.685	0.023	0.019	0.178
FCS	1.494	0.032	0.033	0.937	-1.748	0.010	0.011	0.931
FCSgroup	1.494	0.032	0.033	0.940	-1.748	0.010	0.011	0.929
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.425	0.004	0.004	0.955	1.459	0.033	0.033	0.944
Complete Records	1.337	0.046	0.024	0.000	1.461	0.089	0.088	0.949
FCS	0.427	0.005	0.005	0.938	1.459	0.040	0.041	0.948
FCSgroup	0.427	0.005	0.005	0.940	1.460	0.039	0.041	0.939
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.956	0.029	0.028	0.950	0.485	0.047	0.046	0.954
Complete Records	-0.954	0.078	0.076	0.943	0.490	0.127	0.127	0.955
FCS	-0.956	0.035	0.036	0.941	0.486	0.057	0.058	0.937
FCSgroup	-0.955	0.034	0.035	0.934	0.486	0.056	0.058	0.942

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table C. 11. Simulation results when data missingness (due to mixed type problem) is applied to X_2 from D_2 and D_5. Data generated under scenario 4, e_i: N ~ (0, 2), using equation 5.1.

e_i:N~(0,2)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.501	0.267	0.265	0.945	-1.746	0.086	0.088	0.950
Complete Records	0.832	0.289	0.289	0.368	-1.683	0.088	0.089	0.871
FCS	1.502	0.277	0.274	0.955	-1.745	0.089	0.090	0.951
FCSgroup	1.501	0.271	0.271	0.948	-1.746	0.087	0.089	0.958
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.425	0.043	0.042	0.955	1.457	0.326	0.331	0.944
Complete Records	1.336	0.175	0.174	0.001	1.460	0.336	0.344	0.943
FCS	0.424	0.049	0.050	0.942	1.456	0.339	0.343	0.950
FCSgroup	0.423	0.046	0.045	0.956	1.458	0.331	0.339	0.946
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.958	0.286	0.285	0.950	0.489	0.466	0.465	0.954
Complete Records	-0.957	0.295	0.294	0.952	0.494	0.480	0.483	0.955
FCS	-0.960	0.297	0.293	0.955	0.492	0.483	0.484	0.953
FCSgroup	-0.959	0.291	0.290	0.948	0.492	0.473	0.473	0.954

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table C. 12. Simulation results when data missingness (due to mixed type problem) is applied to X_2 from D_2 and D_5. Data generated under scenario 4, e_i: N ~ (0, 20), using equation 5.1.

e_i:N~(0,20)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.563	2.668	2.649	0.945	-1.729	0.856	0.877	0.950
Complete Records	0.896	2.799	2.803	0.946	-1.667	0.850	0.874	0.955
FCS	1.564	2.673	2.659	0.944	-1.728	0.865	0.885	0.951
FCSgroup	1.563	2.670	2.652	0.946	-1.729	0.859	0.879	0.949
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.420	0.428	0.417	0.955	1.441	3.258	3.313	0.944
Complete Records	1.331	1.693	1.723	0.914	1.447	3.259	3.320	0.945
FCS	0.419	0.523	0.524	0.936	1.436	3.264	3.320	0.940
FCSgroup	0.422	0.466	0.451	0.953	1.443	3.260	3.317	0.945
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.984	2.863	2.849	0.950	0.534	4.655	4.648	-0.984
Complete Records	-0.981	2.864	2.852	0.949	0.539	4.656	4.658	-0.981
FCS	-0.985	2.868	2.856	0.951	0.533	4.663	4.665	-0.985

FCSgroup	-0.984	2.865	2.848	0.951	0.539	4.657	4.652	-0.984

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Simulations with studies of different sizes

Scenario 5 (D=5, N: different per study)

Table C. 13. Simulation results when data missingness (due to mixed type problem) is applied to X_2 from D_4 and D_5. Data generated under scenario 5, e_i: N ~ (0, 0.2), using equation 5.1.

e_i:N~(0,0.2)	**β_0**				**β_1**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.495	0.027	0.026	0.945	-1.748	0.009	0.009	0.950
Complete Records	0.825	0.076	0.071	0.000	-1.685	0.023	0.019	0.178
FCS	1.494	0.032	0.033	0.937	-1.748	0.010	0.011	0.931
FCSgroup	1.494	0.032	0.033	0.940	-1.748	0.010	0.011	0.929
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.425	0.004	0.004	0.955	1.459	0.033	0.033	0.944
Complete Records	1.337	0.046	0.024	0.000	1.461	0.089	0.088	0.949
FCS	0.427	0.005	0.005	0.938	1.459	0.040	0.041	0.948
FCSgroup	0.427	0.005	0.005	0.940	1.460	0.039	0.041	0.939
	β_4				**β_5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.956	0.029	0.028	0.950	0.485	0.047	0.046	0.954
Complete Records	-0.954	0.078	0.076	0.943	0.490	0.127	0.127	0.955
FCS	-0.956	0.035	0.036	0.941	0.486	0.057	0.058	0.937
FCSgroup	-0.955	0.034	0.035	0.934	0.486	0.056	0.058	0.942

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table C. 14. Simulation results when data missingness (due to mixed type problem) is applied to X_2 from D_4 and D_5. Data generated under scenario 5, e_i: N ~ (0, 2), using equation 5.1.

e_i:N~(0,2)	**β_0**				**β_1**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.501	0.267	0.265	0.945	-1.746	0.086	0.088	0.950
Complete Records	0.832	0.289	0.289	0.368	-1.683	0.088	0.089	0.871
FCS	1.502	0.277	0.274	0.955	-1.745	0.089	0.090	0.951
FCSgroup	1.501	0.271	0.271	0.948	-1.746	0.087	0.089	0.958
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.425	0.043	0.042	0.955	1.457	0.326	0.331	0.944

Complete Records	1.336	0.175	0.174	0.001	1.460	0.336	0.344	0.943
FCS	0.424	0.049	0.050	0.942	1.456	0.339	0.343	0.950
FCSgroup	0.423	0.046	0.045	0.956	1.458	0.331	0.339	0.946
	β_4				**β_5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.958	0.286	0.285	0.950	0.489	0.466	0.465	0.954
Complete Records	-0.957	0.295	0.294	0.952	0.494	0.480	0.483	0.955
FCS	-0.960	0.297	0.293	0.955	0.492	0.483	0.484	0.953
FCSgroup	-0.959	0.291	0.290	0.948	0.492	0.473	0.473	0.954

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table C. 15. Simulation results when data missingness (due to mixed type problem) is applied to X_2 from D_4 and D_5. Data generated under scenario 5, e_i: N ~ (0, 20), using equation 5.1.

e_i:N~(0,20)	**β_0**				**β_1**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.563	2.668	2.649	0.945	-1.729	0.856	0.877	0.950
Complete Records	0.896	2.799	2.803	0.946	-1.667	0.850	0.874	0.955
FCS	1.564	2.673	2.659	0.944	-1.728	0.865	0.885	0.951
FCSgroup	1.563	2.670	2.652	0.946	-1.729	0.859	0.879	0.949
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.420	0.428	0.417	0.955	1.441	3.258	3.313	0.944
Complete Records	1.331	1.693	1.723	0.914	1.447	3.259	3.320	0.945
FCS	0.419	0.523	0.524	0.936	1.436	3.264	3.320	0.940
FCSgroup	0.422	0.466	0.451	0.953	1.443	3.260	3.317	0.945
	β_4				**β_5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.984	2.863	2.849	0.950	0.534	4.655	4.648	0.954
Complete Records	-0.981	2.864	2.852	0.949	0.539	4.656	4.658	0.954
FCS	-0.985	2.868	2.856	0.951	0.533	4.663	4.665	0.954
FCSgroup	-0.984	2.865	2.848	0.951	0.539	4.657	4.652	0.953

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Simulations with studies of different sizes

Scenario 5 (D=5, N: different per study)

Table C. 16. Simulation results when data missingness (due to mixed type problem) is applied to X_2 from D_3 , D_6 , D_9 and D_{10}. Data generated under scenario 6, e_i: N ~ (0, 0.2), using equation 5.1.

e_i:N~(0,0.2)	**β_0**				**β_1**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.494	0.011	0.011	0.951	-1.748	0.004	0.004	0.946
Complete Records	0.827	0.032	0.031	0.000	-1.685	0.010	0.008	0.000
FCS	1.495	0.014	0.014	0.933	-1.748	0.005	0.005	0.935
FCSgroup	1.495	0.014	0.015	0.939	-1.749	0.004	0.004	0.939
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.425	0.002	0.002	0.950	1.458	0.014	0.014	0.948
Complete Records	1.337	0.020	0.010	0.000	1.457	0.038	0.039	0.950
FCS	0.426	0.002	0.002	0.935	1.458	0.017	0.018	0.940
FCSgroup	0.426	0.002	0.002	0.929	1.458	0.017	0.018	0.941
	β_4				**β_5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.956	0.012	0.012	0.956	0.485	0.020	0.020	0.955
Complete Records	-0.958	0.033	0.033	0.950	0.484	0.054	0.053	0.947
FCS	-0.956	0.015	0.016	0.943	0.485	0.025	0.025	0.946
FCSgroup	-0.956	0.015	0.016	0.935	0.485	0.024	0.024	0.939

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Scenario 6 (D=10, N: different per study)

Table C. 17. Simulation results when data missingness (due to mixed type problem) is applied to X_2 from D_3 , D_6 , D_9 and D_{10}. Data generated under scenario 6, e_i: N ~ (0, 2), using equation 5.1.

e_i:N~(0,2)	**β_0**				**β_1**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.499	0.113	0.115	0.951	-1.749	0.037	0.037	0.946
Complete Records	0.831	0.123	0.123	0.001	-1.686	0.037	0.037	0.617
FCS	1.499	0.119	0.121	0.946	-1.749	0.039	0.038	0.960
FCSgroup	1.500	0.116	0.117	0.951	-1.749	0.037	0.037	0.944
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.425	0.018	0.019	0.950	1.452	0.139	0.137	0.948
Complete Records	1.339	0.075	0.075	0.000	1.451	0.143	0.142	0.952
FCS	0.425	0.022	0.023	0.938	1.452	0.145	0.146	0.945
FCSgroup	0.425	0.020	0.020	0.952	1.451	0.142	0.139	0.950
	β_4				**β_5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.962	0.122	0.123	0.956	0.486	0.196	0.196	0.955

Complete Records	-0.964	0.126	0.127	0.950	0.485	0.203	0.201	0.954
FCS	-0.962	0.128	0.130	0.949	0.484	0.205	0.210	0.946
FCSgroup	-0.963	0.124	0.126	0.954	0.485	0.200	0.202	0.948

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table C. 18. Simulation results when data missingness (due to mixed type problem) is applied to X_2 from D_3 , D_6 , D_9 and D_{10}. Data generated under scenario 6, e_i: N ~ (0, 20), using equation 5.1.

e_i:N~(0,20)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.542	1.134	1.149	0.951	-1.755	0.366	0.369	0.946
Complete Records	0.869	1.191	1.203	0.912	-1.694	0.363	0.367	0.949
FCS	1.543	1.135	1.151	0.950	-1.755	0.371	0.371	0.953
FCSgroup	1.544	1.135	1.150	0.951	-1.756	0.368	0.371	0.944
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.424	0.183	0.186	0.950	1.389	1.388	1.367	0.948
Complete Records	1.351	0.725	0.747	0.731	1.388	1.389	1.367	0.950
FCS	0.425	0.232	0.239	0.941	1.389	1.389	1.370	0.949
FCSgroup	0.425	0.202	0.206	0.950	1.387	1.389	1.367	0.949
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-1.019	1.218	1.233	0.956	0.499	1.965	1.965	0.955
Complete Records	-1.021	1.219	1.233	0.955	0.498	1.965	1.963	0.954
FCS	-1.019	1.219	1.235	0.956	0.497	1.967	1.969	0.954
FCSgroup	-1.020	1.219	1.233	0.956	0.500	1.965	1.966	0.954

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Studies with different model errors

Scenario 7 (D=2, N=100 per study, e_1: N ~ (0, 1.2), e_2: N ~ (0, 1.3))

Table C. 19. Simulation results when data missingness (due to mixed type problem) is applied to X_2 from D_2. Data generated under scenario 7, e_1: N ~ (0, 1.2), e_2: N ~ (0, 1.3), using equation 5.1.

e_1: N~(0,1.2), e_2: N~(0,1.3)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.493	0.285	0.279	0.955	-1.744	0.091	0.091	0.953
Complete Records	0.825	0.323	0.318	0.455	-1.682	0.098	0.096	0.900
FCS	1.487	0.331	0.318	0.961	-1.741	0.104	0.097	0.961

FCSgroup	1.492	0.306	0.298	0.955	-1.742	0.098	0.095	0.951
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.423	0.046	0.046	0.953	1.464	0.349	0.348	0.953
Complete Records	1.337	0.195	0.187	0.001	1.462	0.376	0.371	0.948
FCS	0.435	0.054	0.058	0.902	1.469	0.405	0.397	0.953
FCSgroup	0.428	0.054	0.056	0.923	1.465	0.374	0.368	0.949
	β_4				**β_5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.946	0.306	0.301	0.945	0.499	0.509	0.478	0.968
Complete Records	-0.946	0.330	0.323	0.946	0.502	0.548	0.531	0.959
FCS	-0.938	0.355	0.344	0.956	0.506	0.589	0.562	0.956
FCSgroup	-0.944	0.329	0.320	0.950	0.497	0.549	0.518	0.953

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Scenario 8 (D=2, N=200 per study, e_1: N ~ (0, 1.2), e_2: N ~ (0, 1.3))

Table C. 20. Simulation results when data missingness (due to mixed type problem) is applied to X_2 from D_2. Data generated under scenario 8, e_1: N ~ (0, 1.2), e_2: N ~ (0, 1.3), using equation 5.1.

e_1: N~(0,1.2), e_2: N~(0,1.3)	**β_0**				**β_1**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.487	0.201	0.202	0.954	-1.750	0.064	0.064	0.952
Complete Records	0.816	0.227	0.225	0.160	-1.688	0.069	0.067	0.858
FCS	1.483	0.232	0.232	0.948	-1.752	0.073	0.069	0.963
FCSgroup	1.487	0.215	0.218	0.947	-1.752	0.069	0.068	0.952
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.426	0.032	0.032	0.952	1.468	0.246	0.241	0.956
Complete Records	1.340	0.138	0.123	0.000	1.468	0.265	0.263	0.955
FCS	0.441	0.037	0.041	0.898	1.471	0.285	0.279	0.959
FCSgroup	0.433	0.038	0.040	0.924	1.467	0.263	0.260	0.951
	β_4				**β_5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.947	0.215	0.210	0.955	0.482	0.350	0.356	0.949
Complete Records	-0.947	0.232	0.227	0.952	0.489	0.377	0.388	0.947
FCS	-0.942	0.249	0.245	0.943	0.488	0.405	0.420	0.942
FCSgroup	-0.947	0.231	0.228	0.945	0.483	0.377	0.386	0.943

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Scenario 9 (D=2, N=500 per study, e_1: N ~ (0, 1.2), e_2: N ~ (0, 1.3))

Table C. 21. Simulation results when data missingness (due to mixed type problem) is applied to X_2 from D_2. Data generated under scenario 9, e_1: N ~ (0, 1.2), e_2: N ~ (0, 1.3), using equation 5.1.

e_1: N~(0,1.2), e_2: N~(0,1.3)	**β₀**				**β₁**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.491	0.126	0.131	0.946	-1.750	0.040	0.041	0.950
Complete Records	0.824	0.142	0.149	0.005	-1.687	0.043	0.043	0.721
FCS	1.489	0.145	0.148	0.932	-1.754	0.047	0.043	0.949
FCSgroup	1.490	0.134	0.137	0.949	-1.752	0.043	0.044	0.950
	β₂				**β₃**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.425	0.020	0.021	0.944	1.458	0.154	0.159	0.942
Complete Records	1.336	0.087	0.082	0.000	1.455	0.166	0.173	0.938
FCS	0.441	0.023	0.025	0.858	1.459	0.177	0.183	0.930
FCSgroup	0.433	0.023	0.025	0.908	1.459	0.165	0.168	0.940
	β₄				**β₅**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.953	0.135	0.142	0.941	0.487	0.218	0.215	0.953
Complete Records	-0.953	0.146	0.154	0.945	0.485	0.236	0.234	0.951
FCS	-0.951	0.156	0.160	0.941	0.491	0.251	0.250	0.945
FCSgroup	-0.952	0.144	0.149	0.941	0.492	0.234	0.231	0.949

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Scenario 10 (D=2, N=1000 per study, e_1: N ~ (0, 1.2), e_2: N ~ (0, 1.3))

Table C. 22. Simulation results when data missingness (due to mixed type problem) is applied to X_2 from D_2. Data generated under scenario 10, e_1: N ~ (0, 1.2), e_2: N ~ (0, 1.3), using equation 5.1.

e_1: N~(0,1.2), e_2: N~(0,1.3)	**β₀**				**β₁**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.490	0.089	0.091	0.939	-1.748	0.029	0.028	0.949
Complete Records	0.819	0.100	0.102	0.000	-1.685	0.031	0.029	0.434
FCS	1.493	0.103	0.107	0.931	-1.753	0.033	0.030	0.969
FCSgroup	1.490	0.095	0.099	0.940	-1.752	0.031	0.030	0.954
	β₂				**β₃**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.425	0.014	0.014	0.952	1.462	0.109	0.111	0.944
Complete Records	1.340	0.061	0.059	0.000	1.463	0.117	0.117	0.945
FCS	0.441	0.016	0.017	0.805	1.461	0.126	0.130	0.936
FCSgroup	0.434	0.016	0.017	0.912	1.462	0.116	0.120	0.935
	β₄				**β₅**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.953	0.095	0.097	0.942	0.486	0.154	0.159	0.948

Complete Records	-0.951	0.103	0.104	0.950	0.484	0.166	0.171	0.938
FCS	-0.955	0.110	0.114	0.935	0.485	0.177	0.187	0.940
FCSgroup	-0.952	0.102	0.105	0.942	0.487	0.165	0.170	0.947

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table C. 23. Coefficients (estimate, standard error, t statistic and p-values) for linear regression model from equation 5.2 in true Full SLE data.

	estimate	**standard error**	**t statistic**	**p-value**
(Intercept)	9.8625	1.1281	8.7430	2.00E-16***
Age	-0.1067	0.0184	-5.7970	0.0000***
Ethnicity				
Caucasian	1.5653	0.5555	2.8180	0.0050**
Other	2.8451	0.5816	4.8920	0.0000***
Creatinine	1.8558	0.4583	4.0490	0.0001***
Gender				
Male	0.4494	0.6316	0.7120	0.4770
BMI	-0.1429	0.0320	-4.4700	0.0000***

*Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Table C. 24. Coefficients (estimate, standard error, t statistic and p-values) for linear regression model from equation 5.2 after applying FCSgroup (5 imputations, 10 iterations) in SLE data.

	estimate	**standard error**	**t statistic**	**p-value**
(Intercept)	9.9518	1.1363	8.7581	0.0000***
Age	-0.1087	0.0196	-5.5363	0.0000***
Ethnicity				
Caucasian	1.5987	0.5585	2.8624	0.0044***
Other	2.8680	0.5827	4.9218	0.0000***
Creatinine	1.9025	0.4588	4.1468	0.0000***
Gender				
Male	0.4584	0.6318	0.7255	0.4685
BMI	-0.1459	0.0320	-4.5641	0.0000***

*Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Table C. 25. Coefficients (estimate, standard error, t statistic and p-values) for linear regression model from equation 5.2 after applying FCS (5 imputations, 10 iterations) in SLE data.

	estimate	**standard error**	**t statistic**	**p-value**
(Intercept)	9.8373	1.1386	8.6399	0.0000***
Age	-0.1058	0.0188	-5.6115	0.0000***
Ethnicity				
Caucasian	1.6092	0.5573	2.8873	0.0040***
Other	2.8860	0.5821	4.9581	0.0000***
Creatinine	1.9031	0.4591	4.1453	0.0000***

Gender				
Male	0.4706	0.6324	0.7441	0.4571
BMI	-0.1464	0.0319	-4.5875	0.0000***

*Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Table C. 26. Coefficients (estimate, standard error, t statistic and p-values) for linear regression model from equation 5.2 after applying complete case analysis (Complete Records) in SLE data.

	estimate	**standard error**	**t statistic**	**p-value**
(Intercept)	8.7826	1.1361	7.7300	0.0000***
Age				
21-40	-2.0013	0.6879	-2.9090	0.0038**
41-60	-4.0000	0.7592	-5.2680	0.0000***
>60	-5.0391	1.7704	-2.8460	0.0046**
Ethnicity				
Caucasian	1.4780	0.5548	2.6640	0.0079
Other	2.8542	0.5850	4.8790	0.0000
Creatinine	1.8614	0.4609	4.0390	0.0001
Gender				
Male	0.5598	0.6317	0.8860	0.3759
BMI	-0.1505	0.0319	-4.7200	0.0000

*Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Appendix D: Combined types of content heterogeneity

This section presents all simulation results for 10 scenarios presented in Chapter 6. Full Data refers to complete true data, Complete Records refers to traditional data integration – mapping to common levels - where mixed type problematic variable is converted to categorical, and the other problematic variable's levels (due to granularity) are mapped to the least granular in order to be included in the analysis model. FCS refers to fully conditional specification - multiple imputation, FCSgroup: imputation model includes only the relevant informative 'group' variable, FCSgroups: both imputation models include both informative 'group' variables, FCS3group2: X_3 is imputed excluding its informative 'group' variable, X_2 is imputed including its informative 'group' variables, e_i refers to model error applied in outcome's data generating mechanism (equation 5.1).

Simulations with studies of same sizes

Scenario 1 (D=2, m=10, it=10, N=200 per study)

Table D. 1. Simulation results when data missingness due to granularity problem is applied to X_3 from D_1 and due to mixed type problem is applied to X_2 from D_2. Data generated under scenario 1, e_i: N ~ (0, 0.2), using equation 5.1.

e_i: N ~ (0, 0.2)	**β₀**				**β₁**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.488	0.032	0.032	0.949	-1.755	0.010	0.010	0.946
Complete Records	1.803	0.069	0.078	0.011	-1.693	0.034	0.030	0.546
FCS	2.055	0.429	0.377	0.835	-1.750	0.025	0.019	0.986
FCSgroup	1.523	0.068	0.055	0.956	-1.752	0.018	0.016	0.973
FCSgroups	1.526	0.071	0.058	0.965	-1.752	0.018	0.016	0.963
FCS3group2	1.575	0.170	0.156	0.993	-1.752	0.020	0.017	0.971
	β₂				**β₃**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.424	0.005	0.005	0.949	1.466	0.039	0.039	0.952
Complete Records	1.331	0.067	0.047	0.000	0.000	0.000	0.000	0.000
FCS	0.427	0.012	0.009	0.986	0.901	0.438	0.380	0.843
FCSgroup	0.427	0.009	0.008	0.963	1.406	0.075	0.060	0.918
FCSgroups	0.427	0.009	0.008	0.957	1.401	0.079	0.065	0.925
FCS3group2	0.428	0.010	0.008	0.976	1.384	0.174	0.156	0.994
	β₄				**β₅**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.957	0.034	0.035	0.947	0.496	0.056	0.057	0.954
Complete Records	-1.936	0.073	0.084	0.000	-0.487	0.163	0.155	0.000
FCS	-1.479	0.418	0.371	0.841	-0.126	0.457	0.380	0.816
FCSgroup	-0.992	0.071	0.059	0.953	0.465	0.103	0.087	0.972
FCSgroups	-0.995	0.074	0.061	0.957	0.460	0.105	0.091	0.965
FCS3group2	-1.027	0.168	0.156	0.991	0.302	0.209	0.184	0.914

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table D. 2. Simulation results when data missingness due to granularity problem is applied to X_3 from D_1 and due to mixed type problem is applied to X_2 from D_2. Data generated under scenario 1, e_i: N ~ (0, 2), using equation 5.1.

e_i: N ~ (0, 2)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.480	0.320	0.312	0.954	-1.756	0.103	0.102	0.953
Complete Records	1.799	0.220	0.226	0.694	-1.694	0.107	0.104	0.912
FCS	1.953	0.785	0.586	0.950	-1.750	0.117	0.109	0.960
FCSgroup	1.476	0.404	0.421	0.934	-1.751	0.108	0.107	0.952
FCSgroups	1.530	0.454	0.433	0.946	-1.752	0.108	0.107	0.953
FCS3group2	1.769	0.601	0.514	0.946	-1.753	0.111	0.108	0.954
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.424	0.051	0.050	0.952	1.478	0.391	0.386	0.950
Complete Records	1.338	0.213	0.201	0.007	0.000	0.000	0.000	0.000
FCS	0.431	0.076	0.078	0.936	1.430	0.968	0.665	0.991
FCSgroup	0.421	0.062	0.062	0.948	1.479	0.522	0.554	0.918
FCSgroups	0.421	0.064	0.062	0.956	1.401	0.603	0.572	0.944
FCS3group2	0.426	0.064	0.064	0.942	1.501	0.713	0.603	0.980
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.946	0.343	0.332	0.954	0.495	0.560	0.547	0.957
Complete Records	-1.932	0.232	0.233	0.008	-0.492	0.519	0.515	0.520
FCS	-1.234	0.802	0.609	0.972	0.244	0.941	0.764	0.972
FCSgroup	-0.942	0.425	0.438	0.932	0.495	0.625	0.630	0.947
FCSgroups	-0.996	0.472	0.451	0.942	0.443	0.661	0.634	0.957
FCS3group2	-1.069	0.618	0.533	0.964	0.414	0.773	0.697	0.961

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table D. 3. Simulation results when data missingness due to granularity problem is applied to X_3 from D_1 and due to mixed type problem is applied to X_2 from D_2. Data generated under scenario 1, e_i: N ~ (0, 20), using equation 5.1.

e_i: N ~ (0, 20)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.553	3.197	3.284	0.938	-1.770	1.027	1.032	0.953
Complete Records	1.901	2.098	2.070	0.949	-1.708	1.019	1.025	0.952
FCS	1.656	4.835	4.628	0.957	-1.752	1.050	1.047	0.953
FCSgroup	1.458	3.761	4.379	0.919	-1.763	1.038	1.037	0.953
FCSgroups	1.470	3.876	4.265	0.921	-1.768	1.039	1.040	0.954
FCS3group2	1.718	4.561	4.534	0.953	-1.763	1.036	1.038	0.952
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.421	0.513	0.511	0.955	1.503	3.916	4.101	0.939
Complete Records	1.314	2.034	2.043	0.918	0.000	0.000	0.000	0.000
FCS	0.416	0.755	0.733	0.947	1.697	6.068	5.878	0.962
FCSgroup	0.423	0.610	0.593	0.956	1.639	4.889	5.957	0.898
FCSgroups	0.433	0.621	0.601	0.953	1.617	5.095	5.760	0.911
FCS3group2	0.419	0.604	0.585	0.951	1.617	5.623	5.784	0.939

	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-1.030	3.432	3.518	0.948	0.185	5.607	5.651	0.959
Complete Records	-2.036	2.218	2.098	0.935	-0.826	4.952	4.869	0.956
FCS	-0.901	5.022	4.858	0.953	0.357	6.705	6.522	0.965
FCSgroup	-0.935	3.964	4.563	0.919	0.282	5.955	6.278	0.946
FCSgroups	-0.947	4.075	4.454	0.923	0.273	6.031	6.183	0.949
FCS3group2	-0.967	4.746	4.758	0.942	0.251	6.476	6.438	0.963

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Scenario 2 (D=2, N=1000 per study, m=10, it=10)

Table D. 4. Simulation results when data missingness due to granularity problem is applied to X_3 from D_1 and due to mixed type problem is applied to X_2 from D_2. Data generated under scenario 2, e_i: N ~ (0, 0.2), using equation 5.1.

e_i: N ~ (0, 0.2)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.488	0.014	0.015	0.937	-1.755	0.005	0.005	0.954
Complete Records	1.799	0.031	0.038	0.000	-1.693	0.015	0.013	0.006
FCS	2.580	0.431	0.248	0.448	-1.752	0.011	0.009	0.972
FCSgroup	1.494	0.023	0.021	0.958	-1.755	0.007	0.007	0.942
FCSgroups	1.495	0.023	0.021	0.950	-1.755	0.007	0.007	0.950
FCS3group2	1.493	0.091	0.084	0.948	-1.755	0.007	0.007	0.940
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.424	0.002	0.002	0.933	1.465	0.017	0.017	0.952
Complete Records	1.332	0.030	0.022	0.000	0.000	0.000	0.000	0.000
FCS	0.425	0.005	0.005	0.974	0.376	0.436	0.251	0.449
FCSgroup	0.425	0.003	0.003	0.919	1.454	0.026	0.025	0.927
FCSgroups	0.425	0.003	0.003	0.926	1.452	0.026	0.025	0.927
FCS3group2	0.425	0.004	0.004	0.926	1.466	0.093	0.084	0.912
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.957	0.015	0.016	0.938	0.498	0.025	0.025	0.953
Complete Records	-1.934	0.033	0.039	0.000	-0.480	0.072	0.066	0.000
FCS	-1.996	0.417	0.242	0.449	-0.717	0.459	0.259	0.426
FCSgroup	-0.962	0.024	0.022	0.950	0.492	0.037	0.034	0.958
FCSgroups	-0.963	0.024	0.022	0.952	0.491	0.037	0.034	0.956
FCS3group2	-0.954	0.090	0.083	0.913	0.420	0.115	0.100	0.859

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table D. 5. Simulation results when data missingness due to granularity problem is applied to X_3 from D_1 and due to mixed type problem is applied to X_2 from D_2. Data generated under scenario 2, e_i: N ~ (0, 2), using equation 5.1.

e_i: N ~ (0, 2)	**β_0**				**β_1**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.485	0.142	0.143	0.947	-1.757	0.046	0.046	0.959
Complete Records	1.801	0.098	0.097	0.110	-1.693	0.048	0.047	0.745
FCS	2.242	0.463	0.310	0.661	-1.762	0.052	0.049	0.962
FCSgroup	1.488	0.177	0.200	0.914	-1.755	0.048	0.048	0.952
FCSgroups	1.520	0.217	0.206	0.950	-1.755	0.048	0.048	0.957
FCS3group2	1.793	0.267	0.238	0.808	-1.758	0.049	0.048	0.957
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.424	0.023	0.022	0.958	1.469	0.174	0.170	0.953
Complete Records	1.331	0.095	0.088	0.000	0.000	0.000	0.000	0.000
FCS	0.442	0.034	0.034	0.911	1.179	0.566	0.334	0.978
FCSgroup	0.423	0.027	0.026	0.958	1.463	0.230	0.256	0.919
FCSgroups	0.421	0.029	0.026	0.967	1.415	0.296	0.270	0.957
FCS3group2	0.428	0.028	0.027	0.956	1.483	0.319	0.277	0.969
	β_4				**β_5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.956	0.152	0.155	0.943	0.504	0.246	0.245	0.945
Complete Records	-1.938	0.104	0.106	0.000	-0.475	0.227	0.221	0.004
FCS	-1.531	0.470	0.319	0.806	-0.026	0.517	0.378	0.876
FCSgroup	-0.959	0.186	0.209	0.917	0.503	0.273	0.286	0.930
FCSgroups	-0.990	0.225	0.216	0.954	0.471	0.301	0.289	0.948
FCS3group2	-1.095	0.275	0.246	0.934	0.417	0.342	0.321	0.953

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table D. 6. Simulation results when data missingness due to granularity problem is applied to X_3 from D_1 and due to mixed type problem is applied to X_2 from D_2. Data generated under scenario 2, e_i: N ~ (0, 20), using equation 5.1.

e_i: N ~ (0, 20)	**β_0**				**β_1**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.460	1.419	1.435	0.947	-1.769	0.457	0.457	0.959
Complete Records	1.802	0.936	0.910	0.938	-1.705	0.454	0.454	0.951
FCS	1.682	2.178	2.172	0.946	-1.765	0.468	0.470	0.954
FCSgroup	1.460	1.655	1.945	0.907	-1.766	0.461	0.461	0.955
FCSgroups	1.484	1.705	1.917	0.914	-1.769	0.462	0.461	0.955
FCS3group2	1.683	2.005	2.131	0.933	-1.766	0.461	0.461	0.957
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.425	0.229	0.218	0.958	1.503	1.737	1.704	0.953
Complete Records	1.325	0.907	0.861	0.856	0.000	0.000	0.000	0.000
FCS	0.421	0.335	0.311	0.962	1.527	2.732	2.582	0.953
FCSgroup	0.421	0.271	0.250	0.969	1.498	2.152	2.577	0.895
FCSgroups	0.427	0.274	0.253	0.970	1.464	2.235	2.541	0.907

FCS3group2	0.419	0.269	0.250	0.970	1.517	2.461	2.536	0.939
	β4				**β5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.954	1.524	1.546	0.943	0.564	2.458	2.452	0.945
Complete Records	-1.959	0.989	0.976	0.823	-0.438	2.167	2.124	0.947
FCS	-0.949	2.257	2.262	0.945	0.576	2.976	2.944	0.947
FCSgroup	-0.954	1.746	2.034	0.910	0.565	2.603	2.767	0.929
FCSgroups	-0.977	1.794	2.010	0.919	0.541	2.636	2.746	0.936
FCS3group2	-0.951	2.089	2.217	0.931	0.572	2.837	2.918	0.937

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Scenario 3 (D=5, N=200 per study, m=10, it=10)

Table D. 7. Simulation results when data missingness due to granularity problem is applied to X_3 from D_2, D_5 and due to mixed type problem is applied to X_2 from D_1, D_4. Data generated under scenario 3, e_i: N ~ (0, 0.2), using equation 5.1.

e_i: N ~ (0, 0.2)	**β0**				**β1**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.489	0.020	0.021	0.944	-1.755	0.006	0.006	0.954
Complete Records	1.799	0.044	0.052	0.000	-1.692	0.021	0.019	0.127
FCS	1.485	0.029	0.027	0.946	-1.755	0.009	0.009	0.943
FCSgroup	1.492	0.027	0.027	0.950	-1.755	0.008	0.008	0.938
FCSgroups	1.492	0.027	0.027	0.928	-1.755	0.009	0.008	0.938
FCS3group2	1.485	0.028	0.027	0.951	-1.755	0.009	0.009	0.950
	β2				**β3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.424	0.003	0.003	0.947	1.465	0.025	0.025	0.950
Complete Records	1.332	0.042	0.030	0.000	0.000	0.000	0.000	0.000
FCS	0.425	0.004	0.005	0.906	1.470	0.035	0.034	0.939
FCSgroup	0.425	0.004	0.004	0.919	1.459	0.033	0.033	0.933
FCSgroups	0.425	0.004	0.004	0.911	1.459	0.032	0.034	0.922
FCS3group2	0.426	0.004	0.004	0.908	1.470	0.033	0.033	0.933
	β4				**β5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.958	0.022	0.022	0.946	0.499	0.035	0.036	0.936
Complete Records	-1.933	0.046	0.055	0.000	-0.478	0.102	0.093	0.000
FCS	-0.952	0.031	0.030	0.940	0.488	0.050	0.049	0.932
FCSgroup	-0.961	0.029	0.030	0.943	0.496	0.046	0.047	0.933
FCSgroups	-0.961	0.029	0.030	0.936	0.496	0.046	0.047	0.932
FCS3group2	-0.952	0.030	0.029	0.936	0.485	0.048	0.049	0.930

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table D. 8. Simulation results when data missingness due to granularity problem is applied to X_3 from D_2, D_5 and due to mixed type problem is applied to X_2 from D_1, D_4. Data generated under scenario 3, e_i: N ~ (0, 2), using equation 5.1.

e_i: N ~ (0, 2)	**β_0**				**β_1**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.484	0.202	0.205	0.946	-1.757	0.065	0.066	0.941
Complete Records	1.805	0.139	0.147	0.388	-1.695	0.068	0.069	0.844
FCS	1.563	0.246	0.250	0.927	-1.754	0.070	0.069	0.948
FCSgroup	1.480	0.222	0.231	0.943	-1.755	0.067	0.068	0.937
FCSgroups	1.481	0.221	0.231	0.944	-1.755	0.067	0.068	0.941
FCS3group2	1.558	0.236	0.240	0.923	-1.754	0.068	0.068	0.945
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.423	0.032	0.032	0.954	1.472	0.247	0.250	0.943
Complete Records	1.328	0.135	0.138	0.000	0.000	0.000	0.000	0.000
FCS	0.420	0.040	0.042	0.928	1.486	0.300	0.305	0.933
FCSgroup	0.421	0.037	0.037	0.940	1.479	0.278	0.284	0.943
FCSgroups	0.422	0.036	0.037	0.942	1.477	0.277	0.288	0.934
FCS3group2	0.421	0.037	0.037	0.950	1.487	0.289	0.287	0.944
	β_4				**β_5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.951	0.216	0.223	0.941	0.509	0.350	0.343	0.958
Complete Records	-1.937	0.147	0.158	0.000	-0.475	0.323	0.313	0.145
FCS	-0.966	0.260	0.269	0.937	0.511	0.392	0.395	0.949
FCSgroup	-0.947	0.236	0.250	0.939	0.514	0.367	0.367	0.947
FCSgroups	-0.948	0.236	0.250	0.932	0.514	0.367	0.367	0.944
FCS3group2	-0.966	0.250	0.260	0.937	0.509	0.380	0.373	0.955

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table D. 9. Simulation results when data missingness due to granularity problem is applied to X_3 from D_2, D_5 and due to mixed type problem is applied to X_2 from D_1, D_4. Data generated under scenario 3, e_i: N ~ (0, 20), using equation 5.1.

e_i: N ~ (0, 20)	**β_0**				**β_1**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.443	2.015	2.047	0.946	-1.773	0.649	0.662	0.941
Complete Records	1.821	1.325	1.392	0.930	-1.711	0.644	0.660	0.942
FCS	1.527	2.268	2.307	0.943	-1.773	0.659	0.673	0.944
FCSgroup	1.436	2.178	2.253	0.941	-1.767	0.652	0.662	0.942
FCSgroups	1.451	2.175	2.254	0.943	-1.769	0.653	0.662	0.945
FCS3group2	1.527	2.263	2.306	0.945	-1.768	0.652	0.662	0.943
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.419	0.324	0.325	0.954	1.534	2.466	2.496	0.943
Complete Records	1.304	1.285	1.336	0.882	0.000	0.000	0.000	0.000
FCS	0.424	0.427	0.443	0.935	1.526	2.781	2.791	0.946
FCSgroup	0.410	0.366	0.370	0.949	1.542	2.755	2.832	0.934
FCSgroups	0.415	0.369	0.369	0.943	1.521	2.748	2.854	0.929
FCS3group2	0.415	0.367	0.369	0.949	1.525	2.777	2.795	0.944
	β_4				**β_5**			

	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.899	2.163	2.230	0.941	0.608	3.497	3.429	0.958
Complete Records	-1.930	1.401	1.471	0.883	-0.419	3.083	3.019	0.940
FCS	-0.902	2.408	2.491	0.936	0.605	3.672	3.602	0.949
FCSgroup	-0.891	2.316	2.429	0.926	0.616	3.596	3.559	0.945
FCSgroups	-0.906	2.314	2.432	0.933	0.605	3.594	3.552	0.949
FCS3group2	-0.901	2.404	2.491	0.937	0.614	3.661	3.593	0.951

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Scenario 4 (D=5, N=1000 per study, m=10, it=10)

Table D. 10. Simulation results when data missingness due to granularity problem is applied to X_3 from D_2, D_5 and due to mixed type problem is applied to X_2 from D_1, D_4. Data generated under scenario 4, e_i: N ~ (0, 0.2), using equation 5.1.

e_i: N ~ (0, 0.2)	**β_0**				**β_1**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.488	0.009	0.009	0.950	-1.755	0.003	0.003	0.950
Complete Records	1.799	0.020	0.024	0.000	-1.693	0.010	0.009	0.000
FCS	1.482	0.013	0.012	0.925	-1.755	0.004	0.004	0.928
FCSgroup	1.489	0.012	0.012	0.936	-1.756	0.004	0.004	0.941
FCSgroups	1.489	0.012	0.012	0.944	-1.755	0.004	0.004	0.933
FCS3group2	1.483	0.012	0.012	0.912	-1.756	0.004	0.004	0.949
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.424	0.001	0.001	0.939	1.466	0.011	0.011	0.947
Complete Records	1.332	0.019	0.014	0.000	0.000	0.000	0.000	0.000
FCS	0.424	0.002	0.002	0.941	1.473	0.015	0.015	0.912
FCSgroup	0.424	0.002	0.002	0.926	1.465	0.014	0.014	0.930
FCSgroups	0.424	0.002	0.002	0.926	1.465	0.014	0.014	0.934
FCS3group2	0.424	0.002	0.002	0.933	1.473	0.015	0.015	0.921
	β_4				**β_5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.957	0.010	0.010	0.945	0.498	0.015	0.016	0.942
Complete Records	-1.933	0.021	0.026	0.000	-0.478	0.045	0.042	0.000
FCS	-0.950	0.013	0.013	0.908	0.490	0.021	0.022	0.910
FCSgroup	-0.957	0.012	0.013	0.942	0.498	0.020	0.020	0.939
FCSgroups	-0.958	0.012	0.013	0.930	0.497	0.020	0.021	0.936
FCS3group2	-0.950	0.013	0.013	0.920	0.490	0.021	0.022	0.906

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table D. 11. Simulation results when data missingness due to granularity problem is applied to X_3 from D_2, D_5 and due to mixed type problem is applied to X_2 from D_1, D_4. Data generated under scenario 4, e_i: N ~ (0, 2), using equation 5.1.

e_i: N ~ (0, 2)	**β_0**	**β_1**

	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.488	0.089	0.088	0.951	-1.757	0.029	0.030	0.951
Complete Records	1.801	0.062	0.063	0.001	-1.694	0.030	0.030	0.469
FCS	1.568	0.109	0.110	0.887	-1.755	0.031	0.031	0.960
FCSgroup	1.488	0.098	0.099	0.952	-1.756	0.030	0.030	0.960
FCSgroups	1.487	0.098	0.099	0.953	-1.756	0.030	0.030	0.955
FCS3group2	1.563	0.105	0.104	0.896	-1.756	0.030	0.031	0.954
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.424	0.014	0.015	0.938	1.469	0.110	0.107	0.950
Complete Records	1.333	0.060	0.061	0.000	0.000	0.000	0.000	0.000
FCS	0.421	0.018	0.018	0.943	1.483	0.133	0.133	0.944
FCSgroup	0.423	0.016	0.017	0.940	1.468	0.123	0.123	0.953
FCSgroups	0.423	0.016	0.016	0.946	1.469	0.123	0.123	0.947
FCS3group2	0.423	0.017	0.017	0.947	1.483	0.128	0.126	0.941
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.957	0.096	0.095	0.950	0.506	0.155	0.158	0.946
Complete Records	-1.936	0.066	0.068	0.000	-0.473	0.143	0.143	0.000
FCS	-0.973	0.116	0.115	0.947	0.509	0.174	0.182	0.943
FCSgroup	-0.958	0.104	0.105	0.950	0.506	0.162	0.168	0.945
FCSgroups	-0.957	0.105	0.105	0.949	0.506	0.163	0.169	0.949
FCS3group2	-0.973	0.111	0.109	0.949	0.508	0.168	0.173	0.948

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table D. 12. Simulation results when data missingness due to granularity problem is applied to X_3 from D_2, D_5 and due to mixed type problem is applied to X_2 from D_1, D_4. Data generated under scenario 4, e_i: N ~ (0, 20), using equation 5.1.

e_i: N ~ (0, 20)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.481	0.895	0.884	0.951	-1.771	0.289	0.295	0.951
Complete Records	1.811	0.591	0.593	0.905	-1.708	0.287	0.293	0.940
FCS	1.563	1.004	0.987	0.957	-1.769	0.294	0.301	0.947
FCSgroup	1.481	0.965	0.963	0.950	-1.770	0.291	0.297	0.948
FCSgroups	1.480	0.961	0.968	0.948	-1.769	0.291	0.297	0.949
FCS3group2	1.565	1.002	0.984	0.954	-1.769	0.291	0.297	0.951
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.426	0.144	0.146	0.938	1.502	1.096	1.072	0.950
Complete Records	1.343	0.573	0.590	0.623	0.000	0.000	0.000	0.000
FCS	0.423	0.191	0.191	0.935	1.498	1.233	1.215	0.958
FCSgroup	0.424	0.164	0.166	0.940	1.501	1.222	1.217	0.951
FCSgroups	0.423	0.164	0.165	0.943	1.504	1.214	1.221	0.940
FCS3group2	0.422	0.164	0.164	0.947	1.497	1.230	1.213	0.955
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.961	0.961	0.949	0.950	0.586	1.550	1.576	0.946

Complete Records	-1.963	0.625	0.636	0.620	-0.416	1.368	1.370	0.904
FCS	-0.960	1.066	1.040	0.956	0.586	1.622	1.658	0.942
FCSgroup	-0.962	1.027	1.024	0.946	0.585	1.593	1.635	0.950
FCSgroups	-0.960	1.023	1.027	0.942	0.586	1.590	1.632	0.955
FCS3group2	-0.964	1.065	1.038	0.957	0.588	1.620	1.644	0.952

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Simulations with studies of different sizes

Scenario 5 (D=5, N: different per study m=5, it=5)

Table D. 13. Simulation results when data missingness due to granularity problem is applied to X_3 from D_4 and due to mixed type problem is applied to X_2 from D_5. Data generated under scenario 5, e_i: N ~ (0, 0.2), using equation 5.1.

e_i: N ~ (0, 0.2)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.487	0.026	0.027	0.949	-1.756	0.009	0.009	0.951
Complete Records	1.798	0.058	0.070	0.000	-1.693	0.028	0.025	0.385
FCS	1.486	0.031	0.031	0.954	-1.755	0.010	0.010	0.944
FCSgroup	1.489	0.030	0.030	0.949	-1.755	0.010	0.010	0.944
FCSgroups	1.489	0.030	0.030	0.945	-1.755	0.010	0.010	0.944
FCS3group2	1.486	0.030	0.030	0.949	-1.755	0.010	0.010	0.947
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.424	0.004	0.004	0.941	1.467	0.032	0.033	0.956
Complete Records	1.332	0.056	0.039	0.000	0.000	0.000	0.000	0.000
FCS	0.424	0.005	0.005	0.939	1.469	0.038	0.037	0.952
FCSgroup	0.424	0.005	0.005	0.944	1.464	0.037	0.037	0.947
FCSgroups	0.424	0.005	0.005	0.939	1.464	0.037	0.036	0.948
FCS3group2	0.424	0.005	0.005	0.946	1.470	0.037	0.037	0.951
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.955	0.028	0.029	0.953	0.499	0.046	0.048	0.930
Complete Records	-1.930	0.061	0.074	0.000	-0.480	0.134	0.128	0.000
FCS	-0.953	0.033	0.033	0.955	0.494	0.055	0.055	0.941
FCSgroup	-0.957	0.033	0.032	0.953	0.497	0.052	0.053	0.950
FCSgroups	-0.958	0.032	0.032	0.948	0.497	0.052	0.052	0.943
FCS3group2	-0.953	0.032	0.032	0.952	0.493	0.054	0.053	0.939

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table D. 14. Simulation results when data missingness due to granularity problem is applied to X_3 from D_4 and due to mixed type problem is applied to X_2 from D_5. Data generated under scenario 5, e_i: N ~ (0, 2), using equation 5.1.

e_i: N ~ (0, 2)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.485	0.266	0.269	0.946	-1.755	0.086	0.085	0.951

Complete Records	1.794	0.183	0.190	0.590	-1.693	0.089	0.086	0.909
FCS	1.529	0.293	0.308	0.930	-1.754	0.088	0.086	0.957
FCSgroup	1.483	0.280	0.289	0.943	-1.754	0.087	0.086	0.954
FCSgroups	1.482	0.280	0.290	0.933	-1.754	0.087	0.085	0.948
FCS3group2	1.525	0.289	0.302	0.933	-1.754	0.087	0.086	0.947
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.421	0.043	0.044	0.934	1.462	0.325	0.332	0.950
Complete Records	1.331	0.177	0.175	0.000	0.000	0.000	0.000	0.000
FCS	0.420	0.047	0.048	0.944	1.474	0.359	0.379	0.932
FCSgroup	0.420	0.045	0.045	0.952	1.465	0.347	0.365	0.934
FCSgroups	0.419	0.045	0.046	0.949	1.465	0.348	0.366	0.934
FCS3group2	0.420	0.045	0.046	0.941	1.475	0.354	0.371	0.930
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.950	0.285	0.290	0.949	0.513	0.461	0.457	0.947
Complete Records	-1.923	0.193	0.198	0.002	-0.464	0.426	0.432	0.370
FCS	-0.955	0.313	0.329	0.929	0.518	0.486	0.493	0.948
FCSgroup	-0.948	0.299	0.310	0.939	0.514	0.472	0.468	0.943
FCSgroups	-0.947	0.299	0.311	0.937	0.515	0.472	0.470	0.941
FCS3group2	-0.951	0.309	0.323	0.936	0.518	0.481	0.484	0.938

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table D. 15. Simulation results when data missingness due to granularity problem is applied to X_3 from D_4 and due to mixed type problem is applied to X_2 from D_5. Data generated under scenario 5, e_i: N ~ (0, 20), using equation 5.1.

e_i: N ~ (0, 20)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.457	2.656	2.693	0.946	-1.754	0.856	0.849	0.951
Complete Records	1.745	1.746	1.772	0.941	-1.703	0.850	0.844	0.952
FCS	1.485	2.855	2.972	0.941	-1.753	0.861	0.854	0.949
FCSgroup	1.444	2.786	2.891	0.943	-1.751	0.858	0.848	0.948
FCSgroups	1.439	2.792	2.889	0.939	-1.749	0.858	0.851	0.948
FCS3group2	1.469	2.853	2.963	0.938	-1.751	0.858	0.851	0.947
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.396	0.428	0.441	0.934	1.430	3.253	3.323	0.950
Complete Records	1.332	1.693	1.728	0.920	0.000	0.000	0.000	0.000
FCS	0.396	0.476	0.486	0.937	1.455	3.506	3.684	0.933
FCSgroup	0.392	0.448	0.453	0.946	1.452	3.484	3.679	0.931
FCSgroups	0.388	0.448	0.454	0.946	1.458	3.493	3.690	0.934
FCS3group2	0.391	0.449	0.459	0.942	1.469	3.499	3.658	0.933
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.892	2.852	2.905	0.949	0.651	4.614	4.575	0.947
Complete Records	-1.845	1.847	1.856	0.919	-0.310	4.067	4.086	0.948
FCS	-0.869	3.043	3.179	0.938	0.678	4.745	4.764	0.943
FCSgroup	-0.877	2.974	3.086	0.935	0.663	4.692	4.683	0.942
FCSgroups	-0.872	2.979	3.090	0.937	0.672	4.695	4.681	0.937
FCS3group2	-0.849	3.043	3.168	0.938	0.677	4.741	4.764	0.937

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Scenario 6 (D=10, N: different per study, m=5, it=5)

Table D. 16. Simulation results when data missingness due to granularity problem is applied to X_3 from D_6, D_{10} and due to mixed type problem is applied to X_2 from D_3, D_9. Data generated under scenario 6, e_i: N ~ (0, 0.2), using equation 5.1.

e_i: N ~ (0, 0.2)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.488	0.011	0.011	0.958	-1.755	0.004	0.004	0.940
Complete Records	1.798	0.025	0.029	0.000	-1.692	0.012	0.011	0.000
FCS	1.487	0.013	0.013	0.953	-1.755	0.004	0.005	0.935
FCSgroup	1.488	0.013	0.012	0.950	-1.755	0.004	0.004	0.933
FCSgroups	1.488	0.013	0.013	0.947	-1.755	0.004	0.004	0.936
FCS3group2	1.487	0.013	0.013	0.946	-1.755	0.004	0.004	0.946
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.424	0.002	0.002	0.939	1.466	0.014	0.014	0.956
Complete Records	1.331	0.024	0.017	0.000	0.000	0.000	0.000	0.000
FCS	0.424	0.002	0.002	0.931	1.467	0.016	0.016	0.957
FCSgroup	0.424	0.002	0.002	0.934	1.466	0.016	0.016	0.945
FCSgroups	0.424	0.002	0.002	0.926	1.466	0.016	0.016	0.948
FCS3group2	0.424	0.002	0.002	0.928	1.467	0.016	0.016	0.939
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.956	0.012	0.011	0.957	0.498	0.020	0.020	0.947
Complete Records	-1.932	0.026	0.031	0.000	-0.478	0.057	0.052	0.000
FCS	-0.955	0.014	0.014	0.948	0.496	0.023	0.023	0.953
FCSgroup	-0.957	0.014	0.013	0.952	0.497	0.023	0.023	0.947
FCSgroups	-0.957	0.014	0.014	0.945	0.497	0.023	0.023	0.940
FCS3group2	-0.955	0.014	0.014	0.951	0.497	0.023	0.023	0.950

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table D. 17. Simulation results when data missingness due to granularity problem is applied to X_3 from D_6, D_{10} and due to mixed type problem is applied to X_2 from D_3, D_9. Data generated under scenario 6, e_i: N ~ (0, 2), using equation 5.1.

e_i: N ~ (0, 2)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.491	0.113	0.111	0.960	-1.756	0.037	0.036	0.947
Complete Records	1.800	0.078	0.078	0.021	-1.692	0.038	0.037	0.613
FCS	1.508	0.120	0.119	0.950	-1.755	0.038	0.037	0.954
FCSgroup	1.491	0.116	0.115	0.956	-1.756	0.037	0.037	0.949
FCSgroups	1.491	0.116	0.115	0.959	-1.756	0.037	0.037	0.946
FCS3group2	1.507	0.118	0.116	0.955	-1.756	0.037	0.037	0.947
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.423	0.018	0.019	0.943	1.462	0.139	0.136	0.952
Complete Records	1.329	0.076	0.074	0.000	0.000	0.000	0.000	0.000

FCS	0.423	0.021	0.023	0.918	1.464	0.147	0.147	0.946
FCSgroup	0.423	0.020	0.021	0.940	1.462	0.144	0.140	0.946
FCSgroups	0.423	0.020	0.021	0.937	1.461	0.144	0.141	0.948
FCS3group2	0.423	0.020	0.021	0.931	1.464	0.144	0.142	0.945
	β_4				**β_5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.961	0.122	0.120	0.948	0.493	0.196	0.190	0.955
Complete Records	-1.935	0.083	0.085	0.000	-0.479	0.182	0.174	0.000
FCS	-0.965	0.129	0.129	0.952	0.494	0.205	0.198	0.958
FCSgroup	-0.961	0.125	0.123	0.954	0.492	0.200	0.193	0.955
FCSgroups	-0.962	0.125	0.123	0.953	0.492	0.200	0.195	0.958
FCS3group2	-0.965	0.126	0.124	0.956	0.492	0.201	0.196	0.955

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table D. 18. Simulation results when data missingness due to granularity problem is applied to X_3 from D_6, D_{10} and due to mixed type problem is applied to X_2 from D_3, D_9. Data generated under scenario 6, e_i: N ~ (0, 20), using equation 5.1.

e_i: N ~ (0, 20)	**β_0**				**β_1**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.510	1.132	1.109	0.960	-1.759	0.366	0.364	0.947
Complete Records	1.811	0.747	0.736	0.928	-1.695	0.363	0.361	0.952
FCS	1.533	1.161	1.142	0.960	-1.761	0.369	0.370	0.948
FCSgroup	1.509	1.153	1.135	0.956	-1.761	0.367	0.365	0.945
FCSgroups	1.518	1.153	1.137	0.957	-1.760	0.367	0.365	0.949
FCS3group2	1.529	1.161	1.142	0.961	-1.761	0.367	0.365	0.947
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.420	0.183	0.190	0.943	1.432	1.387	1.362	0.952
Complete Records	1.306	0.725	0.726	0.777	0.000	0.000	0.000	0.000
FCS	0.424	0.215	0.236	0.924	1.425	1.423	1.395	0.944
FCSgroup	0.425	0.197	0.207	0.942	1.433	1.425	1.398	0.947
FCSgroups	0.424	0.196	0.206	0.940	1.420	1.426	1.400	0.942
FCS3group2	0.425	0.196	0.206	0.941	1.432	1.423	1.392	0.944
	β_4				**β_5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-1.003	1.216	1.195	0.948	0.447	1.963	1.904	0.955
Complete Records	-1.957	0.790	0.786	0.751	-0.504	1.732	1.684	0.908
FCS	-1.008	1.244	1.224	0.957	0.446	1.983	1.925	0.957
FCSgroup	-1.002	1.236	1.217	0.951	0.449	1.975	1.917	0.959
FCSgroups	-1.011	1.236	1.217	0.951	0.440	1.976	1.925	0.955
FCS3group2	-1.004	1.244	1.223	0.958	0.449	1.983	1.935	0.954

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Studies with different model errors

Scenario 7 (D=2, N=100 per study, m=10, it=10, e_1: N~(0,1.2), e_2: N~(0,1.3))

Table D. 19. Simulation results when data missingness due to granularity problem is applied to X_3 from D_2 and due to mixed type problem is applied to X_2 from D_1. Data generated under scenario 7, e_1: N~(0,1.2), e_2: N~(0,1.3)) using equation 5.1.

e_1: N~(0,1.2), e_2: N~(0,1.3)	**β_0**				**β_1**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.485	0.288	0.290	0.943	-1.753	0.091	0.093	0.945
Complete Records	1.802	0.208	0.220	0.654	-1.691	0.101	0.102	0.896
FCS	2.019	0.740	0.555	0.889	-1.748	0.114	0.106	0.963
FCSgroup	1.551	0.399	0.409	0.943	-1.749	0.101	0.100	0.950
FCSgroups	1.621	0.434	0.414	0.944	-1.750	0.101	0.099	0.954
FCS3group2	1.835	0.596	0.495	0.909	-1.752	0.105	0.104	0.955
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.424	0.046	0.046	0.944	1.471	0.352	0.342	0.953
Complete Records	1.335	0.202	0.185	0.005	0.000	0.000	0.000	0.000
FCS	0.442	0.065	0.067	0.915	1.250	0.884	0.614	0.984
FCSgroup	0.431	0.057	0.059	0.940	1.372	0.505	0.514	0.944
FCSgroups	0.430	0.058	0.058	0.944	1.273	0.566	0.521	0.959
FCS3group2	0.441	0.060	0.061	0.929	1.372	0.696	0.572	0.973
	β_4				**β_5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.961	0.309	0.315	0.934	0.495	0.504	0.505	0.951
Complete Records	-1.944	0.221	0.233	0.009	-0.484	0.495	0.484	0.481
FCS	-1.351	0.755	0.581	0.931	0.190	0.886	0.744	0.949
FCSgroup	-1.028	0.416	0.431	0.936	0.445	0.603	0.602	0.948
FCSgroups	-1.097	0.451	0.434	0.942	0.371	0.629	0.608	0.954
FCS3group2	-1.197	0.611	0.517	0.939	0.324	0.747	0.686	0.943

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Scenario 8 (D=2, N=200 per study, m=10, it=10, e_1: N~(0,1.2), e_2: N~(0,1.3))

Table D. 20. Simulation results when data missingness due to granularity problem is applied to X_3 from D_2 and due to mixed type problem is applied to X_2 from D_1. Data generated under scenario 8, e_1: N~(0,1.2), e_2: N~(0,1.3)) using equation 5.1.

e_1: N~(0,1.2), e_2: N~(0,1.3)	**β_0**				**β_1**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.504	0.200	0.202	0.953	-1.756	0.064	0.066	0.949
Complete Records	1.809	0.146	0.157	0.417	-1.693	0.071	0.071	0.863
FCS	2.252	0.584	0.398	0.725	-1.760	0.080	0.075	0.960
FCSgroup	1.549	0.273	0.285	0.928	-1.756	0.071	0.072	0.947
FCSgroups	1.612	0.325	0.301	0.949	-1.756	0.071	0.072	0.944
FCS3group2	1.880	0.461	0.349	0.877	-1.760	0.074	0.074	0.946
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.422	0.032	0.032	0.944	1.449	0.245	0.246	0.946
Complete Records	1.328	0.142	0.134	0.000	0.000	0.000	0.000	0.000
FCS	0.446	0.045	0.047	0.885	1.022	0.690	0.440	0.956

FCSgroup	0.429	0.040	0.041	0.939	1.382	0.348	0.354	0.942
FCSgroups	0.428	0.041	0.040	0.950	1.288	0.432	0.392	0.943
FCS3group2	0.439	0.042	0.042	0.926	1.337	0.526	0.385	0.981
	β_4				**β_5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.977	0.215	0.214	0.954	0.477	0.350	0.348	0.950
Complete Records	-1.945	0.155	0.161	0.000	-0.485	0.345	0.345	0.179
FCS	-1.573	0.594	0.411	0.833	-0.056	0.677	0.519	0.896
FCSgroup	-1.020	0.286	0.299	0.927	0.437	0.413	0.421	0.945
FCSgroups	-1.083	0.336	0.313	0.940	0.375	0.452	0.427	0.943
FCS3group2	-1.233	0.471	0.360	0.936	0.276	0.559	0.474	0.958

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Scenario 9 (D=2, N=500 per study, m=10, it=10, e_1: N~(0,1.2), e_2: N~(0,1.3))

Table D. 21. Simulation results when data missingness due to granularity problem is applied to X_3 from D_2 and due to mixed type problem is applied to X_2 from D_1. Data generated under scenario 9, e_1: N~(0,1.2), e_2: N~(0,1.3)) using equation 5.1.

e_1: N~(0,1.2), e_2: N~(0,1.3)	**β_0**				**β_1**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.489	0.126	0.125	0.955	-1.754	0.040	0.040	0.951
Complete Records	1.800	0.093	0.100	0.105	-1.690	0.045	0.044	0.701
FCS	2.462	0.374	0.282	0.247	-1.761	0.051	0.045	0.966
FCSgroup	1.528	0.171	0.172	0.943	-1.755	0.044	0.044	0.947
FCSgroups	1.568	0.214	0.189	0.949	-1.755	0.045	0.043	0.946
FCS3group2	1.902	0.293	0.225	0.716	-1.760	0.046	0.045	0.957
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.424	0.020	0.021	0.951	1.464	0.154	0.152	0.956
Complete Records	1.332	0.090	0.087	0.000	0.000	0.000	0.000	0.000
FCS	0.452	0.028	0.029	0.796	0.822	0.454	0.299	0.734
FCSgroup	0.432	0.025	0.026	0.921	1.402	0.219	0.213	0.939
FCSgroups	0.431	0.026	0.026	0.929	1.342	0.289	0.242	0.962
FCS3group2	0.442	0.026	0.027	0.868	1.323	0.333	0.243	0.970
	β_4				**β_5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.958	0.135	0.138	0.951	0.496	0.218	0.220	0.955
Complete Records	-1.935	0.098	0.107	0.000	-0.482	0.216	0.215	0.004
FCS	-1.783	0.379	0.294	0.399	-0.264	0.431	0.348	0.561
FCSgroup	-0.998	0.179	0.181	0.945	0.457	0.256	0.265	0.926
FCSgroups	-1.037	0.221	0.197	0.955	0.419	0.290	0.276	0.945
FCS3group2	-1.254	0.299	0.235	0.851	0.262	0.351	0.301	0.916

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Scenario 10 (D=2, N=1000 per study, m=10, it=10, e_1: N~(0,1.2), e_2: N~(0,1.3))

Table D. 22. Simulation results when data missingness due to granularity problem is applied to X_3 from D_2 and due to mixed type problem is applied to X_2 from D_1. Data generated under scenario 10, e_1: N~(0,1.2), e_2: N~(0,1.3)) using equation 5.1.

e_1: N~(0,1.2), e_2: N~(0,1.3)	**β_0**				**β_1**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.487	0.089	0.090	0.949	-1.756	0.029	0.029	0.957
Complete Records	1.801	0.066	0.067	0.002	-1.693	0.032	0.031	0.498
FCS	2.537	0.271	0.208	0.018	-1.767	0.036	0.033	0.959
FCSgroup	1.529	0.120	0.130	0.903	-1.759	0.031	0.031	0.951
FCSgroups	1.553	0.153	0.141	0.949	-1.759	0.032	0.031	0.952
FCS3group2	1.910	0.206	0.166	0.463	-1.764	0.033	0.032	0.952
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.424	0.014	0.014	0.959	1.468	0.109	0.107	0.949
Complete Records	1.331	0.064	0.057	0.000	0.000	0.000	0.000	0.000
FCS	0.454	0.019	0.020	0.646	0.745	0.330	0.231	0.400
FCSgroup	0.432	0.017	0.017	0.934	1.398	0.154	0.160	0.916
FCSgroups	0.431	0.019	0.017	0.948	1.362	0.209	0.182	0.945
FCS3group2	0.442	0.018	0.018	0.821	1.316	0.235	0.184	0.936
	β_4				**β_5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.957	0.095	0.097	0.942	0.502	0.154	0.154	0.946
Complete Records	-1.937	0.069	0.073	0.000	-0.477	0.152	0.146	0.000
FCS	-1.858	0.274	0.213	0.067	-0.337	0.309	0.254	0.205
FCSgroup	-0.999	0.126	0.136	0.911	0.461	0.180	0.187	0.940
FCSgroups	-1.022	0.158	0.147	0.946	0.436	0.205	0.194	0.949
FCS3group2	-1.260	0.210	0.171	0.714	0.261	0.248	0.217	0.862

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Scenario 1' (D=2, N=200 per study, ***m=5, it=5)***

Table D. 23. Simulation results when data missingness due to granularity problem is applied to X_3 from D_1 and due to mixed type problem is applied to X_2 from D_2. Data generated under scenario 1', e_i: N ~ (0, 0.2), using equation 5.1.

e_i: N ~ (0, 0.2)	**β_0**				**β_1**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.489	0.032	0.032	0.950	-1.755	0.010	0.011	0.943
Complete Records	1.804	0.069	0.083	0.020	-1.691	0.034	0.030	0.535
FCS	2.211	0.465	0.318	0.718	-1.748	0.026	0.021	0.961
FCSgroup	1.526	0.070	0.057	0.946	-1.753	0.019	0.017	0.958
FCSgroups	1.533	0.077	0.064	0.950	-1.752	0.019	0.017	0.959
FCS3group2	1.683	0.256	0.212	0.969	-1.751	0.021	0.018	0.970
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.423	0.005	0.005	0.953	1.464	0.039	0.039	0.951
Complete Records	1.330	0.067	0.048	0.000	0.000	0.000	0.000	0.000
FCS	0.426	0.013	0.010	0.982	0.742	0.479	0.323	0.727
FCSgroup	0.427	0.009	0.008	0.953	1.403	0.077	0.064	0.891
FCSgroups	0.426	0.010	0.008	0.961	1.390	0.089	0.076	0.894

FCS3group2	0.427	0.011	0.009	0.955	1.276	0.260	0.211	0.964
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.957	0.034	0.034	0.953	0.497	0.056	0.056	0.951
Complete Records	-1.937	0.073	0.086	0.000	-0.481	0.164	0.150	0.000
FCS	-1.625	0.454	0.315	0.733	-0.271	0.492	0.328	0.701
FCSgroup	-0.992	0.073	0.059	0.955	0.465	0.105	0.090	0.952
FCSgroups	-1.000	0.080	0.066	0.959	0.457	0.111	0.094	0.954
FCS3group2	-1.131	0.252	0.210	0.974	0.181	0.299	0.237	0.867

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table D. 24. Simulation results when data missingness due to granularity problem is applied to X_3 from D_1 and due to mixed type problem is applied to X_2 from D_2. Data generated under scenario 1', e_i: N ~ (0, 2), using equation 5.1.

e_i: N ~ (0, 2)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.485	0.320	0.322	0.948	-1.753	0.103	0.102	0.953
Complete Records	1.801	0.220	0.226	0.701	-1.690	0.107	0.108	0.902
FCS	1.948	0.811	0.611	0.933	-1.747	0.117	0.114	0.954
FCSgroup	1.468	0.409	0.447	0.915	-1.746	0.108	0.109	0.950
FCSgroups	1.524	0.460	0.455	0.932	-1.747	0.109	0.108	0.951
FCS3group2	1.744	0.604	0.530	0.928	-1.749	0.112	0.111	0.948
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.422	0.051	0.052	0.944	1.471	0.391	0.385	0.951
Complete Records	1.329	0.213	0.207	0.006	0.000	0.000	0.000	0.000
FCS	0.426	0.078	0.081	0.924	1.393	1.001	0.694	0.984
FCSgroup	0.417	0.063	0.064	0.931	1.492	0.528	0.578	0.922
FCSgroups	0.418	0.065	0.063	0.943	1.408	0.611	0.592	0.931
FCS3group2	0.421	0.066	0.065	0.945	1.508	0.721	0.615	0.973
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.956	0.343	0.348	0.948	0.499	0.558	0.546	0.960
Complete Records	-1.936	0.233	0.243	0.013	-0.477	0.516	0.490	0.516
FCS	-1.235	0.828	0.633	0.956	0.268	0.964	0.793	0.966
FCSgroup	-0.939	0.430	0.464	0.914	0.515	0.628	0.644	0.943
FCSgroups	-0.996	0.479	0.470	0.936	0.460	0.664	0.653	0.956
FCS3group2	-1.049	0.622	0.549	0.956	0.452	0.780	0.729	0.961

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table D. 25. Simulation results when data missingness due to granularity problem is applied to X_3 from D_1 and due to mixed type problem is applied to X_2 from D_2. Data generated under scenario 1', e_i: N ~ (0, 20), using equation 5.1.

e_i: N ~ (0, 20)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.543	3.184	3.291	0.946	-1.757	1.026	1.006	0.950

Complete Records	1.858	2.092	2.158	0.935	-1.697	1.018	0.999	0.947
FCS	1.772	4.815	4.699	0.952	-1.738	1.053	1.027	0.954
FCSgroup	1.612	3.773	4.395	0.901	-1.746	1.039	1.012	0.957
FCSgroups	1.685	3.878	4.291	0.918	-1.745	1.039	1.010	0.956
FCS3group2	1.731	4.561	4.585	0.957	-1.741	1.036	1.009	0.955
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.428	0.513	0.517	0.945	1.491	3.903	3.852	0.951
Complete Records	1.358	2.033	2.053	0.925	0.000	0.000	0.000	0.000
FCS	0.399	0.770	0.754	0.941	1.425	6.031	5.701	0.961
FCSgroup	0.417	0.616	0.597	0.953	1.388	4.925	5.796	0.893
FCSgroups	0.414	0.623	0.607	0.951	1.281	5.096	5.634	0.908
FCS3group2	0.405	0.613	0.596	0.946	1.411	5.634	5.587	0.952
	β_4				**β_5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.939	3.420	3.534	0.942	0.229	5.596	5.606	0.947
Complete Records	-1.933	2.215	2.278	0.921	-0.770	4.944	4.986	0.946
FCS	-0.927	5.005	4.918	0.956	0.253	6.711	6.566	0.953
FCSgroup	-1.006	3.977	4.610	0.896	0.147	5.964	6.361	0.932
FCSgroups	-1.082	4.079	4.515	0.918	0.083	6.038	6.273	0.939
FCS3group2	-0.890	4.752	4.787	0.951	0.322	6.495	6.476	0.951

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Scenario 2' (D=2, N=1000 per study, ***m=5, it=5****)*

Table D. 26. Simulation results when data missingness due to granularity problem is applied to X_3 from D_1 and due to mixed type problem is applied to X_2 from D_2. Data generated under scenario 2', e_i: N ~ (0, 0.2), using equation 5.1.

e_i:N~(0,0.2)	**β_0**				**β_1**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.488	0.014	0.014	0.950	-1.755	0.005	0.005	0.944
Complete Records	1.798	0.031	0.037	0.000	-1.692	0.015	0.013	0.006
FCS	2.570	0.368	0.228	0.414	-1.751	0.012	0.010	0.945
FCSgroup	1.496	0.023	0.021	0.946	-1.755	0.007	0.007	0.940
FCSgroups	1.496	0.023	0.021	0.939	-1.755	0.007	0.007	0.947
FCS3group2	1.524	0.115	0.106	0.972	-1.755	0.008	0.007	0.942
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.424	0.002	0.002	0.955	1.466	0.017	0.018	0.951
Complete Records	1.331	0.030	0.022	0.000	0.000	0.000	0.000	0.000
FCS	0.424	0.006	0.005	0.959	0.391	0.374	0.230	0.418
FCSgroup	0.425	0.003	0.003	0.926	1.452	0.026	0.026	0.895
FCSgroups	0.425	0.003	0.003	0.917	1.451	0.027	0.026	0.911
FCS3group2	0.425	0.004	0.004	0.951	1.436	0.117	0.105	0.952
	β_4				**β_5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.957	0.015	0.015	0.955	0.498	0.025	0.023	0.957
Complete Records	-1.933	0.033	0.038	0.000	-0.477	0.072	0.066	0.000

FCS	-1.979	0.361	0.225	0.428	-0.704	0.385	0.241	0.389
FCSgroup	-0.964	0.024	0.023	0.934	0.492	0.037	0.036	0.943
FCSgroups	-0.964	0.025	0.022	0.941	0.491	0.037	0.036	0.943
FCS3group2	-0.983	0.113	0.104	0.953	0.369	0.151	0.128	0.833

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table D. 27. Simulation results when data missingness due to granularity problem is applied to X_3 from D_1 and due to mixed type problem is applied to X_2 from D_2. Data generated under scenario 2', e_i: N ~ (0, 2), using equation 5.1.

e_i:N~(0,2)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.494	0.142	0.138	0.953	-1.757	0.046	0.046	0.946
Complete Records	1.806	0.098	0.103	0.106	-1.694	0.048	0.048	0.742
FCS	2.155	0.473	0.308	0.689	-1.760	0.053	0.049	0.962
FCSgroup	1.493	0.180	0.191	0.918	-1.756	0.048	0.049	0.945
FCSgroups	1.533	0.216	0.192	0.960	-1.755	0.048	0.049	0.949
FCS3group2	1.804	0.273	0.232	0.796	-1.758	0.049	0.049	0.953
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.423	0.023	0.023	0.947	1.466	0.174	0.170	0.951
Complete Records	1.327	0.095	0.090	0.000	0.000	0.000	0.000	0.000
FCS	0.437	0.035	0.035	0.919	1.242	0.575	0.335	0.971
FCSgroup	0.422	0.028	0.028	0.941	1.466	0.235	0.251	0.924
FCSgroups	0.420	0.029	0.027	0.949	1.405	0.295	0.256	0.960
FCS3group2	0.427	0.029	0.029	0.942	1.474	0.326	0.276	0.963
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.962	0.152	0.149	0.955	0.495	0.246	0.242	0.956
Complete Records	-1.939	0.104	0.110	0.000	-0.480	0.227	0.227	0.009
FCS	-1.440	0.481	0.319	0.839	0.064	0.529	0.381	0.872
FCSgroup	-0.961	0.189	0.198	0.935	0.499	0.276	0.283	0.949
FCSgroups	-1.002	0.224	0.200	0.948	0.459	0.302	0.282	0.952
FCS3group2	-1.104	0.280	0.241	0.941	0.409	0.348	0.319	0.949

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Table D. 28. Simulation results when data missingness due to granularity problem is applied to X_3 from D_1 and due to mixed type problem is applied to X_2 from D_2. Data generated under scenario 2', e_i: N ~ (0, 20), using equation 5.1.

e_i:N~(0,20)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.541	1.418	1.377	0.953	-1.770	0.457	0.461	0.946
Complete Records	1.874	0.935	0.954	0.915	-1.705	0.454	0.460	0.946
FCS	1.787	2.154	2.065	0.953	-1.773	0.469	0.466	0.953
FCSgroup	1.571	1.679	1.883	0.911	-1.768	0.461	0.466	0.948
FCSgroups	1.572	1.695	1.872	0.914	-1.770	0.462	0.466	0.950
FCS3group2	1.786	2.019	1.986	0.940	-1.766	0.462	0.465	0.951

	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.420	0.228	0.229	0.947	1.467	1.736	1.704	0.951
Complete Records	1.285	0.907	0.883	0.839	0.000	0.000	0.000	0.000
FCS	0.436	0.338	0.339	0.938	1.410	2.670	2.608	0.953
FCSgroup	0.417	0.273	0.277	0.939	1.425	2.191	2.564	0.890
FCSgroups	0.423	0.276	0.272	0.953	1.435	2.330	2.586	0.906
FCS3group2	0.414	0.274	0.267	0.948	1.403	2.488	2.519	0.942
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-1.015	1.523	1.489	0.955	0.467	2.457	2.423	0.956
Complete Records	-1.990	0.988	1.017	0.803	-0.506	2.168	2.194	0.920
FCS	-1.020	2.237	2.133	0.952	0.461	2.955	2.910	0.957
FCSgroup	-1.044	1.770	1.949	0.916	0.441	2.621	2.775	0.936
FCSgroups	-1.038	1.846	1.949	0.936	0.445	2.677	2.773	0.938
FCS3group2	-1.023	2.105	2.053	0.950	0.471	2.843	2.873	0.946

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Scenario 7' (D=2, N=100 per study, ***m=5, it=5****,* e_1*: N~(0,1.2),* e_2*: N~(0,1.3))*

Table D. 29. Simulation results when data missingness due to granularity problem is applied to X_3 from D_2 and due to mixed type problem is applied to X_2 from D_1. Data generated under scenario 7', e_1: N~(0,1.2), e_2: N~(0,1.3)) using equation 5.1.

e_1: N~(0,1.2), e_2: N~(0,1.3)	β_0				β_1			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.500	0.287	0.291	0.947	-1.752	0.091	0.093	0.950
Complete Records	1.805	0.208	0.215	0.651	-1.692	0.101	0.102	0.900
FCS	2.025	0.749	0.559	0.869	-1.741	0.115	0.108	0.960
FCSgroup	1.574	0.401	0.411	0.920	-1.748	0.102	0.102	0.944
FCSgroups	1.644	0.441	0.424	0.924	-1.748	0.102	0.103	0.941
FCS3group2	1.831	0.594	0.502	0.907	-1.750	0.106	0.105	0.948
	β_2				β_3			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.423	0.046	0.045	0.954	1.451	0.351	0.357	0.946
Complete Records	1.336	0.202	0.192	0.002	0.000	0.000	0.000	0.000
FCS	0.433	0.067	0.069	0.909	1.193	0.908	0.634	0.968
FCSgroup	0.428	0.058	0.060	0.924	1.342	0.511	0.516	0.938
FCSgroups	0.427	0.060	0.060	0.936	1.236	0.579	0.537	0.937
FCS3group2	0.436	0.061	0.063	0.924	1.356	0.696	0.576	0.968
	β_4				β_5			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.969	0.308	0.312	0.954	0.503	0.505	0.516	0.940
Complete Records	-1.941	0.220	0.232	0.011	-0.485	0.496	0.498	0.489
FCS	-1.342	0.766	0.576	0.906	0.171	0.902	0.746	0.938
FCSgroup	-1.040	0.420	0.426	0.936	0.429	0.611	0.620	0.943
FCSgroups	-1.111	0.458	0.438	0.928	0.355	0.639	0.634	0.938
FCS3group2	-1.180	0.608	0.515	0.936	0.329	0.753	0.691	0.944

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Scenario 8' (D=2, N=200 per study, ***m=5, it=5****,* e_1*: N~(0,1.2),* e_2*: N~(0,1.3))*

Table D. 30. Simulation results when data missingness due to granularity problem is applied to X_3 from D_2 and due to mixed type problem is applied to X_2 from D_1. Data generated under scenario 8', e_1: N~(0,1.2), e_2: N~(0,1.3)) using equation 5.1.

e_1: N~(0,1.2), e_2: N~(0,1.3)	**β_0**				**β_1**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.487	0.201	0.211	0.943	-1.751	0.064	0.063	0.947
Complete Records	1.805	0.147	0.155	0.454	-1.688	0.071	0.069	0.846
FCS	2.223	0.608	0.424	0.726	-1.751	0.081	0.073	0.971
FCSgroup	1.536	0.280	0.296	0.925	-1.752	0.071	0.069	0.948
FCSgroups	1.620	0.340	0.314	0.937	-1.752	0.071	0.070	0.952
FCS3group2	1.865	0.468	0.392	0.856	-1.755	0.074	0.071	0.963
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.423	0.032	0.031	0.957	1.469	0.246	0.248	0.949
Complete Records	1.331	0.142	0.132	0.000	0.000	0.000	0.000	0.000
FCS	0.444	0.046	0.048	0.891	1.037	0.725	0.481	0.939
FCSgroup	0.432	0.040	0.041	0.935	1.393	0.356	0.366	0.926
FCSgroups	0.429	0.041	0.041	0.939	1.267	0.455	0.399	0.943
FCS3group2	0.441	0.042	0.043	0.915	1.346	0.535	0.435	0.958
	β_4				**β_5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.956	0.216	0.224	0.945	0.491	0.350	0.361	0.944
Complete Records	-1.939	0.156	0.167	0.000	-0.494	0.345	0.347	0.188
FCS	-1.541	0.619	0.436	0.818	-0.030	0.706	0.552	0.881
FCSgroup	-1.004	0.293	0.306	0.929	0.443	0.420	0.436	0.939
FCSgroups	-1.089	0.351	0.323	0.936	0.359	0.464	0.458	0.940
FCS3group2	-1.216	0.478	0.402	0.912	0.290	0.567	0.509	0.935

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Scenario 9' (D=2, N=500 per study, ***m=5, it=5****,* e_1*: N~(0,1.2),* e_2*: N~(0,1.3))*

Table D. 31. Simulation results when data missingness due to granularity problem is applied to X_3 from D_2 and due to mixed type problem is applied to X_2 from D_1. Data generated under scenario 9', e_1: N~(0,1.2), e_2: N~(0,1.3)) using equation 5.1.

e_1: N~(0,1.2), e_2: N~(0,1.3)	**β_0**				**β_1**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.490	0.126	0.131	0.938	-1.754	0.040	0.041	0.946
Complete Records	1.801	0.093	0.100	0.098	-1.692	0.045	0.044	0.716
FCS	2.363	0.412	0.283	0.379	-1.761	0.052	0.047	0.965
FCSgroup	1.533	0.175	0.189	0.908	-1.756	0.045	0.045	0.949
FCSgroups	1.598	0.222	0.199	0.926	-1.755	0.045	0.045	0.949

FCS3group2	1.895	0.304	0.243	0.718	-1.761	0.047	0.047	0.952
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.423	0.020	0.020	0.939	1.463	0.154	0.162	0.938
Complete Records	1.328	0.090	0.085	0.000	0.000	0.000	0.000	0.000
FCS	0.448	0.028	0.030	0.822	0.919	0.496	0.322	0.805
FCSgroup	0.430	0.025	0.026	0.933	1.393	0.225	0.239	0.915
FCSgroups	0.427	0.027	0.026	0.945	1.296	0.303	0.256	0.931
FCS3group2	0.440	0.026	0.028	0.893	1.327	0.346	0.273	0.950
	β_4				**β_5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.959	0.135	0.139	0.937	0.499	0.218	0.211	0.960
Complete Records	-1.933	0.098	0.103	0.000	-0.477	0.216	0.206	0.004
FCS	-1.682	0.418	0.292	0.528	-0.156	0.469	0.349	0.680
FCSgroup	-1.002	0.183	0.197	0.916	0.457	0.259	0.265	0.934
FCSgroups	-1.067	0.228	0.205	0.930	0.390	0.296	0.270	0.946
FCS3group2	-1.245	0.311	0.251	0.842	0.275	0.363	0.310	0.908

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

Scenario 10' (D=2, N=1000 per study, ***m=5, it=5****,* e_1*: N~(0,1.2),* e_2*: N~(0,1.3))*

Table D. 32. Simulation results when data missingness due to granularity problem is applied to X_3 from D_2 and due to mixed type problem is applied to X_2 from D_1. Data generated under scenario 10', e_1: N~(0,1.2), e_2: N~(0,1.3)) using equation 5.1.

e_1: N~(0,1.2), e_2: N~(0,1.3)	**β_0**				**β_1**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	1.493	0.089	0.086	0.955	-1.755	0.029	0.029	0.951
Complete Records	1.803	0.065	0.069	0.003	-1.692	0.032	0.032	0.485
FCS	2.481	0.294	0.213	0.091	-1.764	0.036	0.034	0.956
FCSgroup	1.542	0.122	0.126	0.911	-1.758	0.032	0.032	0.951
FCSgroups	1.596	0.158	0.131	0.910	-1.757	0.032	0.031	0.945
FCS3group2	1.926	0.209	0.170	0.420	-1.763	0.033	0.033	0.952
	β_2				**β_3**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	0.423	0.014	0.014	0.941	1.463	0.109	0.107	0.955
Complete Records	1.331	0.063	0.061	0.000	0.000	0.000	0.000	0.000
FCS	0.451	0.020	0.021	0.684	0.799	0.358	0.235	0.502
FCSgroup	0.430	0.018	0.018	0.925	1.387	0.158	0.160	0.921
FCSgroups	0.427	0.019	0.018	0.958	1.306	0.217	0.172	0.917
FCS3group2	0.441	0.018	0.020	0.819	1.304	0.238	0.188	0.921
	β_4				**β_5**			
	Mean	mSE	EmpSE	Cov	Mean	mSE	EmpSE	Cov
Full Data	-0.963	0.095	0.093	0.953	0.492	0.154	0.153	0.955
Complete Records	-1.939	0.069	0.072	0.000	-0.482	0.152	0.150	0.000
FCS	-1.802	0.298	0.217	0.189	-0.283	0.334	0.261	0.345
FCSgroup	-1.012	0.128	0.132	0.907	0.441	0.184	0.185	0.933
FCSgroups	-1.066	0.163	0.138	0.901	0.387	0.210	0.189	0.933
FCS3group2	-1.277	0.214	0.175	0.653	0.235	0.251	0.222	0.815

Mean: mean estimate over imputed data sets; *mSE:* standard error derived from Rubin's rules (mean over imputed data sets); *EmpSE:* standard deviation of estimate over imputed data sets; *Cov:* coverage of nominal 95% confidence interval.

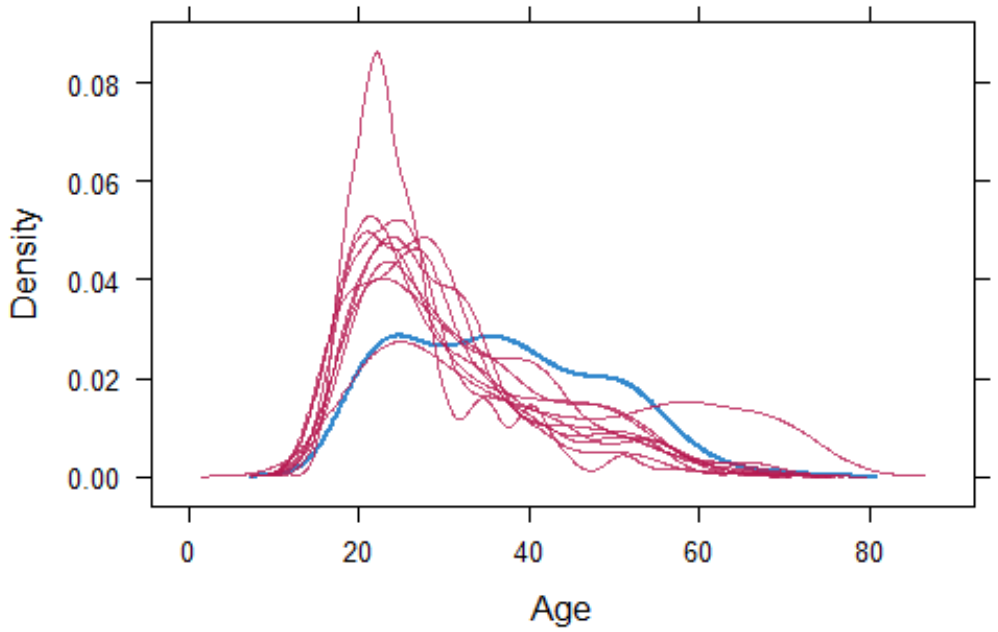


Figure D. 1. Density plot for variables: 'Age' in combined content heterogeneity problems. Blue line shows the observed data and the magenta lines the imputed data from each of the imputations in FCS.

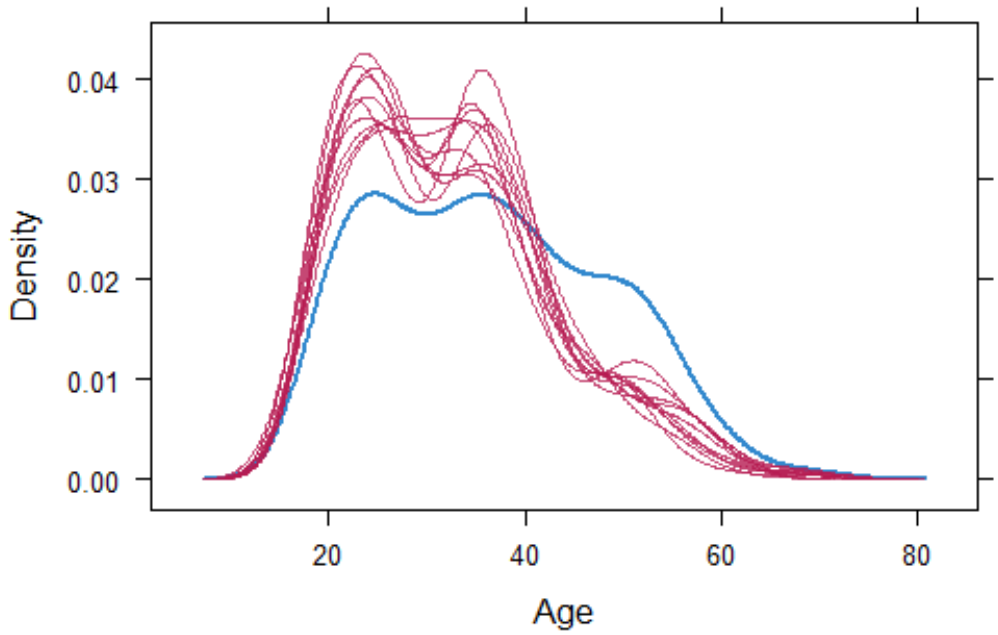


Figure D. 2. Density plot for variables: 'Age' in combined content heterogeneity problems. Blue line shows the observed data and the magenta lines the imputed data from each of the imputations in FCSgroup.

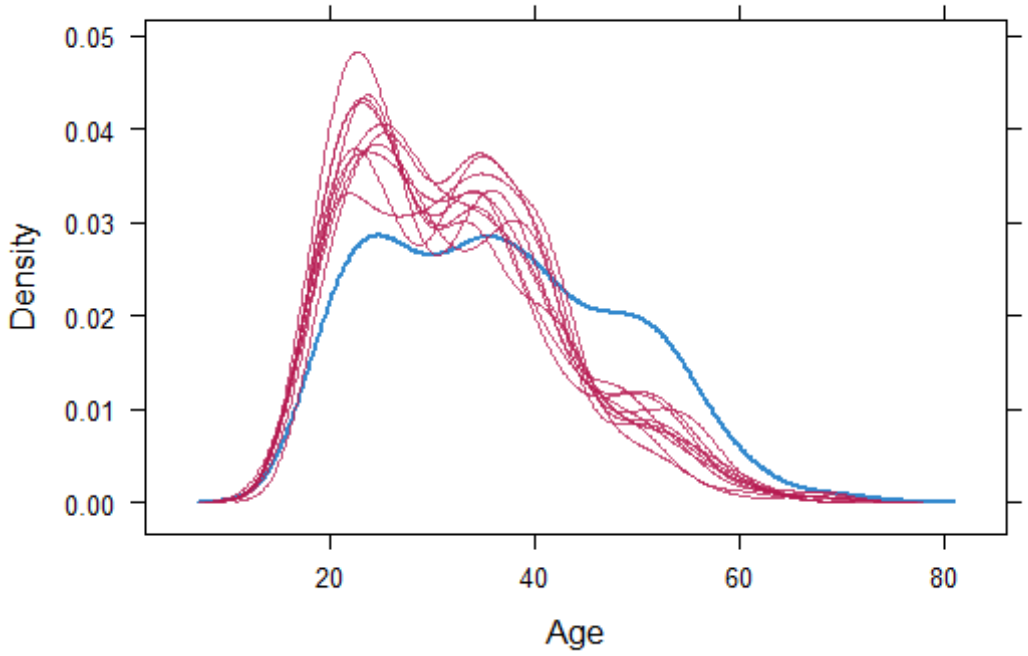


Figure D. 3. Density plot for variables: 'Age' in combined content heterogeneity problems. Blue line shows the observed data and the magenta lines the imputed data from each of the imputations in FCSgroups.

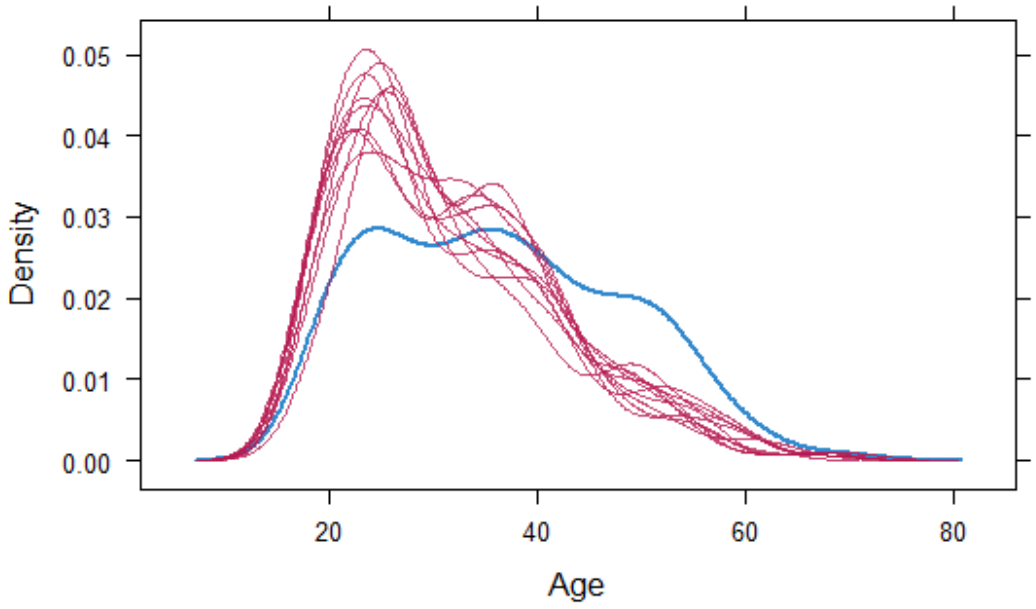


Figure D. 4. Density plot for variables: 'Age' in combined content heterogeneity problems. Blue line shows the observed data and the magenta lines the imputed data from each of the imputations in FCS3group2.

Appendix E: Publication

This section includes a publication resulted from this PhD work.

Digital Personalized Health and Medicine
L.B. Pape-Haugaard et al. (Eds.)

doi:10.3233/SHTI200188

Probabilistic Approaches to Overcome Content Heterogeneity in Data Integration: A Study Case in Systematic Lupus Erythematosus

Alexia SAMPRI[a], Nophar GEIFMAN [a], Helen LE SUEUR [a], Patrick DOHERTY[b], Philip COUCH [a], Ian BRUCE[b,c] and Niels PEEK[a,c] ,
on behalf of the MASTERplans Consortium
[a]*Division of Informatics, Imaging and Data Sciences University of Manchester, Manchester, UK*
[b] *Division of Musculoskeletal & Dermatological Sciences, University of Manchester, Manchester, UK*
[c] *NIHR Manchester Biomedical Research Centre, University of Manchester, Manchester Academic Health Science Centre, Manchester, UK*

Abstract. Integrating data from different sources into homogeneous dataset increases the opportunities to study human health. However, disparate data collections are often heterogeneous, which complicates their integration. In this paper, we focus on the issue of content heterogeneity in data integration. Traditional approaches for resolving content heterogeneity map all source datasets to a common data model that includes only shared data items, and thus omit all items that vary between datasets. Based on an example of three datasets in Systemic Lupus Erythematosus, we describe and experimentally evaluate a probabilistic data integration approach which propagates the uncertainty resulting from content heterogeneity into statistical inference, avoiding the need to map to a common data model.

Keywords. Probabilistic data integration, content heterogeneity, missing data, biomedical data harmonisation

1. Introduction

Integrating data from different sources into a homogeneous data resource creates powerful opportunities to study human health. However, disparate data collections are often heterogeneous, which complicates their integration.

Systemic Lupus Erythematosus (SLE) is a chronic autoimmune disease of unknown aetiology, affecting approximately 16,000 people in the UK. Through the MASTERplans programme [1] we have gained access to data from three cohort studies: Aspreva Lupus Management (ALMS), Lunar, and Exploratory Phase II/III SLE Evaluation of Rituximab (Explorer).

Content heterogeneity refers to the situation in which data items that are not equally represented across different data sources [2]. In the MASTERplans example, content heterogeneity occurs because Lunar and Explorer did not record age at onset of disease,

smoking status, white blood cells (WBC), platelets and lymphocytes, creating missing variables problems. Another form of content heterogeneity occurs because ethnicity was not recorded using the same categories across the three studies.

Traditionally, people solve these problems by mapping all datasets to a common data model. This model would only include variables that are present in all datasets, and, in case of granularity problems, categories that exist in all datasets.

2. Data and Method

First, we select all the datasets that are available and can enable us to answer the research question based on the description in the metadata. For the datasets that do not satisfy all the selection criteria concerning content heterogeneity we build predictive models that estimate the probability that these criteria are present based on other information that is available, using more complete datasets. The probabilistic relationship selection criteria are learned from other datasets. Our approach aims to preserve all the available information in the original datasets by translating content heterogeneity problems into missing value problems, and then solving these missing value problems using established methods (multiple imputation). The method assumes that naming differences are resolved beforehand, that the original datasets were sampled from the same population and that data are missing at random.

We applied our method to the three SLE datasets, to answer the research questions "does ethnicity predict response to treatment?" and "Finding the set of variables that best predict drug response". For the first question a granularity problem occurred, and for the second question a missing variables problem occurred. For both solutions, we applied imputation using multivariate imputation by chained equations (MICE) [3] and a random forest approach called missForest [4].

In order to answer both questions, we included only ALMS-M (maintenance) information from the ALMS study data since we required patients to have a 12-month follow-up visit (with disease severity evaluated) which was not available in the ALMS-I set (which followed patients for only 6 months). We rescaled the first visit as zero days and calculated days for each following visit relative to that. For both questions were used datasets that included patients from all the three studies together, with multiple visits per subject. For the response measure only, in order to have only one visit per patient, we kept the visit that had the least absolute difference from 365 days (12 month). First question's dataset consists of the 545 patients and 9 variables i.e., gender, age, ethnicity, height, weight, drug response (BILAGScore total), creatine, body mass index (BMI), and current treatment. Content heterogeneity occurs because ethnicities' levels are not the same across all the sources. Ethnicity in ALMS has 4 levels, i.e., 'Caucasian', 'Asian', 'Black', and 'Other', where 'Other' are also different levels of ethnicity. On the contrary, ethnicity's levels in Lunar and Explorer are 'Caucasian' and 'Black or African American'. Therefore, in ALMS, ethnicity's granularity is very high. For the granularity problem we add a categorical variable that classifies patients based on the 'Other' ethnicity. With this additional variable we eliminate misclassification.

The dataset with the missing variables problem contains not only the initial variables mentioned earlier but also smoking status, lymphocytes, WBC, and platelets. Missing variable problem occurs because these extra variables were included only in Lunar and Explorer data. We remove records for 7 patients with missing value for height and 9 with

missing creatine leading to 530 patients. This way makes it better and clearer to understand how the method works in application.

The chosen parameters for missForest imputation are 100 trees and a maximum of 50 iterations for both research questions. For MICE the chosen parameters are 30 and 20 imputed datasets, 50 and 40 iterations, and 10,000 and 15,000 seed for the research questions respectively. At the end, we fit linear regression models to the complete datasets that resulted from both imputation methods.

3. Results

In Table 1, we see the coefficients for the linear regression model that estimates the drug response based on ethnicity. We compare the coefficients from both imputation methods; MICE and random forest. Table 2 shows the coefficients for the linear regression model that estimates the drug response based on ethnicity, gender, smoking status, BMI, treatment, creatine, lymphocytes and platelets. Figure 1. shows the distributions

Table 1. Coefficients for linear regression model in granularity problem

	estimate		std error		p_{value}	
	MICE	**MissForest**	**MICE**	**MissForest**	**MICE**	**MissForest**
Intercept	1.1917	1.0000	4.9643	6.1870	0.8104	0.8720
Asian	3.8140	4.6710	5.0076	6.2090	0.4467	0.4520
Black or African American	5.8783	6.0200	5.0009	6.2180	0.2404	0.3330
Cape Coloured	-0.9833	-1.0000	7.5161	8.7500	0.8960	0.9090
Caucasian	6.9757	7.2210	4.9789	6.1990	0.1618	0.2450
East Indian	10.3623	5.0000	5.4224	8.7500	0.0567	0.5680
Eritrean	12.4797	11.0000	5.7933	8.7500	0.0319	0.2090
Hispanic	4.1158	4.5000	6.2853	7.5780	0.5129	0.5530
Mexican Mestizo	2.8448	3.2610	5.0588	6.3200	0.5741	0.6060
Middle Eastern	-0.8167	-1.0000	7.5825	8.7500	0.9143	0.9090
Mixed	2.7500	1.0000	6.5575	8.7500	0.6751	0.9090
Moroccan	5.9950	5.0000	7.4419	8.7500	0.4210	0.5680
Native American	4.0167	2.0000	6.9421	8.7500	0.5632	0.8190
Nicaraguan	9.7708	6.0000	6.7114	7.1440	0.1462	0.4010

Table 2. Coefficients for linear regression model in missing variables problem

	estimate		std error		p_{value}	
	MICE	**MissForest**	**MICE**	**MissForest**	**MICE**	**MissForest**
(Intercept)	5.4982	6.2212	1.89E+00	1.9747	5.76E-03	0.0017
Caucasian	1.2472	1.2619	7.47E-01	0.7485	1.22E-01	0.0924
Other	-0.5385	-0.5426	8.00E-01	0.7867	4.37E-01	0.4907
Male	-2.0917	-1.9727	8.64E-01	0.8238	1.89E-02	0.017
Never smoker	-0.4917	-0.2473	8.73E-01	0.8431	5.57E-01	0.7694
Previous smoker	-1.0473	-1.0190	1.02E+00	0.9793	3.10E-01	0.2986
BMI	0.0402	0.0465	4.33E-02	0.0421	4.17E-01	0.2692
Placebo + MMF OR MMF	-0.2093	0.1268	7.95E-01	0.6919	8.74E-01	0.8547
Placebo + MTX OR MTX	3.4207	3.8444	1.54E+00	1.5080	2.73E-02	0.0111
RITUX + AZA	1.4513	1.5648	1.18E+00	1.1227	2.00E-01	0.1634
RITUX + MMF	0.6037	0.8455	9.02E-01	0.7961	5.08E-01	0.2888
RITUX + MTX	2.0002	2.2410	1.24E+00	1.1906	1.20E-01	0.0604
Creatine	1.0438	1.0278	9.24E-01	0.8998	3.23E-01	0.2539
Lymphocytes	-2.3277	-2.9186	5.60E-01	0.5525	2.40E-05	1.89E-07
Platelets	0.0090	0.0068	3.40E-03	0.0036	1.11E-02	0.05723

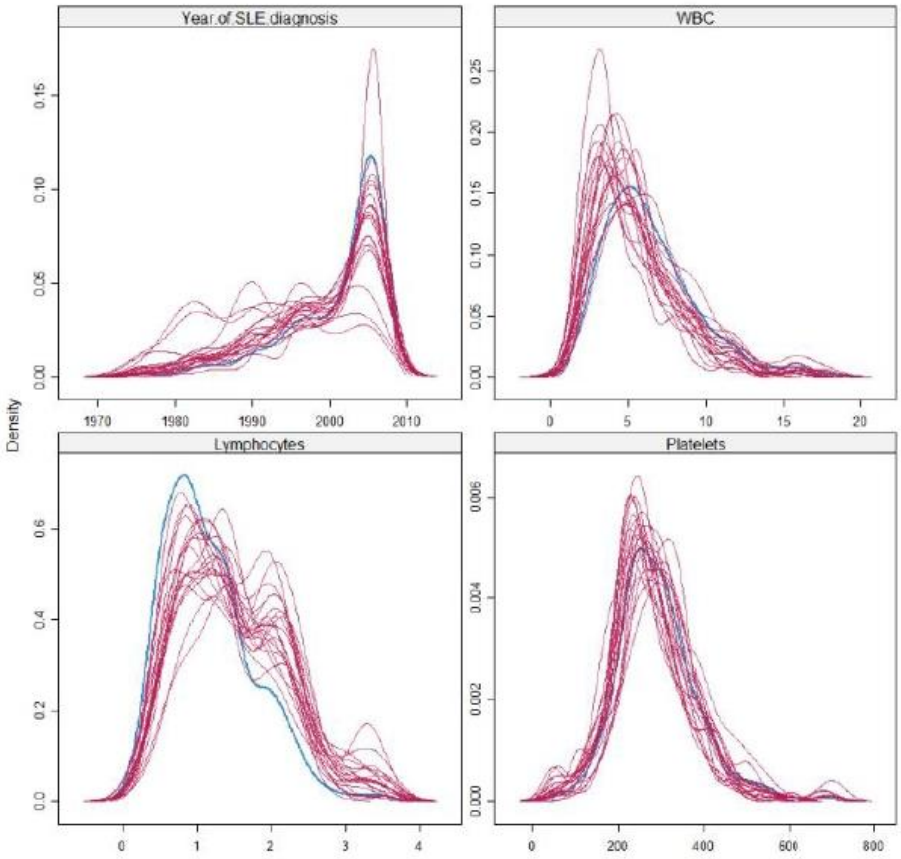


Figure 1. Density plots for year of SLE diagnosis, WBC, lymphocytes and platelets, in missing variables problem. Blue line shows the observed data and the red lines the imputed data from each of the imputations in MICE.

4. Discussion

The impact of ethnicity on drug relationship has the huge potential of achieving and improving the efficacy of results in precision medicine and differences in recommended drug doses [5, 6]. In table 1 we see that 'Eritrean' is significant and it implies that ethnicity is associated with drug response. Concerning the imputation methods, MICE did not have any misclassifications. However, missForest misclassified 1 patient as 'Caucasian' and 2 as 'Black of African American'. Further analysis is suggested such as application of our method to other data sources.

In Table 2, the coefficients of linear regression models, after imputation using MICE, show that all the variables included in the model are significant. Nonetheless, imputed values with missForest imputation show less significance. Gender, treatment and lymphocytes are the most significant with missForest imputation. Figure 1 also shows the distributions of the observed and imputed values for missing variables problem. As we see in figure 1 extreme values affect the shape of the plots. However, the central tendencies of the density plots of imputed data appear relatively similar to the observed ones.

In general, our approach contains a number of important ingredients. First, it insists on the future existence of health data heterogeneity. Therefore, our approach strives for

post alignment rather than pre-alignment of Big-health/bio datasets. Second, as a post-alignment of heterogeneous data sources will be always imperfect and it is not a problem that datasets are not content equivalent, if they estimate the probability that they are. Third, this approach is pragmatic in the sense that always provides an answer- although it might not better when the source datasets do not provide useful information to answer the research question. However, the results of the probabilistic data integration would be the same as those that would result from analysing an integrated dataset.

5. Conclusion

Existing methods for dataset integration rely on mapping to common data models, often resulting in a substantial loss of information that is present in the source datasets. One promising alternative relies on probabilistic methodologies. This paper has illustrated this approach using a real-world example from Lupus cohort studies. Rather than relying on perfectly harmonised data items, our method propagates the uncertainty that results from imperfect harmonisation into the statistical analysis, thus obviating the need for data integration through a common data model.

Ideally, shared data models would be implemented at source, enabling uniform data collection at different sites and studies. But in reality, data standardisation is always imperfect, and our approach embraces this imperfection rather than trying to extinguish it. Future work includes expanding of the general applicability of the method, and comparing results of the proposed integration techniques with gold standard results through statistical simulation studies. Moreover, to demonstrate its utility the developed approach will be applied to real world biomedical and health datasets such as the studies in SLE that were used for our examples.

Acknowledgement

This study received financial support from the Engineering Physical Sciences Research Council (EPSRC) Doctoral Training Partnerships.

References

[1] MASTERPLANS Maximising SLE Therapeutic Potential by Application of Novel and Systematic Approaches [Internet]. Lupusmasterplans.org. 2019 [cited 14 October 2019]. Available from: http://www.lupusmasterplans.org/home.html

[2] W. Sujansky, Heterogeneous Database Integration in Biomedicine, Journal of Biomedical Informatics, 34 (2001), 285–298.

[3] S. van Buuren, K. Groothuis-Oudshoorn, mice: Multivariate Imputation by Chained Equations in R, Journal of Statistical Software 45 (2011).

[4] MissForest—non-parametric missing value imputation for mixed-type data, Bioinformatics 28 (2011), 112–118.

[5] S.U. Yasuda, L. Zhang, S.M. Huang, The Role of Ethnicity in Variability in Response to Drugs: Focus on Clinical Pharmacology Studies, Clinical Pharmacology 84 (2008), 417–423.

[6] R.R. Shah, A. Gaedigk, Precision medicine: does ethnicity information complement genotype-based prescribing decisions? , Therapeutic Advances in Drug Safety 9 (2018), 45–62.

Appendix F: Research data repository

I am in the process of uploading my code, figures etc in a public repository. I am making it available in GitHub via this link: https://github.com/alexiasampri/Doctoral-thesis.