A Big Data Analytics Short Course for Epidemiology in Africa

Sickle
In Africa

Final project report

Team members

Training & Curriculum Development

Jack Morrice

Gaston Mazandu

Arthemon Nguweneza

Mario Jonas

Kenneth Babu

Annemie Stewart

Mamana Mbiyavanga

Christian Bope

Emile Chimusa

Nicola Mulder

Syntia Munung

Victoria Nembaware

Raphael Sangeda

Ambroise Wonkam

Logistics & Admin Support

Gabby Emjedi

Mzo Tutuka

Khuthala Mnika

Noluthando Manyisa

Chandré Oosterwyk

Victoria Nembaware

Executive summary

Both molecular data and clinical data for genetic disorders such as Sickle Cell Disease will be high volume, high variety, high velocity, and in need of experts to collect, store, clean and analyse. The Sickle Africa Data Coordinating Center (SADaCC)—a member of the Sickle In Africa (SIA) consortium—recently developed and implemented a pilot course on Big Data Analytics in Epidemiology for SIA research centers in Ghana, Nigeria, Tanzania, and South Africa; this report covers the basic course design, highlights some of the key successes and shortfalls, and provides a list of recommendations, based on the findings, for the expansion of this training programme in the near future.

Methods

Training topics included programming, statistics, database harmonization, and others central to addressing the challenges of working with big shared datasets. Enrolled learners came with a range of skills, occupations and schedules; the course combined a range of online and traditional classroom methods to optimise the learners' experience given this heterogeneity. Instructors were given targeted learning outcomes and competencies to guide development of their course material.

Findings

Instructors responded well to the invitations, producing excellent lectures, tutorials, and exercises, which have been archived for future training iterations. Student engagement was positive overall. Assignment submission rates were low, from 25-60%, but the assignments themselves were generally of good quality: no mark fell below 50%, and many were as high as 80% or 90%. A classroom survey revealed that most participants recognised the need for data quality, but were not clear on standard definitions or assurance procedures. We had mixed experiences with the online classroom technology, and in response we modified our online classroom design during the course. Making sure that all participants had the prerequisite software on their personal machines was a significant challenge.

Conclusions & Recommendations

There is considerable interest in Big Data for Epidemiologists but, as of today, not many courses available in Africa address this demand. We expect in general that applicants will come with a range of computer proficiency, which any course on this subject must accommodate. We found that a blended learning approach was possible, and we provide recommendations for optimising its impact. To scale up the course, it seems necessary that learning personas are designed to represent the key SPARCo roles, and these are given distinct, but interacting, learning paths through the course. We also recommend the purchasing of computing infrastructure and IT support staff for SPARCo sites to facilitate such training and improvement of data quality and management.

Contents

Team members **2**

Executive summary **3**

Methods 3

Findings 3

Conclusions 3

Contents **4**

A | Introduction **6**

Project aims 7

Comparison with similar courses 7

B | Course overview **9**

Participants 9

Situational constraints 10

Learning objectives 10

Evaluation methods 10

Instructional strategies 11

Online sessions 12

Workshop 12

Course content 12

Statistics & computational methods 12

Data quality & harmonisation 13

C | Findings **14**

Formative assessments 14

Multiple-choice quizzes 14

Linux & R assignments 15

Summative assessments 15

Next generation sequencing project 17

Data quality survey 17

End-of-course evaluations 18

D | Discussion **19**

Recommendations 21

Conclusions 22

Acknowledgements 22

E | References **23**

F | Appendix **26**

Online sessions 27

Day 1 | March 29th 27

Day 2 | May 3rd 27

Day 3 | May 31st 27

Workshop program 28

Helpful tags: 28

Session abbreviations 28

Monday 28

Tuesday 29

Wednesday 30

Thursday 31

Friday 32

Summative assignment examples 34

Data quality survey 36

Course evaluation 37

A | Introduction

The Sickle Africa Data Coordination Center (SADaCC)[1] was established to provide technical support and training for the Sickle Pan-African Research Consortium (SPARCo) across Africa—at present Ghana, Nigeria, and Tanzania—working to reduce the burden of Sickle Cell Disease (SCD)[2] on the continent; while SPARCo has a very specific set of aims, it also exists against a backdrop of generally low state finding levels, poor access to resources, and deep social inequalities[3]. This report describes the design and implementation of a short pilot course by the members of SADaCC: we directly targeted the training requirements of SPARCo, with sensitivity to the African research context.

A key goal of SPARCo is the development of a SCD database for a large multinational African cohort[4]. The phenotypic variations in SCD cases are not well mapped out across Africa, and what information does exist is not centralised, or easily accessible. SPARCo also aims to strengthen skills in health and research and to plan research studies of its own, so that this data may be effectively utilized by researchers from the same communities as the patients described in the database.

SADaCC aims to support SPARCo, building capacity in the following areas:

- Data collection and management (databases)
- Analysis of large diverse data sets (Big Data analytics)
- Design of studies that provide results to shape health policy (epidemiology study design)

For SCD research in Africa these three areas intersect in unique ways and, though considered highly computer-intensive in Europe and the US, they are tightly constrained in this context. As we will discuss in this report, the lack of sufficient computer resources available to members enrolled in this course proved a significant and recurring barrier to learning.

Any courses that SADaCC develops will fit within a rich global network of online classes, training events and workshops, and degree programmes. Though currently not operating any pan-consortium courses of their own, a subset of SPARCo's training needs can already be met to varying degrees by elements of this global network. For example, there are many good courses on R programming with a slant toward public health[5–8] that have been designed to suit a range of learners around the world. Closer to home, initiatives like H3Africa (H3A)[9], the African Centers of Excellence (ACE)[10] and African institutes of Mathematical Sciences (AIMS)[11] offer courses in health science and bioinformatics tailored to raising *African* capacity. However there are key elements of the SPARCo enterprise not captured by any combination of these courses, as well as important perspectives unique to SPARCo that they miss. The course we describe here attempts to fill some of these crucial gaps.

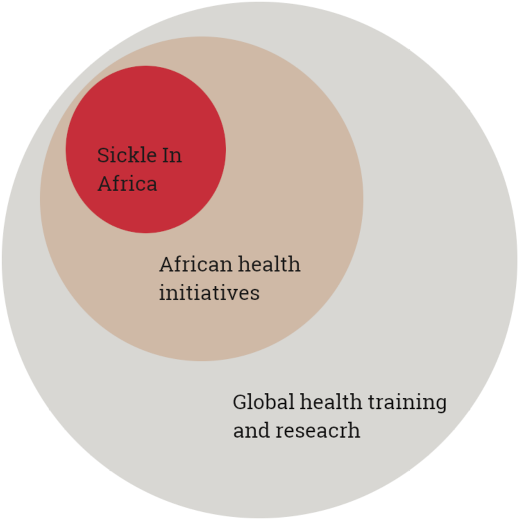

Sickle In Africa
African health initiatives
Global health training and reseacrh

figure 1: Sickle In Africa is part of African Health research, itself part of the global context. While a number of training opportenities are aimed at the African context, a much larger but less relevant pool is open to global health

Project aims

It is important to design a *scalable* course, since SPARCo itself is intended to expand to all sickle-affected parts of the continent, and a *reusable course,* as new researchers will be recruited all the time. This is the first course of its type, and very few courses have been designed specifically to address the demands of Big Data or public health training for pan-African audiences, so it is important that the course be *flexible* since we are still exploring the territory, and still investigating what ideal form such a course should take.

Large amounts of data have already been collected by SPARCo researchers and affiliated clinicians, and are ready for thorough cleaning and analysis; the pilot course must also serve to guide this process. Young data managers and analysts competent in the most up to date methods are needed now to shape the course of the consortium from the start, and as their feedback shapes the development of the course, so too the course will steer the progress of the consortium.

Finally, we aim for the course to provide a catalyst for collaboration. Students should be encouraged to work together on course assignments and projects, and these interactions will then form a base for collaboration in the real world.

With the above taken into consideration, the SADaCC Big Data short course was initiated as a project with the following aims:

- to create a scalable course that could be reused by SPARCo members as a core training instrument, and improved—using assessment, evaluation, and student feedback—over each successive iteration;
- to engage current SPARCo members with modern Big Data principles, tools, and methods;
- to facilitate intra-continental collaboration by connecting researchers across SPARCo sites.

Since this course was a pilot, it mainly focused on resolving logistic obstacles, and finding optimal paths to teaching, though assessments and evaluations were also used to look at the impact of the content and content organisation.

Comparison with similar courses

As discussed above, SPARCo members present and prospective need access to training in the management of data, the analysis of data, and the design of epidemiology studies. In more detail, we can see this will require the communication of key skills in programming and statistics, server computing and data security, analysis methods and study design principles. R is a very popular language in the health sciences[12–14]; servers ubiquitously rely on Linux- and Unix-like environments[15,16]; databases are often constructed and queried with SQL[17,18]; many studies in public health use REDCap[19–21] to create surveys and collect data. Data steward, data analyst, system administrator, clinician—many roles comprise the successful SPARCo enterprise, and in this environment these roles closely interact.

The explosive rise of online learning technology over the recent decades, much of it free, goes a long way to covering these needs. Online learning platforms like Coursera[22], Khan academy[23], edX[24], and Future Learn[25], offer flexible training in clinical data management, introductions to R and SQL, REDCap, and many more; the list is endless, and a good deal of the opportunities are free, or at least cheap. They are typically informed by the most up to date evidence-based pedagogy research, and often the only resource requirements are time, and a strong internet connection. The European Bioinformatics Institute[26] and other health related initiatives provide regular training events, or online courses and webinars, that vary from an hour to a month of required interaction time on the part of the students. Summer and winter schools, on epidemiology or genomics for example[27,28], are also available for registration, though these are often the most expensive option.

On the continent, initiatives that are leading the way in terms of health science focused capacity building are H3Africa (and H3BioNet), ACE, AIMS, and the Mandela Rhodes and MINDS scholarships[29,30]. Though the choice within this restricted pool naturally is less wide, the focus is squarely on Africa: the combat of Tuberculosis and malaria; health research and education in low resource settings; training events held in Abuja, Nairobi, Cape Town. On their website, H3bioNet list their past training events as well as those upcoming, and online courses on Bioinformatics, which have archived content accessible via website links[31]. The AIMS centers offer intensive one-year masters degrees in varied aspects of the mathematical sciences to students from the African continent, and the ACE project channels funding into promising university departments and institutes in west and central Africa along certain thematic lines, health being a major theme.

Given this extensive list, we might hope that some well curated amalgamation of the above opportunities could be used to provide a patchwork covering of SPARCo's training requirements. Certainly it will be hard to develop a basic introduction to R (or SQL) that matches the best open courseware currently available, for example through datacamp[32] or code academy[33], and who better to learn about REDCap than from the source itself (see the coursera course "Data Management for Clinical Research"[34] convened by Vanderbilt university). What this patchwork would miss, however, is a unifying coherence of content, the platform for collaboration and networking between SPARCo sites, and course ownership; by creating a course of our own that takes influence from the wealth of available material online, rather than simply directing students to it, we are able to blend the content together, use examples taken directly from our work in SCD as pedagogic aids during instruction, encourage students to work together across SPARCo sites, and develop and change content for future iterations, as the SPARCo project matures.

In this report we describe the first implementation of such a course, and lay the groundwork for a library of courses on all aspects of SADaCC/SPARCo operations that we hope will in turn be of great benefit to the international health research community.

B | Course overview

We designed a short course to specifically address the training requirements of SPARCo—data management, data analysis, study design—and asked members of the various sites to enroll and provide feedback for development. Since a key aim of SPARCo is the assembly of a large, multisite, diverse database of sickle cell phenotype information across Africa, this data will be complex in the *Big Data* sense, and so part of the course aimed at skills specific to handling of Big Data. The long name of the course was:

Big Data analytics and multisite epidemiology studies.

In this section we describe the learning objectives, content structure, and teaching strategies of the course. Our findings are presented in the following section.

Participants

We intended the course for early career researchers, data managers, clerks, and clinicians from SPARCo, with an interest in, or commitment to, reducing the burden of SCD on the continent. To facilitate collaboration, we aimed to enroll participants on this specific range of career paths so that shared languages develop between, for example, data managers and study designers. (Typically, a single participant may occupy more than one of these roles herself). We sent out the call to all current SPARCo sites—Ghana (Kumasi), Nigeria (Abuja), and Tanzania (Dar es Salam)—and asked that all candidates submit their CVs.

The call was also extended to early career researchers in the division of human genetics, University of Cape Town Medical School, where the course was designed, to test the appeal of the content more widely. Any early career researchers in the division with an interest in Big Data Epidemiology and Sickle cell disease, and enough free time to complete the course, were encouraged to apply. The enrollment criteria was also widened in this way to observe how effectively the course could scale up to larger classrooms. The call was by invite only, and no posters or advertisements were circulated.

A basic proficiency in computer programming was required of candidates; the course was intended to teach skills which used basic coding to produce results in epidemiology, and we were aware that even the most basic coding exercises can be a significant challenge to someone who has, for example, never seen a program variable before. All candidates were required to have an MSc in Public Health, Biomedical Sciences, Computer Science, Biostatistics or similar field.

figure 2: Samuel Mawuli Adadey, University of Ghana, learning to write R code at the workshop

Amongst the candidates who might apply and be accepted, based on how we spread awareness of the course and on our chosen entrance criteria, we anticipated a significant variety in computer proficiency, statistical knowledge, and chosen career path. We also allowed for our participants to live in different time zones, and work to very different schedules.

Situational constraints

The key constraints, like any project, were of time and resources. The instructors and organisers had limited time to plan and create content, and computing resources could only be made available to participants while they were at the University of Cape Town.

The instructors, convenors, and organisers all volunteered their time for the project, and all had to balance the course-related work with their own. All the content—lecture slides, assignments, tutorials—was created from scratch, on donated time. There was sufficient funding to hold a week-long face-to-face workshop, and so, to get maximum benefit from this time, three extra online sessions were held before this workshop on aspects of the course well suited to distance learning.

Strong internet connections could not always be relied on, therefore prerequisite software packages had to be chosen carefully so as to be downloadable by all course participants. No proprietary software could be used in lessons (expect basic cases like Microsoft Windows) since licence access could not be ubiquitously assumed. Gotomeeting[35] was the platform used to host the online classrooms where the online pre-workshop sessions took place. The GoToMeeting license used by the course instructors was shared with other research groups, which occasionally created scheduling conflicts.

Learning objectives

We derived a set of learning objectives for the course from the SPARCo training requirements, and the project aims, outlined in the introduction of this report. A concrete aim of the course was that participants who successfully completed the programme would be able to provide important assistance to a research meeting taking place in November of the same year, 2019. In addition, we aimed to meet the SPARCo training needs through a set of objectives that centered on data cleaning, analysis, and study design, from a Big Data perspective, using the SPARCo sickle cell database as a key case study.

In the planning phase of the project, it was initially intended that, by the end of the course, participants would be able to:

1. describe the basics of Big Data analytics and tools;
2. draft a proposal for ethics clearance for a multi-site retrospective study;
3. apply basic epidemiological methods;
4. design a multi-site retrospective epidemiological study;
5. deal with missing data;
6. process epidemiological data from multi-site observational studies;
7. attend the November meeting (on the Sickle Cell Disease Ontology) and to provide study design support.

Given the pilot nature of the course, this set of objectives allowed some flexibility so that the course could adapt to hidden challenges, or address unanticipated training needs.

Evaluation methods

The course employed formative and summative assessments in its design[36]. Formative assessment was used during the online sessions, particularly in the computational parts of the course, to assess how well participants were responding to the online content, and to locate any weak areas to be improved during the online course and the face-to-face workshop; summative assessments were set at the end, to measure the overall impact of the course. A data quality survey was administered towards the end of the workshop, and course evaluation forms were made available after the course had finished.

Since the course was a pilot, the assessments did not extensively cover all aspects of the course taught, but were intended mostly to measure general engagement, and to provide the students with some means of practicing the skills they were learning. In future iterations, we intend to monitor the suitability and difficulty of the content more closely with more extensive testing, and to provide the students with opportunities to practice the new skills obtained from all aspects of the course, in particular, practice in areas linked to the course learning objectives outlined above.

The formative assessments covered the Linux and R content in the pre-workshop online sessions (we will give a full breakdown of the curriculum in the next subsections). They consisted of short multiple choice questions at the end of short content sections, followed by longer assignments to be submitted to the course convenor for grading. The multiple choice quiz questions were not designed to be challenging, but to test recall of simple aspects of the content provided, and in this way stimulate learning and engagement without adding pressure or stress to the student learning experience. The assignments were a little more challenging, and each question was marked simply out of 2 (0 for no attempt, 1 for an attempt, 2 for the correct answer). Written feedback was returned to the students with their marks.

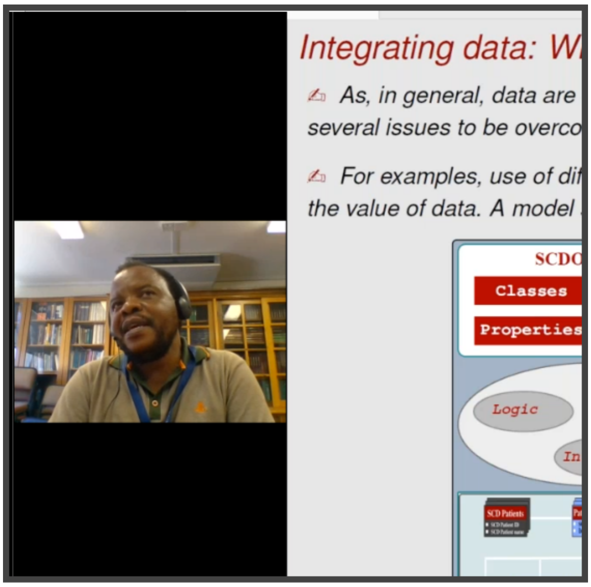
The summative assessments set at the end of the course tested the content of the face-to-face workshop. There was a short set of questions assigned for each session, and again each question was marked out of 2 (0 for no attempt, 1 for an attempt, 2 for the correct answer). Since some of the material in the online sessions reappeared in the workshop (the use of R, for example) this in principle could allow us to crudely track progress of the students who submitted this work. Feedback was the only incentive offered for participants who completed the assignments. Examples of these assignments are included in the appendix of this report.

The data quality survey was created using REDCap, and links were sent to all participants to allow access to the instrument. The questions were anonymous, no names or identification information was attached to the responses, and 15 minutes was given during the workshop for participants to complete the survey, which helped guarantee a higher response rate. The links were sent out after most workshop sessions on data quality had been given, so as to measure their impact on the audience. These survey questions can also be found in the appendix.

After the course had finished, participants were asked to review the course using electronic evaluation forms. Most of the questions took the form of a statement—"the online classroom service (GoToMeeting) was used effectively"—and students were asked to agree or disagree, on a scale from "strongly disagree" to "strongly agree". In this way leading questions were avoided, and ambiguity minimised. All the evaluation form questions and results can be found in the appendix.

Instructional strategies

As we have described, the course was organised in blended learning structure[37–39]. SPARCo is a distributed consortium, with members spread across Africa, and our course had to work equally well for all members. A face-to-face workshop lasting a week was held, with 3 days of online sessions leading up to the workshop, each day separated by a month, with assignment work set for the days in between.



Integrating data: W

As, in general, data are several issues to be overco

For examples, use of dif the value of data. A model

figure 3: Dr Gaston Mazandu lectures on integrating Big Data sets in Epidemiology via gotomeeting

All the content, for both the online sessions and the workshop, was centralised online, using the University of Cape Town's Learning Management System: Vula[40] (a fork of the popular open source system Sakai[41]). A private course intranet website was created on Vula, course participants were added to the site as learners, and all the course content was stored here in resource folders and on specially designed course webpages. For the online sessions, we used the GoToMeeting platform to host the virtual classrooms. Course announcements were circulated using Vula, and forums—subject specific and general—were opened on the Vula site for students to engage with each other remotely. The GoToMeeting classroom is shown in figure 3, and the homepage of the course website in figure 4.

Online sessions

The dominant strategy used in the online sessions was webinars hosted on GoToMeeting[1]. Announcements were circulated to all participants with links and instructions to the GotoMeeting platform. Guest instructors from the Division of Human Genetics, Department of Pathology, University of Cape Town were given a few weeks to prepare webinar sessions on topics selected from the course curriculum, and a time slot to present. Live webinars were recorded using GoToMeeting's record feature, and these recordings with their associated slides (pdfs) were uploaded to the course Vula site. Questions by participants could be asked using GoToMeeting's text chat feature, or via a participant's computer microphone. The webinars were broadcast from a physical seminar room at the University, and local participants were invited to attend in person.

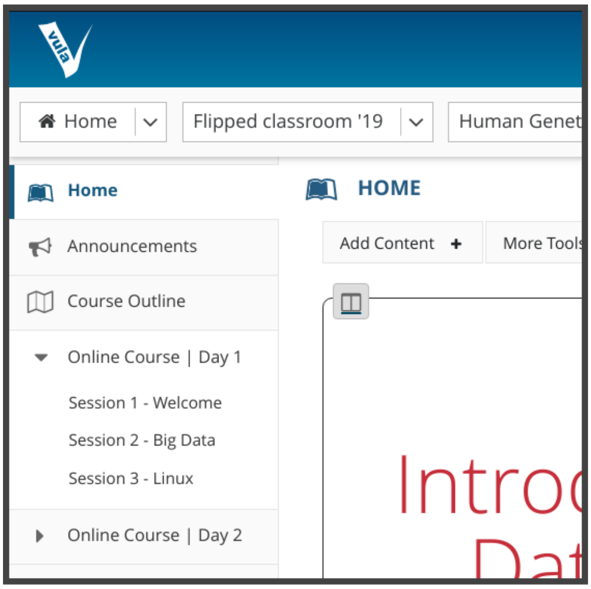


figure 4: home page of the course website, hosted by the learning management system Vula

In addition to the webinars, we also designed tutorials. The students were asked to read some content, text we had written and uploaded to Vula, and then answer questions on the content. The answers to the assignments, which we discussed in the subsection on evaluation methods above, were then to be submitted to us for marking. students were given until the next course day, approximately one month, to complete the exercises.

Workshop

Guest lecturers from the department were invited to design lessons that spoke to a particular part of the curriculum and, as for the webinars described above, the instructors were given freedom in how they interpreted their designated part of the course. Many instructors, particularly those teaching computer-based skills, chose to use a lecture session, followed by a hands-on session that often involved group work. A great example was the session on machine learning. The instructor wrote a tutorial using R markdown[42], which they demonstrated in class. Participants in groups were able to execute the tutorial code in real time along with the teacher.

Alongside these lectures, a small amount of group work was organised—the participants were divided into 3 groups, and asked to conceive an epidemiology study based on their own research interests. In their groups they consolidated the knowledge and skills they learned by incorporating them into their study designs. The groups then presented their study designs to the rest of the workshop and some of the course organisers and instructors, who asked probing questions and offered study design feedback.

Vula has inbuilt features for designing course evaluation forms. We used this tool to create a general evaluation form for the course.

Course content

The content was designed to meet the learning objectives outlined above. It was grouped into 2 modules: *Statistics & computational methods*, and *Data quality & harmonisation*. The context of each was epidemiology and Big Data. We include a full catalogue of sessions that comprised the course in the appendix of this report.

Statistics & computational methods

In this module, guest instructors designed and gave courses on statistical distributions, power calculations, R, python, SQL, as well as machine learning, and Linux: participants were asked to install Linux operating systems on virtual machines using the software Virtual Box[43] as well as to explore a real high performance Linux server.

[1] https://www.gotomeeting.com/en-za

In most cases these were introductory level; though some students had seen these topics before, they were not assumed as course prerequisites.

The above topics were linked; for example power calculations were demonstrated in R, and the participants often used their own Linux distributions they had installed on their laptops to download and use the database environment *mySQL*[44]. Distributions were introduced first by their mathematical definitions, and then as probabilistic functions in R.

The R, bash, and python programming languages were made central to the computing side of the course, since reliable Big Data methods—in both handling and analysis—do not exist for GUI technologies. We note that Big Data files can not be opened in excel, and so it is important to familiarise students with these alternate tools right from the start. In line with our first learning objective, students were expected to be able to list these methods and explain why they are particularly important to Big Data.

A workshop on Next Generation Sequencing that followed the one described in this report, and was attended by the same participants, used Linux servers and the command line entirely; participants were exposed to a week-long sink-or-swim walkthrough of genomics pipelines using the bash shell.

Data quality & harmonisation

The other side of the course focused on the collection, storage, and cleaning of data sets in epidemiology: electronic data capture with REDCap; the importance of standardising sets of data elements across sites of a consortia, and methods for retrospectively harmonising data; the use of ontologies[45,46] in standardising not only data elements but also research practice and publications across whole fields; the standard dimensions of data quality[47], including the ethical dimension; and how to clean a data set in REDCap, or deal with missing data in R (learning objective number 5). Though we touched on more advanced aspects of data quality, such as ontologies and the FAIR[48] principles, we kept the hands-on demonstrations and exercises to more pragmatic aspects of the SPARCo research programme, like the use of REDCap in electronic data capture, and the use of python to retrospectively harmonise distinct datasets with distinct code books.

We emphasized the dimensions of data quality—completeness, consistency, conformity, accuracy, integrity, timeliness—as a means to ensure that participants designed studies to collect good quality data, but the ethical dimensions were also highlighted, in line with our second learning objective.

The importance of data set harmonisation, retrospective and preemptive, was stressed in several sessions of the course, and was key to addressing our sixth learning objective. Some instructors had first hand experience with retrospective data harmonisation (and some horror stories to tell) while others were central in the development of the Sickle Cell Disease Ontology[49].

C | Findings

The epidemiology course ran from the 29th of March, to the 14th of June 2019, and most participants went on to attend a related workshop the following week, 17th till the 21st June (inclusive), on Next Generation Sequencing. This second workshop also used our Vula course website to register students and host content. The online sessions took place on the 29th of March, 3rd of May (the original date for day 2 lay on a public holiday in Tanzania so was subsequently changed) and the 31st of May. The workshop took place in a computer lab of the University of Cape Town's medical school library, from the 10th to the 14th of June (inclusive).

In total, the course website had 46 registered users. Of these, 3 were site owners (i.e. administrators) and 6 were support staff (able to directly add and change content). These Vula-defined roles, however, did not perfectly reflect the roles of everyone involved in the course. Of the 46 registered, a total of 34 were learners; 9 of these learners were from the SPARCo sites and 14 were only registered for the NGS workshop that followed. The remaining 12 participants registered to the Vula site (46 minus 34) were volunteer instructors. A team of support members provided logistic support, some of whom were registered on Vula but not all.

These categories overlapped, as some invited instructors also joined other sessions to learn, or to help with the catering and organisation. Some learners also provided invaluable logistic support throughout the course. The 9 participants from the SPARCo sites joined the online sessions remotely from Nigeria, Ghana, and Tanzania, and 7 of these participants flew to Cape Town in June to participate in the workshop.

The forums we added to the Vula course website were not utilised: most sought advice via private email. We collected no data on visits to the Vula course website, nor did we take strict attendance during the webinars or workshop sessions, which means we can not present accurate measures of general attendance in this report. Our sole data source that reflects the engagement levels of the participants is the assignment submissions which we summarise in the rest of this section. The names are removed to respect privacy.

Formative assessments

As described in the Course Overview section of this report, the formative assessments, in the form of tutorials, covered material in the online sessions. For each tutorial the students were given several short pages of online text to read, with one or two multiple choice questions at the bottom of each page, and a longer assignment at the end of the tutorial. There were two tutorials of this kind, one on Linux and another on R. The Linux tutorial was set on online day 1, and due before day 2. The R tutorial went live after day 2 and was due before day 3. We describe the results below.

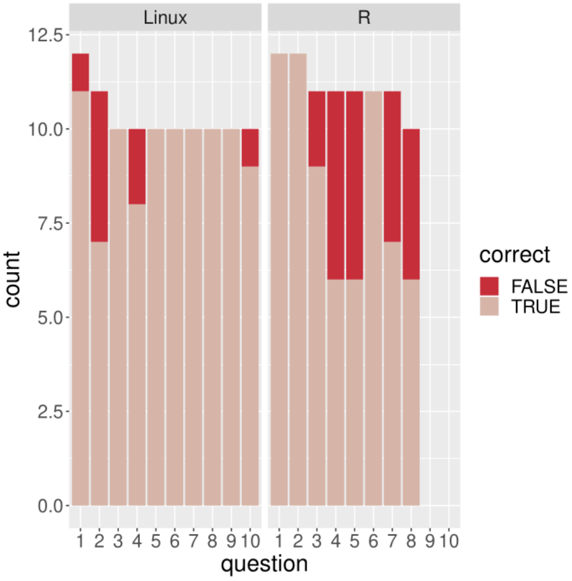


figure 5: Number of attempts per multiple choice question (red bars); number of correct attempts per question (beige bars). Left hand panel shows the Linux questions, and R questions shown on the right.

Multiple-choice quizzes

There were a total of 18 multiple choice questions, 10 on Linux and 8 on R. On average 11 participants attempted an answer per question with a standard deviation of 1; an average of 9 of these attempts were correct, per question, with a standard deviation of 2. An average of 6 of these attempts were by SPARCo members.

Comparing the tutorials, the Linux quiz questions had an average of 10 attempts, while the R questions had on average 11 attempts; the spread (standard deviation) was 1 in both tutorials. However, when comparing the average number of *correct* answers per question, there was a slightly more pronounced difference between the tutorials: 9 participants answered the Linux questions correctly per question, with a spread of 1, and only 8 answered the R questions correctly but with a spread of 3. This information is summarised in figures 5, 6, and 7.

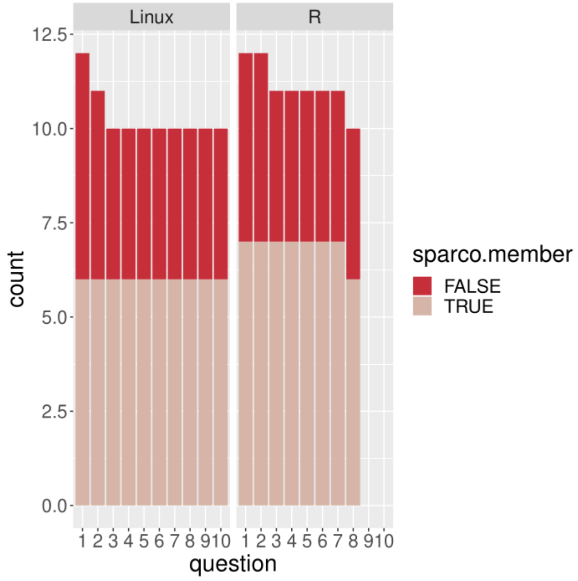


figure 6: Number of attempts per multiple choice question (red bars); number of attempts per question by SPARCo members (beige bars). Left hand panel shows the Linux questions, and R questions shown on the right.

Linux & R assignments

The response rates for the tutorial assignments were much lower than for the quiz questions: only 6 participants submitted answers for the Linux assignment, and 5 submitted answers for the R assignment; of these, 5 and 5 submissions were from SPARCo members respectively. The quality of these responses however was very good: the mean mark, as a fraction of the obtained marks over the maximum number possible, was 0.93 for Linux, with a spread of only 0.1, and 0.85 for R, with a wider spread of 0.2.

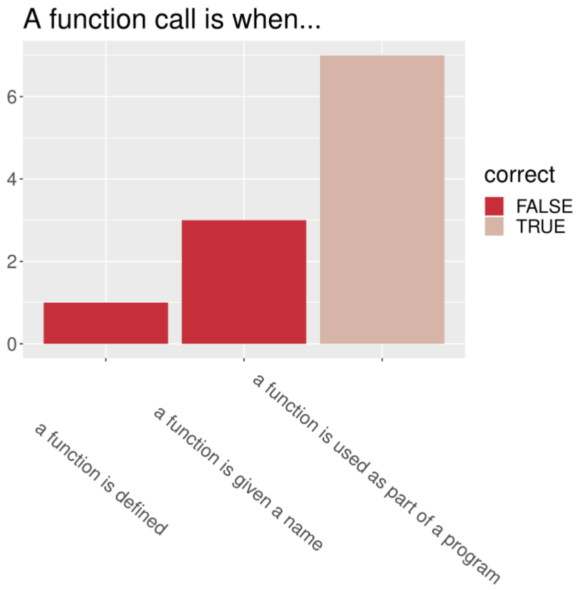


figure 7: Number of attempts per possible answer for the multiple choice question "a function call is when...". Red bars are incorrect answers, the beige bar is the correct choice.

A common mistake in the Linux submissions was in the writing of simple shell scripts (in fact, all mistakes on this assignment made were on this question): not everyone understood that the example script given had to be saved as a bash script, compiled, and executed. The most common source of error in the R assignment was the question that asked participants to write and use their own mean function.

The difference between SPARCo and non SPARCo member response rates, for both the multiple choice questions and assignments, are shown in figure 8. The distribution of responses for the Linux and R assignments are displayed as a violin plot in figure 9.

Summative assessments

There were many sessions during the week-long workshop covering many different aspects of the course content described in the Course overview section of this report, however we only provided monitored assessment for 6 of them: Databasing, Linux servers, Power, Python, Statistics, Tidyverse[50]. In addition, many of the participants (including all those that were also SPARCo members) went on to complete a course on Next Generation Sequencing the following week; this second workshop had a significant project component and we have access to the submission numbers for this project but, unfortunately, not the final marks.

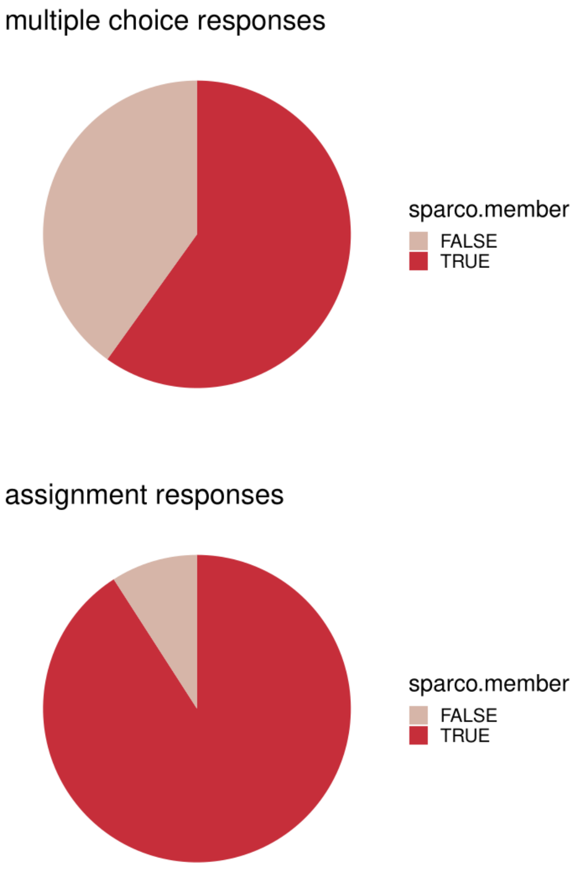


figure 8: ratio of the total number of attempts by sparco members (red sector) to the total number of attempts for non sparco members (beige sector) for all multiple choice question attempts (top) and assignment submissions (bottom). Both Linux and R tutorials are combined.

In table 1 we show the number of submissions received for each of these assessment units. Completion of the Linux server assignment relied on access to a high performance computing server to complete, which was provided during the workshop by our group and the University of Cape Town, however this access was for a very limited period for each user, and the access period ended before most students could complete their assignment, which is the main reason why so few Linux server submissions were received. The Statistics assignment was not hosted in the same location on Vula as the others, and many students may have missed it because of this.

unit	**number of submissions**
Databasing	4
Linux servers	2
Power	5
Python	5
Statistics	1
Tidyverse	5

table 1: total number of submissions for each summative assignment unit.

The results for these submissions are shown in table 2 and figure 10. Overall—neglecting the Linux servers and Statistics units for lack of responses—the participants performed best, with least spread, on the Databasing assignment, and worst, with much greater spread, on the Tidyverse unit. The unit on Power, actually a unit on the 'pwr' package for power computations in R[51], was the most varied in terms of submission marks.

unit	**mean mark**	**spread of marks**
Databasing	0.917	0.096
Linux servers	0.938	0.088
Power	0.725	0.347
Python	0.85	0.224
Statistics	0.75	NA
Tidyverse	0.72	0.249

table 2: mean fractional mark (mark obtained over total possible mark) for all summative assignment submissions. NA - not applicable. The ‘spread’ refers to the standard deviation.

The main source of errors made in answering the Power unit questions was in the interpretation of a particular plot; participants were asked to interpret the change in a graph that plotted sample size (x-axis) against statistical power (y-axis), with effect size—the main difficulty seemed to be the *interpretation* of this change in statistical terms. Drawing the plot relied on the use of some R plotting functions, and some functions in the pwr package, but this task was managed successfully by almost all participants who attempted the assignment.

Next generation sequencing project

A total of 22 participants submitted their work for the next generation sequencing project. Of these, only 4 were submitted by SPARCo members. We do not have access to the results of these submissions.

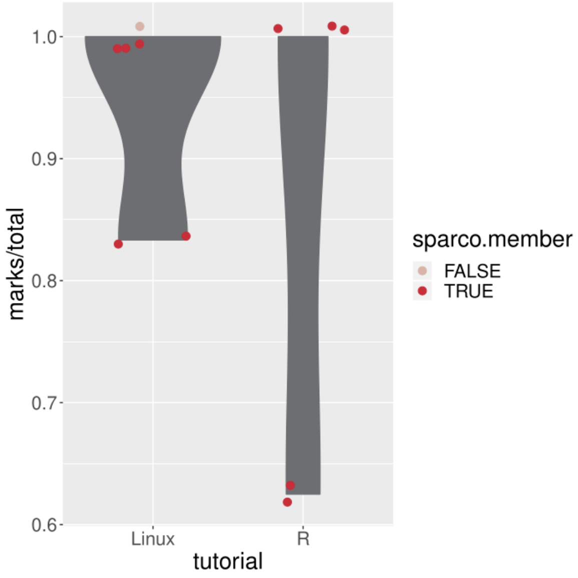


figure 9: distribution of marks, shown as marks obtained over total possible mark, for all Linux (left) and R (right) assignment submissions by sparco members (red dots) and non sparco members (beige dots). Violins are drawn for sparco and non sparco members combined.

Data quality survey

A total of 13 participants completed the survey. The survey was 7 questions, one for each of the following dimensions of data quality: accuracy, completeness, integrity, consistency, validity, timeliness, and security. Each question asked "How would you improve the <blank> of the data at your site?" where <blank> refers to each dimension listed above. We briefly summarise the answers given to some of these questions below.

Accuracy: the responses were very varied. Some mentioned concrete steps to take based on an understanding of REDCap: "I would [...] encourage raising and resolving quarries of questionable data", whereas others either simply mirrored the question: "Make sure what data captured at the site is the truth", or offered ambiguous advice: "Cleaning the data".

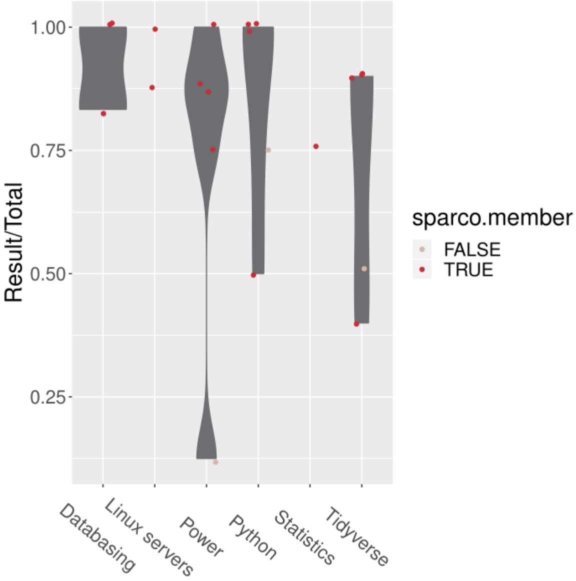


figure 10: distribution of marks, shown as marks obtained over total possible mark, for all summative assignment submissions by sparco members (red dots) and non sparco members (beige dots). Violins are drawn for sparco and non sparco members combined.

Completeness: more concrete advice was given in response to this question than the previous one, and more direct references to REDCap and its functionality specifically were made. For example: "Prevent collection of new variable if the preceding one is vacant by adding a clause to the variables itself", "By adding some validation to the form", "Implement systems checks e.g.Redcap, Regular reports (Weekly and monthly)".

Consistency: some answers were insightful: "By developing ontologies and metadata information for dataset and data collection tools", but many others were vague, and did not display an understanding of data consistency or why consistency is distinct from the other dimensions: "By setting goals", "clearing the data", "Robust training and retraining".

Timeliness: many responses did not directly address this dimension. Some offered advice that might indirectly help improve timeliness: "The use of Technology(e.g . From paper based systems to Electronic", but again many answers did not distinguish timeliness from the other dimensions: "Firstly, data must be collected on time from the right source. Ensuring that necessary rights and privileges are given to the data personel" whereas other responses again mirrored the question: "To ensure that the data input and processing is done on time ".

Security: this question received the most focused responses; participants spoke directly to issues of data security, and made clear allusions to REDCap security features as well as other cybersecurity tools: "By restricting everyone from having full admin privileges to the data", "establishing control access and define a secure confidential waste bin", "having a privacy clause, information can only be assessed by asking for permission.", "Regular Backup of data (weekly), and security measures to the servers and systems".

End-of-course evaluations

In total we had 11 responses to our evaluation form, which was available for responses after the workshop, from the 7th to the 17th of August 2019. The responses to each question were on the whole positive: most were either "neutral" to "strongly agree", given that the questions were worded positively rather than negatively ("the lessons were all at the right difficulty for me"). All feedback was de identified.

The most positive responses were for the questions "I had access to the necessary computing resources throughout the course", and "the course website (Vula) was easy to navigate"—both questions received 73% "strongly agree".

The least positive responses were for the questions: "the instructors returned assignments and exams in a timely manner" (30% disagree), "the lessons were at the right level of difficulty for me" (9% strongly disagree), and "I enjoyed the online component of the course" (9% strongly disagree).

While only 9% strongly agreed with the statement "I enjoyed the online component of the course", 55% said they strongly agreed with: "I enjoyed the face-to-face workshop". Most seemed to agree that the course content was aligned to the advertised learning objectives, and further that these learning objectives were well aligned with their own private research needs and goals.

> **"I wish, put together, it would have been at least a month long face to face session for both multi site epidemiology and Next Generation Sequencing."**

A general theme in the comments was that participants would have liked the face-to-face workshop to be longer, some responses quoted “a month”, and that the online component be shorter, or not existent: "I wish, put together, it would have been at least a month long face to face session for both multi site epidemiology and Next Generation Sequencing." Participants said they would have liked to see more "Hands on exercise with the programming languages", more "Practical face to face session with with some theory on specific data analysis(More of Bioinformatic).", and less "online session".

D | Discussion

We here discuss our main findings, and the implications of these for the design of future iterations. In this final section, we will also compile some recommendations derived from our experience that we feel are important for future designers, planners, and instructors to be aware of, and make our conclusions. Alongside these discussions we emphasise caution to the reader: our enrollment numbers were small, and the numbers of engaged participants (i.e., those who submitted assignments) were even smaller. Therefore any patterns we discuss or extract advice from are merely suggestive.

We had a total of 21 registered learners on the course, and 12 guest instructors. This implies a ratio of almost 2 learners to every teacher, though in practice it did not work out quite this way; most instructors were only available for the short period of time that was their allocated session. The instructors did however all create their own content and plan their lessons, which were all of very good quality. Further, the teaching aids—slides, questions, example codes, and so on—are all now archived to be changed and reused in future course iterations (with the appropriate credit). This reusability will save a lot of time, since lessons no longer need to be built from scratch.

Of the learners, 9 were initially SPARCo members, but only 7 of these were able to attend the workshop. This means around a third of the learners were SPARCo participants, yet as we saw in the previous section, over 50% of the submitted work for each assignment was from SPARCo members, in some cases this was as high as 80% or even 100%. Clearly these students were the most engaged of the learners, and derived the most from attending the course. Much of the course was directed specifically at SPARCo needs, managing the SCD database or standardising the registry data elements for example, and this is probably for the better. It is not clear if opening the call more widely to other early career researchers was beneficial for these researchers or to the SPARCo members, but it appears that the former group did not feel like an integral part of the program.

The response rate for the multiple choice quizzes was around 50% of all the learners. Although low, this was higher than the other assignments, and the quizzes provided a good source of information on who was reading the content, how well they understood it, and where the weaker areas were; multiple choice questions are not hard to write, they take little time to answer, and they are a great way to get snapshots in time of the student proficiencies as the course progresses. By monitoring responses we were able to see for example that coding experience, in line with our initial expectations, was varied and low on the whole, and this led us to place more emphasis on programming in the face to face workshop.

For the formative assignments on Linux and R, the response rate was much lower, around 25-30%, and this time the significant majority of submissions came from SPARCo members. We again found that coding was an issue with most participants, and the issues worsened with code abstraction—while perfectly able to *use* R functions, for example to compute means or draw simple histograms, writing their own simple examples proved a challenge. Because the response rate from non SPARCo members was so low, we cannot know how much these learners took away from this part of the course, but we can certainly see their engagement levels were not high. The same patterns were observed in the summative assessments at the end of the course: response rates were around 25%, and the majority of the submissions came from SPARCo members. The questions with the lowest success rates and most variability were those related to the use of the R language, which again supports our initial intuition about coding proficiency.

The data quality survey had a higher response rate, around 60%, and this is most likely because the survey was administered *in class*. The responses showed that the importance of data quality was appreciated, but the classification of the various dimensions of quality was not familiar, and very few concrete quality assurance measures were cited. Answers displayed a lack of understanding of the standard definitions, and the different data quality dimensions were not well distinguished in the responses. This perhaps indicates that our learning objectives relating to data quality were not well communicated through our choice of content, and that this aspect of the course needs some revision and more focus in future iterations.

Users seemed on the whole to appreciate the Vula system. In the evaluation, it was voted as being useful and easy to navigate. This is backed up by the instructors, who found it easy to add content to, and create webpages for. The evaluation tool itself, provided by Vula, worked very well, and the form has been archived as a template for reuse in future course iterations.

Though reporting the ease of use of the course website, participants routinely avoided the forums; participants prefered instead private email contact with the instructors. This was a problem, as the help given to one student could not been seen by others, possibly with the same issue. For a course of 21 learners this is not so bad, particularly if less than 50% are attempting the exercises, however it is bad for scalability. For the course to scale up effectively, forums must be a key part of the learning environment, so the students are encouraged to help each other, and problems need solving only once, on a public forum, rather than many times in private correspondances.

It is curious that many learners reported in the evaluation they had access to suitable computer resources, when several instructors observed the exact opposite. In fact, the lack of suitable computer access, particularly for the remote learners, proved to be the most significant barrier to learning that we encountered, and a great deal of time and attention was misspent by both learners and instructors attempting to install prerequisite software (for example: Virtual box, Linux, tidyverse) on very old laptops. For example, almost an entire day of the face to face workshop was spent (unsuccessfully) installing Linux for the few learners in the room who were having difficulties. During this time one participant, in an attempt to resolve the issue, even removed a processing chip from their laptop by hand to clean it.

It was also reported in the course evaluation that several participants did not enjoy the online sessions, and felt that they were unorganised. This may have been due, in parts: to internet connectivity trouble both in Cape Town and the remote centers; to the scheduling conflicts of the GoToMeeting platform; and to the blended classroom model we initially adopted. The first few webinars we organised were held in a seminar room with a live *and* a remote audience; all participants in Cape Town at the time were encouraged to attend in person, and all remote participants joined the room via GoToMeeting. This meant first that we relied, unnecessarily, on the availability of a seminar room shared widely by a department but, more importantly, the voices of the remote participants tended to be lost in the room. This was exacerbated by connectivity issues. The questions after the webinars came from the room only, and not from remote learners. This improved when we decided, on the third day of online sessions, to only stream the content online.

Students also reported in the evaluation that their assignment marks were not promptly returned—one of the main challenges that we will encounter on scaling up the course is marking assignments. For small numbers, it can be done in the small amount of time the guest instructors can devote to their course segment, but as the numbers grow, this workload will grow exponentially. Strict marking Standard Operating Procedures (SOPs) must be available to streamline this process and to standardise the marking, when many instructors are responsible for marking the same assignment. Since the assignments are such an important aspect of the evidence-based development of the course, standardised marking expectations are crucial.

One aspect of the course that we did not address adequately was the heterogeneity of learners' previous experience. We anticipated a wide variety of computer proficiency, and of learners' current career paths. We saw this reflected in the variance of assignment marks. In the evaluations, some learners reported that the course material was too hard, or not at the right level for them. But in the course design we did not plan for this. We will comment on this further in the Recommendations subsection below.

In summary, we observed that while assignment response rates were not high, the submissions we did get were of good quality, and we highlighted patterns that can help guide future iterations of the course. Programming experience varies widely, however the development of this experience takes up the most teaching time. The importance of data quality was appreciated by the students, but data quality assurance as a practice was not well understood. The online components had severe logistical challenges, even though the Vula and GoToMeeting platforms did work quite well. The heterogeneity in learners' backgrounds was anticipated, present, but not handled well in the planning.

Recommendations

Based on the discussion above, we present here some recommendations for how to improve the course. These are to be taken as suggestions, and may be adapted depending on how the needs of SPARCo evolve.

1. Seek out younger instructors, and offer participation in the course itself as an incentive for them to get involved. It was a finding of note that some invited instructors joined in as learners in other sessions. This seems more likely to occur if the instructors are also early-career, and training regularly themselves.

2. Quantify your expected participants' diverse backgrounds by developing learning personas[52]. These personas will dictate what will be possible and suitable to teach, and so they should be drafted *before* the learning objectives and course content. Use these personas to plan strategies for peer to peer learning: a diverse background of experiences means that students can teach each other, thus alleviating the pressure on the instructors, and facilitating collaboration. For an online course, full use of the forums is required.

3. Outsource the basic coding lessons to available online courses, for example linkedin learning[53]. Introductions to programming, for example in R, are often quite generic, and are done exceptionally well by some online course providers. Make sure the students are well prepared in these auxiliary skills before attending the workshop, and dedicate more time to group discussions and relevant case studies.

4. Provide course specific repositories for all the prerequisite software. This includes any R packages you might want to use, or Linux distributions (most of which are around 2GB in download size). Do not include packages in lesson plans unless they are absolutely necessary, and make sure detailed installation instructions for many different computer models and operating systems are clearly posted before the course starts, so no course time is wasted on installation. An alternative is to provide special classrooms near the locations of the remote learners, and arrange local tech support.

5. Approach potential guest instructors early, say a few months in advance, and give them learning objectives to meet, rather than dictating any specific content. It is then up to them to plan their lessons as they feel comfortable, but the take-away skills will be in line with the main objectives of the course.

6. If a lesson is to be streamed online as a webinar, it should only be online. Do not organise a room of participants as part of the audience, or they will steal focus from the remote ones.

7. Establish detailed marking SOPs and assessment standards before the assignments are written and sent out. Remember, the instructors may be planning their lessons a few weeks or even months in advance.

8. Address learner heterogeneity with multiple learning paths[54] and learning personas. Some participants will be present as data managers, and keen to understand the details of mySQL. Others might be data analysts, possibly with an academic paper to write in their immediate future. Others may be clinicians, with clear research questions they have gathered from years of experience with patients. Design lessons where everyone has a place, using personas, and provide alternate learning paths through the course with optional modules and coursework.

9. Rely less on live webinars, and provide more recorded lectures, with written content that can be read in the students' own time (with deadlines). Many learners and instructors will have widely differing schedules beyond your control and, in organising too many live webinars, many participants might miss important information due to schedule clashes.

10. Be clear on who the intended audience is at the earliest planning stage and what the entrance criteria are, then write realistic learning objectives based on this intended audience. Try to avoid opening up the course to anyone interested, especially after the course has properly started, as it may affect whether the learning objectives are met by everyone, and will reduce your ability to objectively assess the successes or failures of aspects of the course based on any collected data.

11. Make note of the public holidays in all participating countries that may conflict with your teaching. Keep aware of any changes in the state of affairs in these countries, for example political protest or power cuts, that may affect your participants' ability to access information.

Conclusions

In this report we have described a pilot course on Big Data methods in Epidemiology, designed to address the specific training requirements of the Sickle Pan African Research Consortium, that ran from March to August 2019. The course focused on tools indispensable to Big Data research, including R and python programming, Linux servers and the bash shell, mySQL, machine learning, and more. The SPARCo database was used as a key case study, and a key example of the ways in which Big Data sets arise in epidemiology.

The course used distance learning methods to support a face to face workshop. It was found that a lack of sufficient access to the necessary computer resources provided the most significant barrier to learning, and that varying degrees of learners' previous experience in computer programming translated into a wide spread of assignment success and participant experience of the course as a whole. Data quality was recognised as important, but the various dimensions of data quality were not well discerned by the participants even after course completion, and practical methods of data quality assurance were not well known. We provided a list of recommendations for future planners and convenors of the course based on evidence collected throughout the course duration, and our experiences.

Big Data courses are seldom available to medical researchers in Africa, and even more rarely are they given an Epidemiology context. We believe that pan-African consortia like SPARCo will soon need Big Data experts working alongside data managers and clinicians to shape government health policy with evidence based research, and these analysts must have access to the most up to date tools and evidence-based pedagogy. This course is the first of its kind, and one of very few to address this niche; we hope it will pave the way for a library of courses aimed at raising African capacity in the health sciences, and at reducing the burden of Africa's most destructive diseases with cutting edge research and exceptional standards of patient care, though African initiatives.

Acknowledgements

We would like to thank the National Institutes of Health, in particular the National Heart, Lung, and Blood Institute; their generous guidance and financial support made this project possible. We are grateful to all our instructors who donated their time to design and teach such great sessions, to all our hardworking team members who assisted with the coordination of all aspects of the course, to everyone in SPARCo who gave their time and input, and finally to all our students, whose enthusiastic engagement and feedback will guide future developments of the course for years to come.

E | References

1. SickleInAfrica Consortium | https://www.sickleinafrica.org. Available at: https://sickleinafrica.org/. (Accessed: 25th September 2019)
2. Jackson, A. Sickle cell disease: Africa's most prevalent 'invisible condition'. *The M&G Online* Available at: https://mg.co.za/article/2019-04-04-00-sickle-cell-disease-africas-most-prevalent-invisible-condition/. (Accessed: 25th September 2019)
3. Makani, J., Ofori-Acquah, S. F., Tluway, F., Mulder, N. & Wonkam, A. Sickle cell disease: tipping the balance of genomic research to catalyse discoveries in Africa. *Lancet Lond. Engl.* **389**, 2355–2358 (2017).
4. Makani, J. The Sickle Pan-African Research Consortium (SPARCO).
5. Statistical Analysis with R for Public Health | Coursera. Available at: https://www.coursera.org/specializations/statistical-analysis-r-public-health. (Accessed: 25th September 2019)
6. https://plus.google.com/u/0/+Datacamp. Health Survey Data Analysis of BMI - R - Online Project. Available at: www.datacamp.com/projects/677. (Accessed: 25th September 2019)
7. Developing R Programming Skills (2 day course) — University of Oxford, Medical Sciences Division. Available at: https://www.medsci.ox.ac.uk/study/skillstraining/calendar/introduction-to-programming-in-r-2-day-course. (Accessed: 25th September 2019)
8. UCL. Introduction to R for Healthcare Researchers. *UCL Institute of Health Informatics* (2019). Available at: https://www.ucl.ac.uk/health-informatics/introduction-r-healthcare-researchers. (Accessed: 25th September 2019)
9. H3Africa | Human Heredity & Health in Africa | Human Genomic Research. *H3Africa* Available at: https://h3africa.org/. (Accessed: 25th September 2019)
10. ACE – African Higher Education Centres of Excellence – Association of African Universities. Available at: https://ace.aau.org/. (Accessed: 25th September 2019)
11. AIMS | Building Science in Africa. Available at: https://www.nexteinstein.org/. (Accessed: 25th September 2019)
12. R: What is R? Available at: https://www.r-project.org/about.html. (Accessed: 25th September 2019)
13. Bioconductor - Home. Available at: https://www.bioconductor.org/. (Accessed: 25th September 2019)
14. Jalal, H. *et al.* An Overview of R in Health Decision Sciences. *Med. Decis. Mak. Int. J. Soc. Med. Decis. Mak.* **37**, 735–746 (2017).
15. Linux Servers. (2019). Available at: https://www.ibm.com/uk-en/it-infrastructure/servers/linux. (Accessed: 25th September 2019)
16. Linux on Azure | Microsoft Azure. Available at: https://azure.microsoft.com/en-gb/overview/linux-on-azure/. (Accessed: 25th September 2019)
17. SQL Tutorial: Learn SQL For Free. *Codecademy* Available at: https://www.codecademy.com/learn/learn-sql. (Accessed: 25th September 2019)
18. Intro to SQL: Querying and managing data. *Khan Academy* Available at: https://www.khanacademy.org/computing/computer-programming/sql. (Accessed: 25th September 2019)
19. Harvey, L. A. REDCap: web-based software for all types of data storage and collection. *Spinal Cord* **56**, 625–625 (2018).
20. Patridge, E. F. & Bardyn, T. P. Research Electronic Data Capture (REDCap). *J. Med. Libr. Assoc. JMLA* **106**, 142–144 (2018).
21. REDCap.
22. Introduction to Computer Science and Programming | Coursera. Available at: https://www.coursera.org/specializations/introduction-computer-science-programming?utm_source=gg&utm_medium=sem&utm_content=Deg-04-ComputerScience&Programming-UoL-US/UK&campaig

nid=2042754710&adgroupid=74983312929&device=c&keyword=computer%20programming%20online%20courses&matchtype=b&network=g&devicemodel=&adpostion=1t1&creativeid=357404780323&hide_mobile_promo&gclid=Cj0KCQjwoKzsBRC5ARIsAITcwXG_vXw4K1Sqmm-8jr1OI7aALL-hfMN4YZogGW2O8Ej-kBIxzyoTxUwaAoTKEALw_wcB. (Accessed: 25th September 2019)

23. Khan Academy. *Khan Academy* Available at: http://www.khanacademy.org. (Accessed: 25th September 2019)
24. edX. *edX* Available at: https://www.edx.org/. (Accessed: 25th September 2019)
25. FutureLearn. Free online courses from Top Universities. *FutureLearn* Available at: https://www.futurelearn.com/. (Accessed: 25th September 2019)
26. Training | European Bioinformatics Institute. Available at: https://www.ebi.ac.uk/training. (Accessed: 25th September 2019)
27. Swiss Epidemiology Winter School | Institute of Social and Preventive Medicine Bern.
28. Summer School in Bioinformatics. *Wellcome Genome Campus Advanced Courses and Scientific Conferences* Available at: https://coursesandconferences.wellcomegenomecampus.org/our-events/bioinformatics2019/. (Accessed: 25th September 2019)
29. Home. *The Mandela Rhodes Foundation* Available at: https://mandelarhodes.org/. (Accessed: 25th September 2019)
30. MINDS – Mandela Institute for Development Studies.
31. Home. *H3ABioNet* Available at: https://www.h3abionet.org/. (Accessed: 25th September 2019)
32. https://plus.google.com/u/0/+Datacamp. Learn R, Python & Data Science Online. Available at: www.datacamp.com/. (Accessed: 25th September 2019)
33. Learn to Code - for Free. *Codecademy* Available at: https://www.codecademy.com/. (Accessed: 25th September 2019)
34. Data Management for Clinical Research | Coursera. Available at: https://www.coursera.org/learn/clinical-data-management. (Accessed: 25th September 2019)
35. Video Conferencing Software | Web Conference | GoToMeeting. Available at: https://www.GoToMeeting.com/en-gb. (Accessed: 25th September 2019)
36. University, C. M. Design & Teach a Course - Eberly Center - Carnegie Mellon University. Available at: https://www.cmu.edu/teaching/designteach/index.html. (Accessed: 25th September 2019)
37. FutureLearn. The pedagogy of blended learning. *FutureLearn* Available at: https://www.futurelearn.com/courses/blended-learning-getting-started/0/steps/7848. (Accessed: 25th September 2019)
38. 12 Different Types of Blended Learning. *TeachThought* (2019).
39. Garrison, D. R. & Kanuka, H. Blended learning: Uncovering its transformative potential in higher education. *Internet High. Educ.* **7**, 95–105 (2004).
40. Vula : Gateway : Welcome. Available at: https://vula.uct.ac.za/portal. (Accessed: 25th September 2019)
41. Sakai Learning Management System | Higher Education. *Sakai LMS* Available at: https://www.sakailms.org. (Accessed: 25th September 2019)
42. R Markdown. Available at: https://rmarkdown.rstudio.com/. (Accessed: 25th September 2019)
43. Oracle VM VirtualBox. Available at: https://www.virtualbox.org/. (Accessed: 25th September 2019)
44. MySQL. Available at: https://www.mysql.com/. (Accessed: 25th September 2019)
45. Kramer, F. & Beißbarth, T. Working with Ontologies. *Methods Mol. Biol. Clifton NJ* **1525**, 123–135 (2017).
46. Rubin, D. L., Shah, N. H. & Noy, N. F. Biomedical ontologies: a functional perspective. *Brief. Bioinform.* **9**, 75–90 (2008).
47. Feder, S. L. Data Quality in Electronic Health Records Research: Quality Domains and Assessment Methods. *West. J. Nurs. Res.* **40**, 753–766 (2018).
48. Boeckhout, M., Zielhuis, G. A. & Bredenoord, A. L. The FAIR guiding principles for data stewardship: fair enough? *Eur. J. Hum. Genet. EJHG* **26**, 931–936 (2018).
49. Mulder, N. *et al.* Proceedings of a Sickle Cell Disease Ontology workshop - Towards the first comprehensive ontology for Sickle Cell Disease. *Appl. Transl. Genomics* **9**, 23–29 (2016).
50. Tidyverse. Available at: https://www.tidyverse.org/. (Accessed: 25th September 2019)
51. Getting started with the pwr package. Available at: https://cran.r-project.org/web/packages/pwr/vignettes/pwr-vignette.html. (Accessed: 25th September

2019)

52. Wilson, G. Ten quick tips for creating an effective lesson. *PLOS Comput. Biol.* **15**, 1–12 (2019).
53. R Online Courses | LinkedIn Learning, formerly Lynda.com. Available at: https://www.linkedin.com/learning/topics/r. (Accessed: 25th September 2019)
54. The Modern Learning Path: Building a Flow in Digital Learning. *Training Industry*

F | Appendix

Included in this appendix are:

1. A full list of planned course sessions for both online sessions and the workshop (those that did not take place are marked)
2. Examples of the summative assignment questions
3. Data quality survey questions
4. Full evaluation form questions and responses

Online sessions

Day 1 | March 29th

Session	**Description**
Welcome address	Introduction of members to the course, an other members (webinar)
Introduction to Big Data in Epidemiological studies	Overview of Big Data methods (webinar)
Data Manipulation with Linux	Reading material, quiz and assignments (tutorial)

Day 2 | May 3rd

Session	**Description**
Preparing data for analysis	Overview of data types, cleaning methods, and basic summary statistics (webinar)
Statistical methods for Big Data	Mathematical theory of probability and statistics: including sample spaces and distributions (webinar)
Introduction to R	Introduction to R programming, including data types, variables, functions, and reading and plotting data (webinar, tutorial)

Day 3 | May 31st

Session	Description
Searching, Summarising, and Assessing Epidemiology Publications (CANCELLED)	Overview of useful tools and techniques for keeping on top of the scientific literature (webinar)
Multi-site data elements consolidation: A Horror story	Introduction to the SADaCC standard data elements, retrospective data set integration case study (webinar)
Statistical Power and Sample Size Estimation	Introduction to the R package pwr, and computing statistical power (tutorial)

Workshop program

Helpful tags:

talk	this session unit is a presentation/talk/lecture
tutorial	this session unit is interactive, but involves answering simple questions. Does not use participant's own project data
project	this session unit uses participant's own data. It is a piece of the overall project that students work on during the workshop.

Session abbreviations

W - Welcome	T - Tidyverse
LS - Linux Servers	DH - Data Harmonization
D - Databasing	SD - Study Design
I - Interoperability	ML - Machine Learning
DQ - Data Quality	P - Power
CD - Cleaning Data	E - Ethics

Monday

Time	Session	Speaker/Facilitator
8.00 - 8:30	Registration	
8.30 - 9:00	Hardware/software diagnostics	
9.00 - 9.15	Introductions (W) talk	Jack Morrice
9.15 - 9.45	Recap of online course (W) talk	Gaston Mazandu
9.45 - 10:30	Description of Course Structure, Resources and Activities (W) talk	Jack Morrice
10.30 - 11.00	Coffee	
11.00 - 11.30	Big Data in Genetic Epidemiology (LS) talk	Emile Chimusa

11.30 - 11.50	Introduction to a Linux Server and PBS (LS) talk	Christian Bope
11.50 - 12.20	CHPC Linux server (LS) tutorial	Lead: Jack Morrice Assist: Christian Bope
12.20 - 12.30	Accessing project data sets (LS) project (CANCELLED)	Lead: Jack Morrice
12.30 - 13.30	Lunch	
13.30 - 14.00	Introduction to Databasing (D) talk	Kenneth Babu
14.00 - 14.30	Construct SQL queries of existing databases (D) tutorial	Lead: Kenneth Babu Assist: Mario, Arthemon
14.30 - 15.00	Creating relational databases (D) project	Lead: Kenneth Babu Assist: Mario, Arthemon
15:00 - 15:30	Tea	
15.30 - 16.00	Data Repository Interoperability between MySQL and REDCAP Systems in Muhimbili Sickle Cell Cohort (I) talk (CANCELLED)	Raphael Sangeda
16.00 - 17.00	Data Repository Interoperability (I) tutorial (CANCELLED)	Lead: Raphael Assist: Jack Morrice

Tuesday

Time	Session	Speaker/Facilitator
8.30 - 8.50	Morning exRcise - I/O with files	Jack Morrice
8.50 - 9.00	Summary + feedback: Monday	Jack Morrice
9.00 - 9.30	Proposed SickleInAfrica Data Quality Assurance Framework (DQ) talk (CANCELLED)	Vicky Nembaware, Gaston Mazandu, Raphael Sangeda
9.30 - 10.00	SickleInAfrica Standard Operating Procedures (DQ) talk	Annemie Stewart

10.00 - 10.30	Data quality assurance methods and processes (DQ) talk	Arthemon Nguweneza
10.30 - 11.00	Coffee	
11.00 - 11.30	Cleaning & filtering using REDCap - demo (CD) talk	Arthemon Nguweneza
11.30 - 12:30	Cleaning & filtering using REDCap (CD) tutorial	Lead: Arthemon Nguweneza Assist: Annemie Stewart, Mario Jonas
12.30 - 13:30	Lunch	
13.30 - 14.00	Tidyverse (T) talk	Jack Morrice
14.00 - 14.30	Tidyverse (T) tutorial	Lead: Jack Morrice Assist: Gaston Mazandu
14.30 - 15.00	Data Summary (T) project (CANCELLED)	Jack Morrice
15:00 - 15:30	Tea	
15.30 - 15.40	Description of SickleInAfrica Standardized Data Elements (DH) talk	Mario Jonas
15.40 - 16.10	Mapping scripts (DH) tutorial (CANCELLED)	Lead: Gaston Mazandu Assist: Annemie Stewart
16:15 - 17.00	Data harmonization, ontologies and FAIR data (DH) talk	Nicola Mulder

Wednesday

Time	Session	Trainer/Facilitator
8.30 - 8.50	Morning exRcise - installing packages, introducing bioconductor	Jack Morrice
8.50 - 9.00	Summary + feedback: Tuesday	Jack Morrice

9.00 - 9.30	Study design (SD) talk	Arthemon Nguweneza
9.30 - 10.00	Overview of machine learning methods in Big Data epidemiology (ML) talk	Mamana Mbiyavanga
10.00 - 10.30	Linear Regression (ML) tutorial	Lead: Mamana Mbiyavanga Assist: Jack Morrice
10.30 - 11.00	Coffee	
11.00 - 11.30	Principal Component Analysis (ML) tutorial	Lead: Mamana Mbiyavanga Assist: Jack Morrice
11.30 - 12.30	Machine Learning methods & project data (ML) project	Lead: Mamana Mbiyavanga
12.30 - 13.30	Lunch	
13.30 - 14.00	Statistical power recap (P) talk	Gaston Mazandu
14.00 - 14.30	Example power calculations (P) tutorial	Lead: Gaston Mazandu Assist: Jack Morrice
14.30 - 15.00	Project power calculations (P) project	Lead: Gaston Mazandu Assist: Jack Morrice
15.00 - 15.30	Tea	
15.30 - 16.00	Designing Big Data epidemiological studies (SD) tutorial	Lead: Arthemon Nguweneza Assist: Mario Jonas
16.00 - 17.00	Draft project study design (SD) project	Lead: Arthemon Nguweneza Assist: Mario Jonas

Thursday

Time	Session	Trainer/Facilitator
9.00 - 9.10	Summary + feedback: Wednesday	

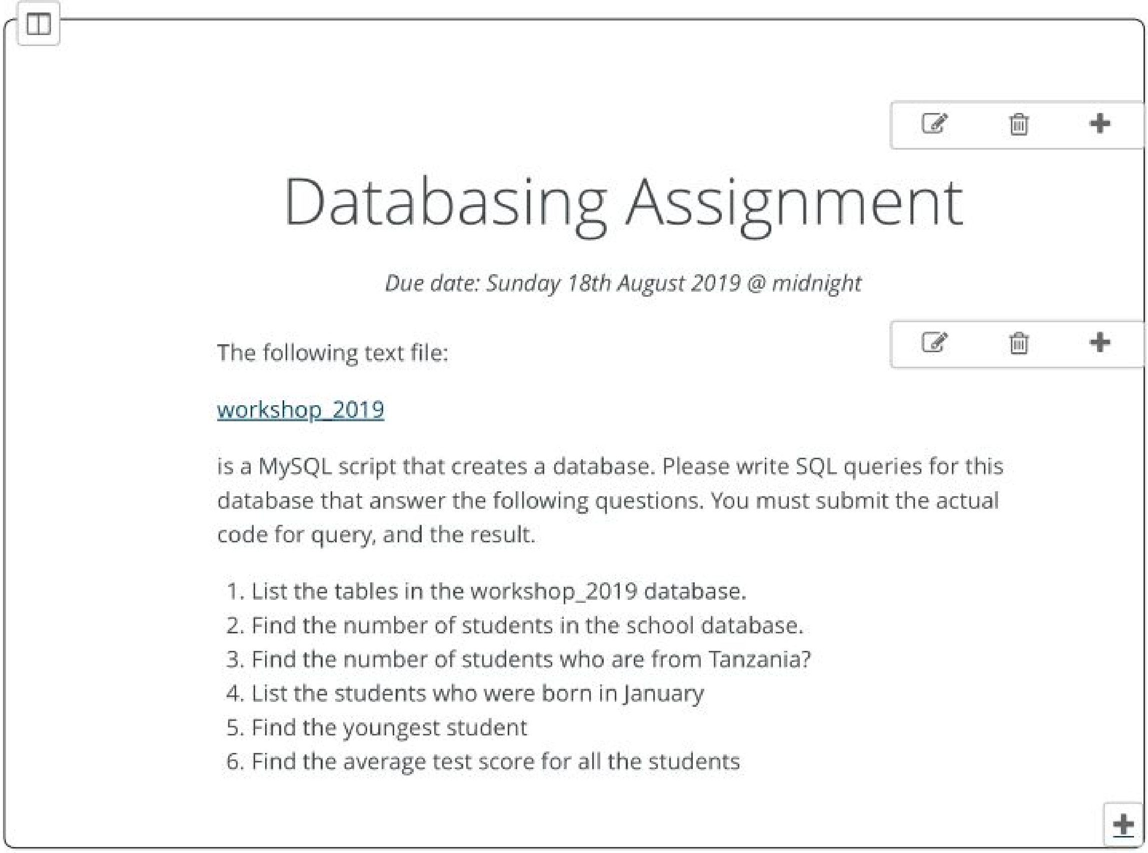
9.10 - 10.30	R, tidyverse, and the CHPC server	Jack Morrice
10.30 - 11.00	Coffee	
11.00 - 12.30	Project proposal work	Jack & Arthemon
12.30 - 13.30	Lunch	
13.30 - 14.15	Multi-site study ethics talk (CANCELLED) - ethical guidelines for designing multi site retrospective studies - clear accessible ethical clearance approval as aspect of data quality	Nchangwi Syntia Munung
14.15 - 15.00	Project ethical clearance proposals tutorial (CANCELLED) - Participants start to draft their own clearances, based on the types of data they have, and the types of research questions they are now considering	Nchangwi Syntia Munung
15.00 - 15.30	Tea	

Friday

Time	Session	Trainer/Facilitator
9.00 - 9.10	Summary + feedback: Thursday	
9.10 - 10.30	Python, from the beginning	Gaston, Jack Assist
10.30 - 11.00	Coffee	
11.00 - 12.30	Python, from the beginning	Gaston, Jack Assist
12.30 - 13.30	Lunch	
13.30 - 14.00	Effective presentations talk (CANCELLED)	Jack

14.00 - 15.00	Writing project presentations (CANCELLED)	Arthemon and Jack
15.00 - 15.30	Tea	
15.30 - 17.00	Groups present their projects project	Entire team

Summative assignment examples



Databasing Assignment

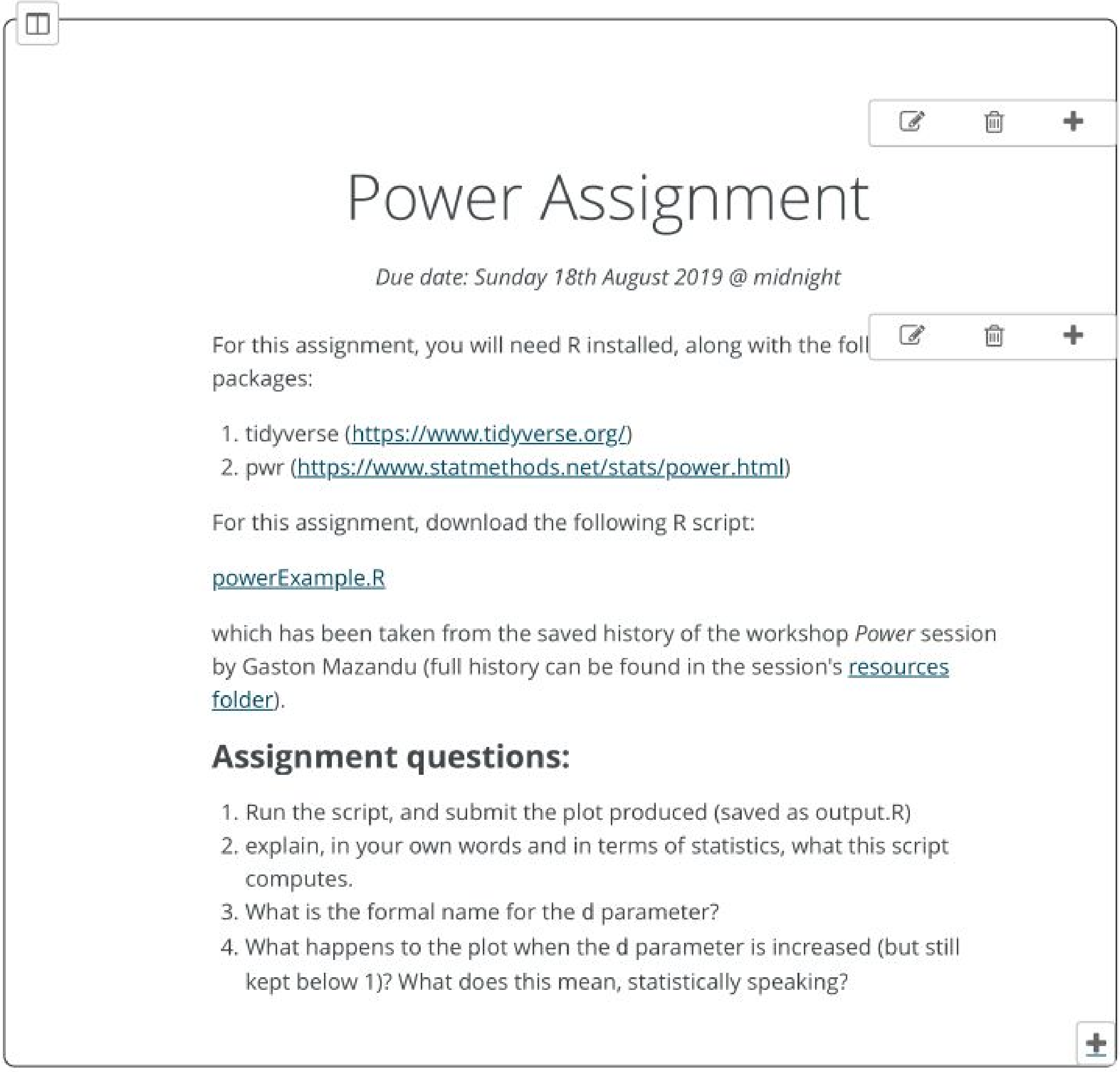
Due date: Sunday 18th August 2019 @ midnight

The following text file:

workshop_2019

is a MySQL script that creates a database. Please write SQL queries for this database that answer the following questions. You must submit the actual code for query, and the result.

1. List the tables in the workshop_2019 database.
2. Find the number of students in the school database.
3. Find the number of students who are from Tanzania?
4. List the students who were born in January
5. Find the youngest student
6. Find the average test score for all the students



Power Assignment

Due date: Sunday 18th August 2019 @ midnight

For this assignment, you will need R installed, along with the foll packages:

1. tidyverse (https://www.tidyverse.org/)
2. pwr (https://www.statmethods.net/stats/power.html)

For this assignment, download the following R script:

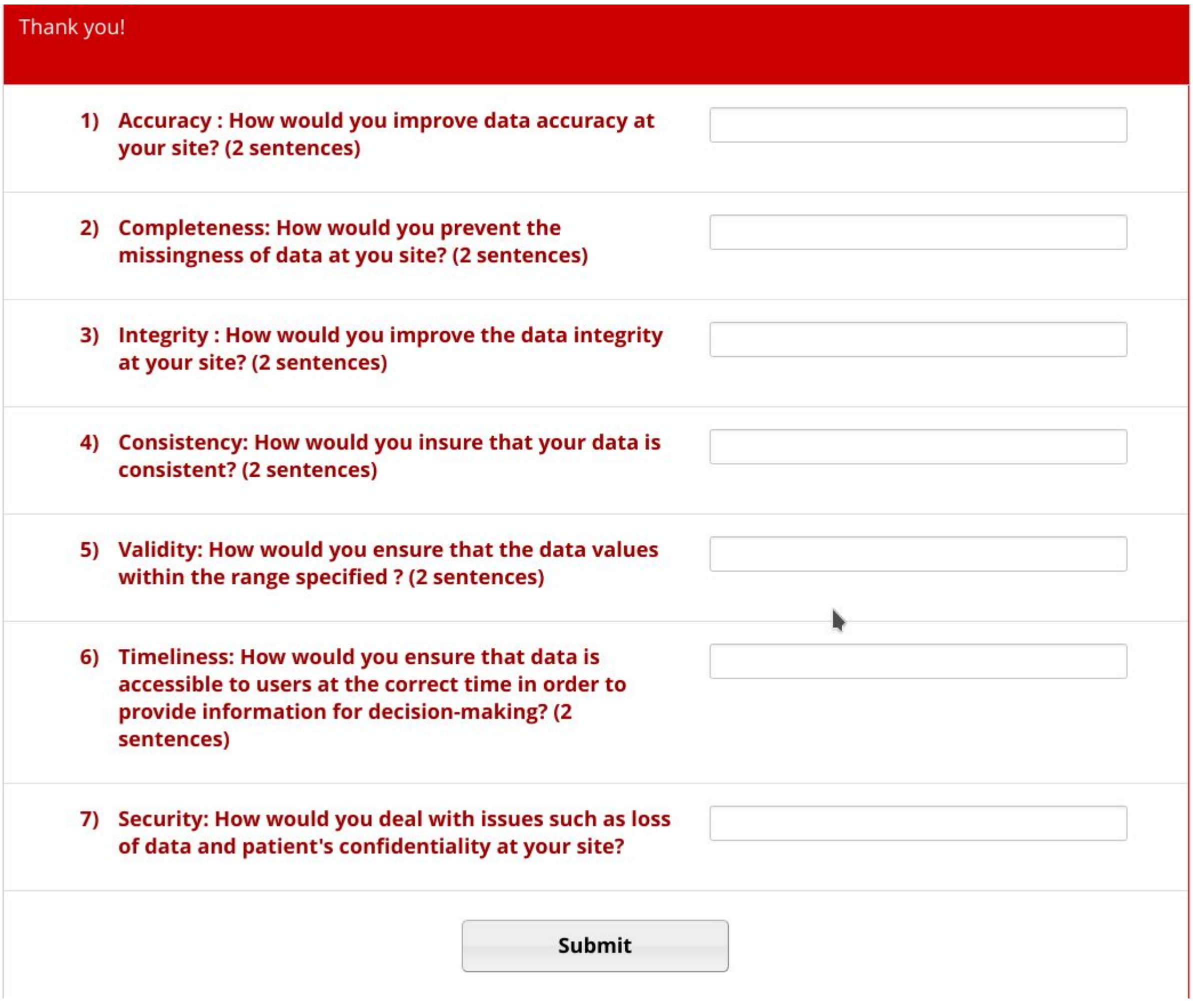
powerExample.R

which has been taken from the saved history of the workshop *Power* session by Gaston Mazandu (full history can be found in the session's resources folder).

Assignment questions:

1. Run the script, and submit the plot produced (saved as output.R)
2. explain, in your own words and in terms of statistics, what this script computes.
3. What is the formal name for the d parameter?
4. What happens to the plot when the d parameter is increased (but still kept below 1)? What does this mean, statistically speaking?

Data quality survey



Thank you!

1) Accuracy : How would you improve data accuracy at your site? (2 sentences)

2) Completeness: How would you prevent the missingness of data at you site? (2 sentences)

3) Integrity : How would you improve the data integrity at your site? (2 sentences)

4) Consistency: How would you insure that your data is consistent? (2 sentences)

5) Validity: How would you ensure that the data values within the range specified ? (2 sentences)

6) Timeliness: How would you ensure that data is accessible to users at the correct time in order to provide information for decision-making? (2 sentences)

7) Security: How would you deal with issues such as loss of data and patient's confidentiality at your site?

Big Data Analytics for multi-site epidemiology

SADACC-Big Data Analytics

Survey results

Started: 07 August 2019

Ended: 17 August 2019

Reply rate: 28% (11 / 40)

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Vula Online Evaluation System

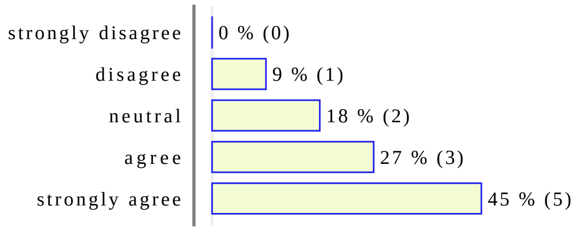
Big Data Analytics for multi-site epidemiology

Thank you for completing the Big Data short course!Please complete our feedback form. Your input will be invaluable when we improveit for the next iterations, and your effort will be greatly appreciated :)Warm regards,the SADaCC team

Course/Group Items:

Course materials

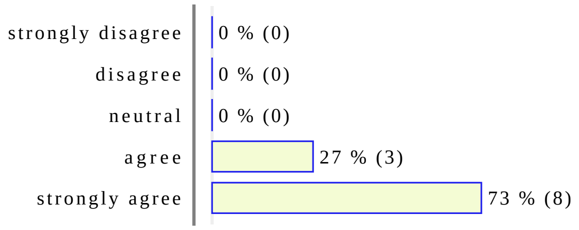
1. the online classroom service (gotomeeting) was used effectively



Number of answers: 11

Weighted mean: 4.09

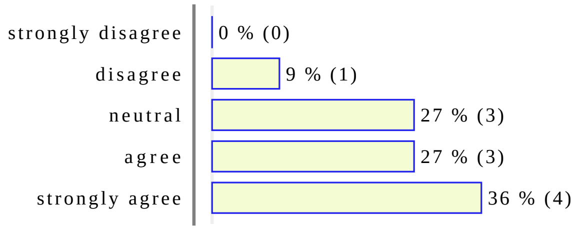
2. the course website (Vula) was easy to navigate



Number of answers: 11

Weighted mean: 4.73

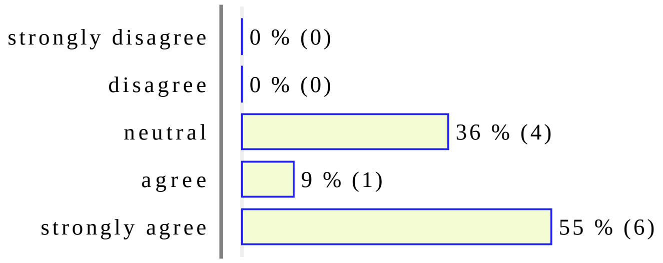
3. the assignment instructions were clear



Number of answers: 11

Weighted mean: 3.91

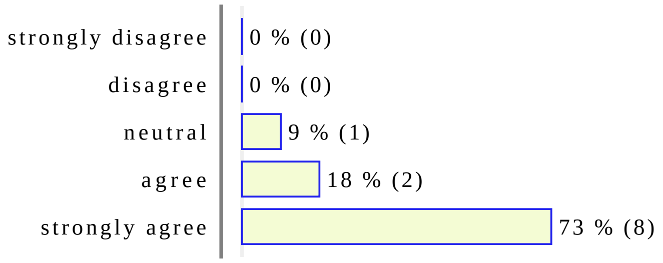
4. there were sufficient online notes and slides to complete the assignments



Number of answers: 11

Weighted mean: 4.18

5. I had access to the necessary computing resources throughout the course

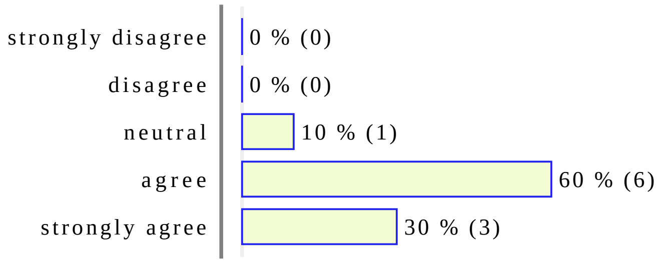


Number of answers: 11

Weighted mean: 4.64

Instructors & assistants

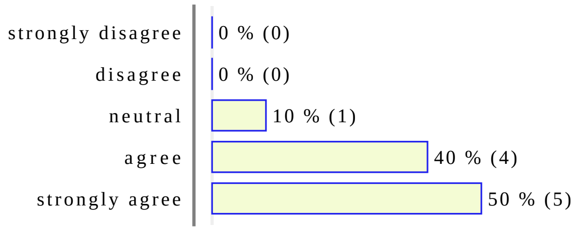
6. The instructors communicated clearly



Number of answers: 10

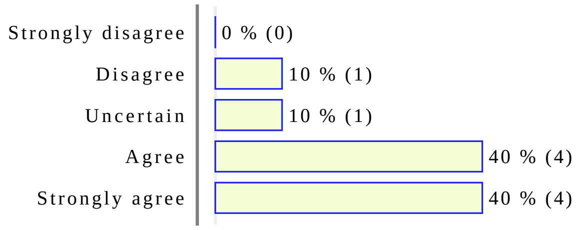
Weighted mean: 4.2

7. I found the instructors engaging



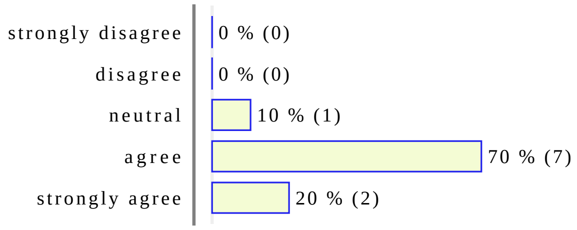
Number of answers: 10
Weighted mean: 4.4

8. The online classrooms were well prepared



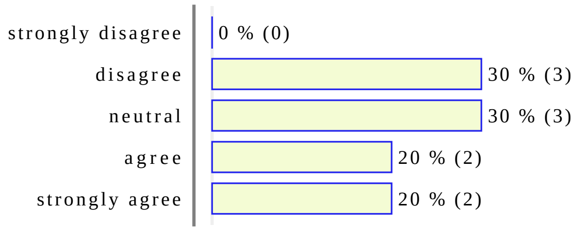
Number of answers: 10
Weighted mean: 4.1

9. I had adequate learning support during the course



Number of answers: 10
Weighted mean: 4.1

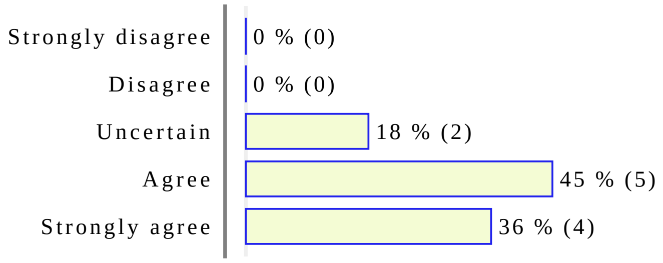
10. the instructors returned assignments and exams in a timely manner



Number of answers: 10
Weighted mean: 3.3

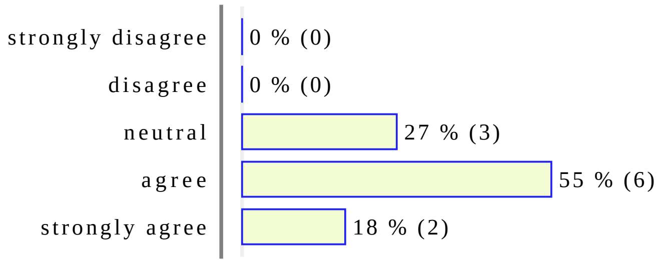
Course content

11. the course learning outcomes were clear



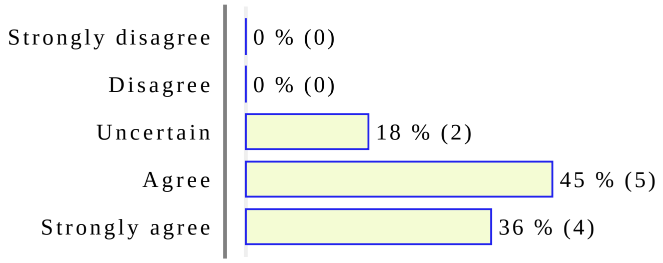
Number of answers: 11
Weighted mean: 4.18

12. the course helped me complete the advertisedlearning outcomes



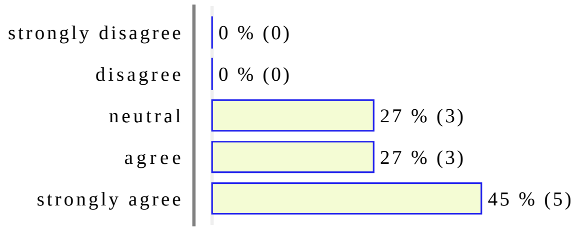
Number of answers: 11
Weighted mean: 3.91

13. the course content was aligned with my personal learning goals



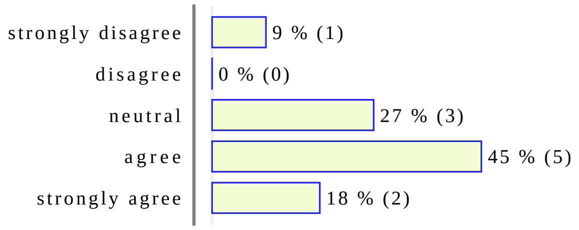
Number of answers: 11
Weighted mean: 4.18

14. the online content and workshop content complemented each other



Number of answers: 11
Weighted mean: 4.18

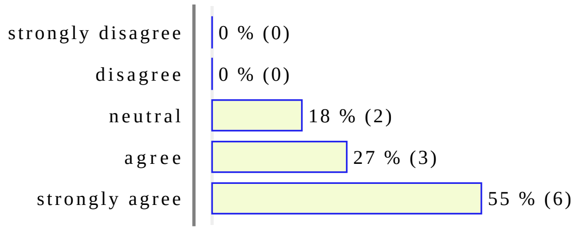
15. the lessons were at the right level of difficulty for me



Number of answers: 11
Weighted mean: 3.64

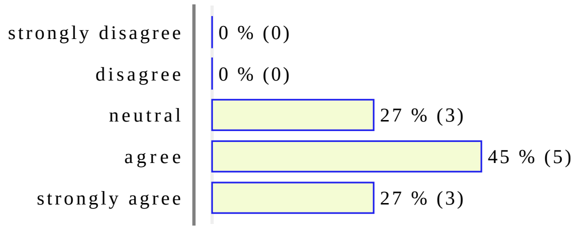
Your experience

16. I attended lessons regularly



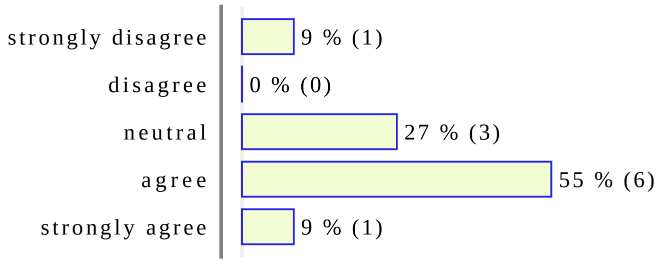
Number of answers: 11
Weighted mean: 4.36

17. I was able to manage the workload well



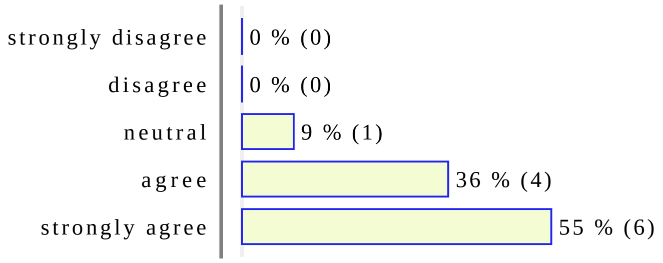
Number of answers: 11
Weighted mean: 4

18. I enjoyed the online component of the course



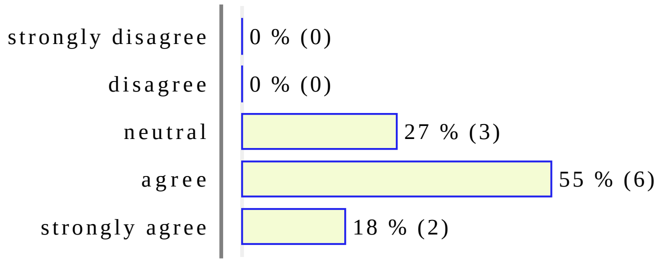
Number of answers: 11
Weighted mean: 3.55

19. I enjoyed the face-to-face workshop



Number of answers: 11
Weighted mean: 4.45

20. I feel more confident working with Big Data methods now



Number of answers: 11
Weighted mean: 3.91

Open-ended questions

21. I would like to have had more of:

- Practical face to face session with with some theory on specif data analysis(More of Bioinformatic).
- Hands on exercise with the programming languages
- when will you organize the next session?
- data basing, python and NGS
- Data quality and statistical analysis
- more of the NGS training. The expected results or output of the NGS data analysis was not clear.
- The NGS pipeline tutorials
- The same workshop again
- Python lessons

Number of answers : 9

22. I would like to have had less of:

- online session
- NA
- training
- N/A
- nono
- None
- -

Number of answers : 7

23. Any additional comments

- I want to emphasize that giving enough time for the workshop and it would be better if it is more specific ,so that the participants will be more specialized in the area.
- The workshop should be extended to cover at least one month.
- no comment
- I wish, put together, it would have been at least a month long face to face session for both multi site epidemiology and Next Generation Sequencing.
- none
- The training was good. I can now appreciate Big Data and acquired the skills of handling Big Data.

Number of answers : 6