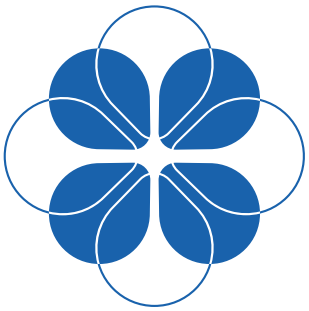

PRAKSISNETT

Data management plan

Version 1.3
Authors: Johan Gustav Bellika / Snow team
Date: 2018.11.29

Revision history:

Issue	Details	Who	Date
1.3	Revised the plan to reflect changes necessary to support GDPR.	JGB	2018.11.29
1.2	Minor changes based on feedback from Professor Guri Rørtveit.	JGB	2018.06.14
1.1	Revised the plan to reflect the policy decisions made by the PraksisNett management board.	JGB	2018.06.12
1.0	Revised the plan to reflect requirements from GDPR. Added a background section.	JGB	2018.06.05
0.9	Added TODO table. Added description of the storage and backup section. Added description of selection and preservation. Added description of data sharing.	JGB	2018.03.13
0.1	Initial version	JGB	2017.09.05

1 Introduction

1.1 Summary

This document describes how data will be handled in the PraksisNett research infrastructure. This is a dynamic document that will need to be updated regularly as details and agreements on how to handle research data are clarified. The latest version of the data management plan will always be available on www.praksisnett.no.

The document is based on a checklist for data management plans[1], provided by the Digital Curation Centre in UK.

The current version of the Data Management Plan only consider GP office solutions with one EHR server. Cloud based EHR solutions will be covered in a future revision of this document.

1.2 Audience

Who are the readers for this document?
This document is written assuming an audience of clinical research informatics officers (CRIOs) [2], decision makers, researchers and general practitioners with special interest in how data is managed in the PraksisNett research infrastructure.

What background knowledge about the research infrastructure do we assume?
This document assumes minimal knowledge about electronic health records (EHR) data reuse and its challenges, and the data reuse infrastructure. However, interested readers may find detailed description about the background and rationale for the proposed data management in the list of references provided at the end of the document.

1.3 Background

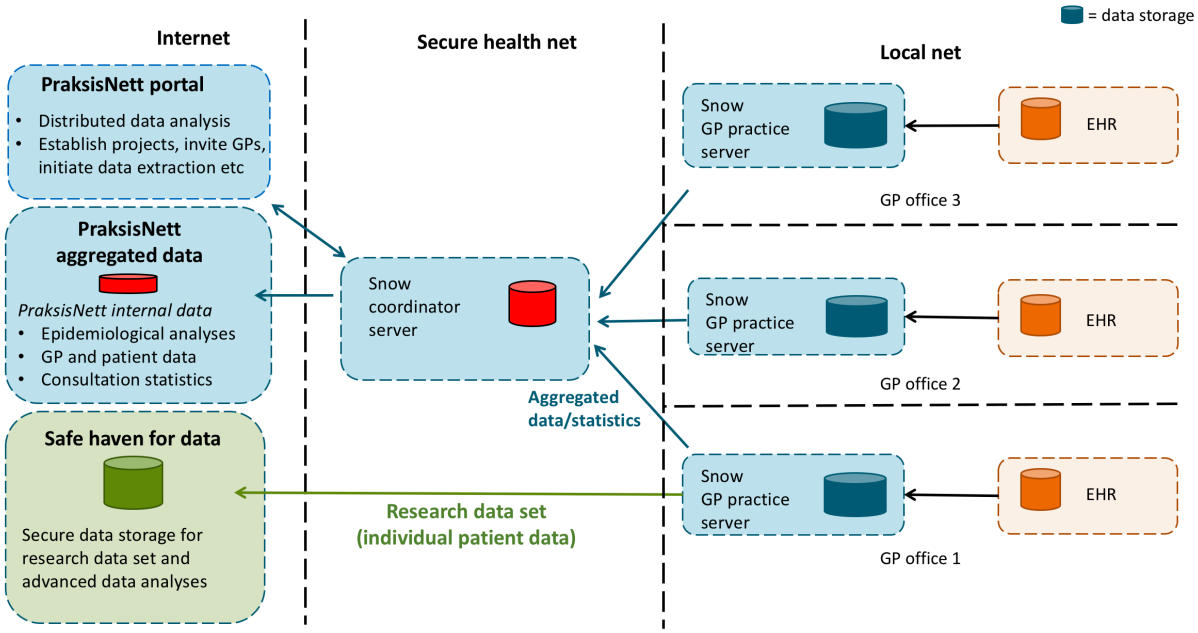

Figure 1. Dataflow in PraksisNett IT infrastructure

The PraksisNett IT infrastructure contains three types of data resources. That is:

1. Research datasets stored inside GP practice servers (blue barrels in Figure 1)
2. Aggregated and anonymous data generated from all participating GP practices (red barrels in Figure 1)
3. Complete research datasets stored inside safe havens (green barrels in Figure 1).

The orange barrels are the EHR database used daily by general practitioners in GP practices to document patient treatment. Before data can be extracted and stored on the Snow GP practice server (in the blue barrels), the GP need to consent to usage of the EHR data. When data is stored on the GP practice server, patients can be identified as potential research subject candidates. Data about both patients and health workers stored in the blue barrels are pseudonymized when leaving the EHR server (black arrows). Pseudonymized research datasets stored inside GP practice servers can only be accessed by personnel with access to the Snow GP practice server. Patients can only be re-identified by the general practitioner or by personnel authorized by the general practitioner, from inside the GP practice.

Based on data in the blue barrels, patient data is first aggregated locally at the GP practice and then aggregated across all participating GP practices. This data is stored inside the red barrel, the report database, at the Snow coordinator server. From there aggregated data can be downloaded using a web service interface to the report database. The PraksisNett aggregated data is a subset of the data stored in the snow coordinator server report database. The flow of aggregated data is shown as blue arrows in Figure 1.

The distributed data analysis client, available from the PraksisNett portal, use the data inside the Snow GP practice servers to produce aggregated data, following the procedure explained above. All results from this processing is stored inside the report database at the snow coordinator server (red barrel) and is available in the distributed data analysis client user interface.

When data collection in a research project is initiated at the GP practice servers, the datasets containing individual patient data will be transferred to the green safe haven. Only data about consenting patients will be transferred. This is shown as the green arrow in Figure 1. The researcher will get access to the complete datasets for advanced data analysis inside the safe haven.

2 Data Collection

What data will you collect or create?
EHR data (orange barrels in Figure 1) from consultations between patients and physicians in general practice offices.

What type, format and volume of data?
All kinds of data that are created and stored in electronic health record systems.
The format of extracted data will be published in reports and made available on the web at www.praksisnett.no.
All data stored on the Snow GP practice server will be pseudonymized.
Extracted data will be available in csv, xml or json format.

Do your chosen formats and software enable sharing and long-term access to the data?
The format of the data will enable local and distributed use of data for computations while stored in the Snow GP practice server. The data will be transferred to safe havens for advanced statistical analysis.

The first version of the infrastructure will use a proprietary data format for storage of the EHR data on the Snow GP practice server. The data will be stored in a traditional database. The long term goal is to store EHR data in OpenEHR archetype format.

Are there any existing data that you can reuse?
Data stored on the Snow GP practice server will reuse existing EHR data and data collected as part of the GP practice participating in research studies (orange barrels in Figure 1).

How will the data be collected or created?
Data will be extracted from electronic health record (EHR) systems in primary care and through electronic questionnaires.

What standards or methodologies will you use?
The long term goal is to store EHR data in OpenEHR archetype format. For systems based on this format, no transformation of EHR data will be necessary. For systems not using OpenEHR we will transform extracted EHR data to OpenEHR archetype format. This transformation aims at creating generic software for analysis of any EHR data, and standardization of EHR data in primary care.

What quality assurance processes will you adopt?
We aim at leveraging data quality assessment using standardized data quality checks and giving the clinicians in the GP practice the possibility to fix data quality by modifying data on the Snow GP practice server and at the source of the data, i.e. in the EHR system, if possible.

The data quality checks will be implemented as generic rules using the OpenEHR guideline definition language (GDL). Aggregated statistics about the data quality of EHR data elements will provide guidance on what data elements that has sufficient data quality for research purposes.

3 Documentation and Metadata

What documentation and metadata will accompany the data?
When the data is in OpenEHR archetype format all documentation and metadata will be available in the Norwegian clinical Knowledge Manager at Arketyper.no [3]. Research data will also be described using the NESSTAR[4] / DDI [5] data format.

The format of available aggregated data (the red barrels in Figure 1) will be described in system documentation available as part of the report database application programming interface (API) description on www.praksisnett.no.

4 Ethics and Legal Compliance

How will you manage any ethical issues?
There are many ethical and especially privacy issues involved in using medical record data for research. The need for access to medical record data spans the whole research cycle, from generating the research idea, identifying the patient cohort, getting consent for participation, extracting the datasets and finally analysis and publishing of research results.

Idea generation and feasibility assessment: In this phase, pseudonymous data (the blue barrels) extracted daily from electronic health record (EHR) systems (the orange barrels), stored inside each primary care office in the Snow GP practice server, will be used to explore potential study cohorts using distributed privacy preserving statistical processing tools. Using these tools, the researcher or the PraksisNett staff can perform study feasibility assessment before or after specifying the research protocol.

The benefit of this approach is that no patient identifiable data leaves the general practice during the study specification phase. The GP office (and the GPs) are the controllers (behandlingsansvarlig) of the Snow GP practice server, according to Norwegian law and GDPR, and decides for what purposes EHR data can be processed.

Project approval by GP´s: When the research project is approved by the PraksisNett board, it needs the consent from each participating GP to get access to information about the patient the GP has the responsibility for. When this approval is registered, pseudonymized data about the patients can be extracted from the EHR (the orange barrels) to the Snow GP practice server (the blue barrels). In this process, the data is pseudonymized. That is, all identifiers, both patients and physicians are replaced by pseudonyms. The only place where the pseudonyms can be re-identified to the real patient ID, is at the GP practice, inside the GP practice EHR system (shown as the orange barrels in Figure 1).

Obtaining consent from patient: After approval by the GP the potential cohort is identified. Now the GP need to assess which patients that should be invited to the study. The GP then need to send study invitation letters to their patients. When the patient confirm participation, using a written consent form prepared by the researcher, the patient becomes part of the study cohort. The non-consenting patients now need to be removed from the study cohort using the IT tool available from the local Snow GP practice server.

Extracting the research dataset: When patient consent has been registered in the local research dataset at the GP practice server (the blue barrels), the PraksisNett staff may initiate transfer of research data directly to a safe haven. All local datasets with data from consenting patients will then be encrypted and transferred daily to a safe haven. In the safe haven all partial datasets may be merged together to form a complete research dataset (the green barrel). This data is the responsibility of the researcher and the responsible research institution. As the safe haven is a secure environment, the research dataset stays in the safe haven environment until deleted. Advanced statistical analysis can be performed on the research dataset within the safe haven environment.

Have you gained consent for data preservation and sharing?
GPs approve use of their patients' pseudonymized data for study feasibility assessment and preliminary study cohort identification. Patients provide written consent for study participation, based on the invitation letter from their GPs for extraction and use of their data for research purpose.

According to GDPR, patients must be informed about what their data is used for. Therefore, each GP practice participating in the research network must inform their patients that their EHR data is used to identify them as potential research subjects. Patients have the right to exempt from such usage of their EHR data. Patients that want to exempt from being invited to research studies can contact their GP. The GP will then ensure that the patient is not invited to research, using the local PraksisNett software.

How will you protect the identity of participants if required?
The identity of all patients and physicians will be removed from the datasets stored in the Snow GP practice server and replaced by pseudonyms. The pseudonyms of patients and physicians can only be mapped back to real identities in the EHR system at the GP practice.

How will sensitive data be handled to ensure it is stored and transferred securely?
Sensitive data will be kept at the GP practice until transferred to a safe haven. During transfer the data will be encrypted in a way that makes it impossible for any entity other than the safe haven to decrypt the data.

How will you manage copyright and Intellectual Property Rights (IPR) issues?
As the underlying technology have been developed inside UiT and UNN, the usual procedures for inventions have been and will be used. The consortia agreement specifies how IPR established by the project are handled.

Who owns the data?
The are several data owners involved. The data in the GP practice server is owned by the GPs in each office. The patients have several rights in relation to use of this data, according to GDPR and Norwegian law.

The aggregated data (the red barrels), generated though distributed computations on the local GP practice data, will be owned by the PraksisNett consortia. The data generated in each research project is owned by the responsible research institution. The research datasets extracted from the Snow GP practice server and saved within safe havens (green barrel) is owned by the responsible research institution for the research project.

How will the data be licensed for reuse?
It is the intention of the PraksisNett consortia to allow open and free access to the aggregated data generated by the Snow/Emnet system. However, open access can only be allowed as long as no patient, health personnel or health institutions can be exposed through the aggregated data. We will therefor evaluate both the methods and the data generated before free access to the data and the tools is permitted.

Access to aggregated data will be made freely available to the PraksisNett partners through the report database interface. It is the intention that the aggregated data (the red barrels in Figure 1), can be used without a specific license.

The responsible research institutions own the research projects and the research datasets (the green barrels). These institutions decides how the research datasets can be reused. Reuse of data in the research datasets for other purposes than the research project, need to be approved by each study participant (patient) before data collection, according to GDPR.

Will data sharing be postponed / restricted e.g. to publish or seek patents?
As the data (in the green barrel) is owned by each research institution, it is up to the study to decide how data should be handled after arriving in the safe haven. However, patients and GPs need to approve processing of data collected as part of a study performed within PraksisNett, before extraction of data from the GP practice. This should be done as part of data owner approval and the process of collecting patient consent for participation.

5 Storage and Backup

How will the data be stored and backed up during the research?
At this stage, it is too early to be specific about how research datasets may be backed up and restored. A dedicated specification and implementation for this functionality will be available before becoming operational. A risk assessment will be performed on the selected approach, producing a risk assessment report.

Research datasets will be stored on the Snow GP practice server until they are transferred to a safe haven. Backup of research datasets can be done using several methods:

1. While the dataset is stored on the Snow GP practice server backups can be stored on dedicated backup media on each Snow GP practice server (external USB sticks or dedicated storage media on the GP practice server). This alternative do not provide a reliable backup solution for some types of threats (physical theft, fire, hacking, cryptovirus, for instance).
2. For projects using a safe haven, or alternatively for all projects, backup of the research datasets may be sent and stored in the safe haven.
3. Project datasets may be backed up to the EHR server. For mapping data managed by the data reuse component (DRC), the ordinary backup of the EHR system must be used. This alternative require additional collaboration with the system administrators of the GP practice.

After transfer of the research datasets to a safe haven, backups of the research datasets may be deleted from the backup media, as the safe haven will have the responsibility for the research dataset.

Who will be responsible for backup and recovery?
The Snow team.

How will the data be recovered in the event of an incident?
At this stage, it is too early to be specific about how research datasets may be backed up and restored. A dedicated specification and implementation for this functionality will be available before becoming operational. A risk assessment should also be performed on the selected approach, producing a risk assessment report.

Some of the data in research datasets can be regenerated by extraction of data from the EHR server. If backups are stored elsewhere, the last stored state of the research dataset may be restored.

How will you manage access and security?
According to the Snow team security policy document and “The Code of Conduct for information security in the healthcare and care services” [6].

What are the risks to data security and how will these be managed?
The risks to information security have been identified through seven risk assessments performed on the Snow system and its components. These are available from PraksisNett web pages. The risk assessments are also summarized in an appendix to the DPIA document, also available from the PraksisNett web pages. Additional risk assessment activities will be

performed when new and substantial functionality are added to the solution. This must ensure that we maintain a high level of protection against adverse events and threats.

How will you control access to keep the data secure?
Authentication before access to the Snow GP practice server and the PraksisNett tools will be provided by the Helseid(https://www.nhn.no/helseid/). System administration logon to the Snow GP practice server will be done according to the Snow team security policy document.

The research infrastructure use a dedicated authorization server to control access to the resources in the research infrastructure. Authorizations will be managed by the PraksisNett staff in the coordinating node.

How will you ensure that collaborators can access your data securely?
PraksisNett follows the European Charter for Access to Research Infrastructures [7]. Collaborators can come in many roles:

1. As users of the aggregated data (the red barrels in Figure 1) produced by the research infrastructure,
2. as users of the client for distributed privacy preserving statistical computations on the GP practice data, getting access to the data in the red barrels in Figure 1,
3. as collaborators to each individual research project (access to the data in the red barrel at the Snow coordinator server) and,
4. as collaborators when data is stored in a safe haven (the green barrel in Figure 1).

Users of aggregated data owned by PraksisNett. It is the intention of the PraksisNett consortia to allow free access to the aggregated data (the red barrels). As the data is aggregated and re-identification of patients or health workers based on the aggregated data should not be possible.

Users of the client for distributed privacy preserving statistical computations. The authentication of collaborators will be based on Helse ID (https://www.nhn.no/helseid/). The data provided by the client is aggregated and re-identification of patients or health workers based on the aggregated data should not be possible. We will adopt a stepwise implementation of access. First users will be the PraksisNett staff, which will help the researchers in getting the answers they need while planning and managing their research projects. After this step an evaluation of access policy will be performed. When the scalability of the infrastructure is known, new users will be given user credentials for access to the distributed privacy preserving statistical computation client (that provide access to the data in the red barrel at the Snow coordinator server).

Collaborators of individual projects. The last step in the deployment plan will be giving collaborators of individual projects authenticated access to the client.

Collaborators when data is stored in a safe haven. When data is stored in a safe haven (the green barrel), access to collaborators is handled by the safe haven administration.

If creating or collecting data in the field, how will you ensure its safe transfer into your main secured systems?
Data will be transformed to the appropriate data format, encrypted and transferred to the safe haven environment. Inside this environment the researcher will get access to the research dataset.

6 Selection and Preservation

Which data should be retained, shared, and/or preserved?
The Snow infrastructure can redo a data extraction from the EHRs at any time. Only questionnaire data, measurement data registered in the research dataset in the Snow GP practice server, and actions performed on the research dataset cannot be recreated. When the research project is completed, all data will be transferred to the safe haven. After the final data transfer, the research dataset will be deleted from the Snow GP practice server, except the log used to satisfy the GDPR requirements. As the data created during the research project is the property of the researcher and his organization, the policy adopted by that institution determines how the data is handled. It is important to understand that the owner of the data, the patient, must consent to usage of the data. The data cannot be used for other purposes, without new consent from the patient.

What data must be retained/destroyed for contractual, legal, or regulatory purposes?
Data managed by the data reuse component (DRC) and the processing logs belonging to each project must be retained in each Snow GP practice server to support GDPR. If a patient ask what his data have been used for, we will be able to provide a correct answer.

How will you decide what other data to keep?
Not applicable.

What are the foreseeable research uses for the data?
Data collected for research and consented to by the patient must only be used for the purpose it was collected for. After this usage (and publication of results) data should be handled according to GDPR and the Health research act (Helseforskningsloven) and the rules given by the regional ethics committee that approved the project. The normal rule is to delete the data after it has served its purpose in the research project.
Aggregated data generated ad hoc or periodically by the research infrastructure can also be used for research.

How long will the data be retained and preserved?
No specific GDPR requirements on timeframe for storing research datasets and access/processing logs exists, as far as we know. Use of data must be approved by the patient, before processing can take place. The need for long time storage of access and processing logs is therefor limited, as long as the patient is aware of the processing. Our policy is to keep the access and processing logs for 5 additional years after deletion of the research datasets.

What is the long-term preservation plan for the dataset?
According to GDPR, data should only be stored until it has served its purpose. After that it should be deleted. Research datasets may need to be stored for several years, ensuring that verification of the research results is possible.

Where e.g. in which repository or archive will the data be held?
Data will be stored in safe havens until deleted. See above.

What costs if any will your selected data repository or archive charge?
It is the intention of the PraksisNett consortium not to charge for storage of data in the research infrastructure. Storage of research datasets in safe heavens is outside our responsibility. Each safe haven may charge the individual research project for storage.

Have you estimated the cost in time and effort to prepare the data for sharing / preservation?
No, not yet. The research datasets will be transferred to safe havens. The cost for supporting this functionality will be estimated during planning meetings in the Snow team.

7 Data Sharing

How will you share the data?
The data handled in the infrastructure are the following:
Aggregated data generated directly from the deployed Snow GP practice servers (the red barrels).
Research datasets created and stored in each GP practice (the blue barrels).
Research datasets stored in safe havens (the green barrel).

It is the intention of the PraksisNett consortia to provide free access to the aggregated data generated from the deployed snow GP practice servers. After a trial period with internal use, we will evaluate whether it is possible to follow this policy. Disclosure control need to be performed before data is made available by the IT infrastructure. A specification for disclosure control will be created and provided by the PraksisNett management board.

Research datasets stored on the Snow GP practice servers. Access to analyze the research datasets stored in the Snow GP practice servers (the blue barrels) can be shared between the PraksisNett staff, the researchers and their collaborators using the distributed privacy preserving statistical processing tool client. Only aggregated statistics about the datasets (with disclosure control embedded in the IT infrastructure) will be available.

Research datasets stored in safe havens. Data stored in safe havens is outside the responsibility of our research infrastructure. The researchers need to follow the rules and regulations for sharing the data adopted by the safe havens.

How will potential users find out about your data?
The PraksisNett web portal (www.praksisnett.no) will provide information about the IT infrastructure and our data.

With whom will you share the data, and under what conditions?
It is the intention of the PraksisNett consortia to share access to aggregated data generated by the research infrastructure with anyone for free. As research datasets is own by the research institution responsible for the research projects, each project is free to decide how their datasets can be shared. However, according to GDPR patients must be informed, before collection of the data is performed, about who will get access to their data, and for what purposes.

Will you share data via a repository, handle requests directly or use another mechanism?
Aggregated data generated from the deployed Snow GP practice servers will be stored in a report database and made available through a web service interface. From there the aggregated data can be downloaded according to the PraksisNett policy.

When will you make the data available?
Aggregated data will be updated regularly (daily, weekly, monthly, depending on the PraksisNett policy) and will be available for download at any time.

Will you pursue getting a persistent identifier for your data?
Pseudonyms will be generated based on the specification described in [8]. Each project will have a unique pseudonymization process that will map patient identifiers (personnummer) to the same identifiers across health institutions. This enables identifying duplicate entries across health institutions and tracking patients pathways across our health service. Re-identification of patient can only be done inside the health institution that produced the pseudonym.

Are any restrictions on data sharing required?
As the research datasets contains highly sensitive health information about patients, access to the research datasets are restricted. It is the intention of the PraksisNett to provide free access to aggregated data generated from the deployed Snow GP practice servers.

What action will you take to overcome or minimise restrictions?
It is the intention of the PraksisNett to provide free access to aggregated data generated from the deployed Snow GP practice servers.

For how long do you need exclusive use of the data and why?
It is the intention of the PraksisNett to provide free access to aggregated data generated from the deployed Snow GP practice servers.

Will a data sharing agreement (or equivalent) be required?
There will be an agreement between each GP practice and UNN/NSE about duties of the data processor (UNN/ NSE), processing, management, sharing and ownership of the data.

8 Responsibilities and Resources

Who will be responsible for data management?
The Clinical Research Informatics Officer (CRIO) for the PraksisNett will be responsible for data management. The responsibility ends when data is delivered at safe havens or downloaded from the report database interfaces.

Who is responsible for implementing the DMP, and ensuring it is reviewed and revised?
The CRIO is responsible for implementing the DMP and ensuring that it is revised and reviewed.

Who will be responsible for each data management activity?
CRIO will be the leader of a team of system administrators and software developers that builds and perform system administration of the PraksisNett IT infrastructure.

How will responsibilities be split across partner sites in collaborative research projects?
Research projects that span several regional networks will be coordinated by the regional network where the responsible researcher belongs. The collaborative regional networks will be responsible for recruiting the GP offices requested by the coordinating regional network.

Will data ownership and responsibilities for research data management (RDM) be part of any consortium agreement or contract agreed between partners?
See contract between GP practice and PraksisNett for details. In short, the PraksisNett consortia owns the aggregated data (the red barrels), GP practices own data extracted from the EHR (the blue barrels). The research project owns the dataset generated by the research project that is transferred to the safe haven (the green barrel).

What resources will you require to deliver your plan?
The resources required to perform system administration is estimated to be 15 hours for installing, system and security monitoring, backup, logging, reporting, performing system administration and providing support per Snow GP practice server per year. In addition, the IT infrastructure will be built by a team of software developers.

Is additional specialist expertise (or training for existing staff) required?
The system administration and software development skills within clinical research informatics area need specialized training (3-6 months) before being productive, depending on the experience of the person.

Do you require hardware or software which is additional or exceptional to existing institutional provision?
We use standard appliance GP practice server computers (small square shaped GP practice servers) for the GP practice servers. For she Snow coordinator servers we can use both physical or virtual servers running any operating systems.

For the software system no other alternative (that we know of) exists, requiring specialized training that cannot be found outside the Snow team.

Will charges be applied by data repositories?
No. Not while being stored within the PraksisNett infrastructure (the blue barrels). For the safe haven (the green barrel) additional charges may apply. This is outside the responsibility of the PraksisNett consortia and must be paid by each individual research project.

9 References

[1] Digital Curation Centre. Checklist for a Data Management Plan., http://www.dcc.ac.uk/resources/data-management-plans/checklist (accessed 12 March 2018).

[2] Sanchez-Pinto LN, Mosa ASM, Fultz-Hollis K, et al. The Emerging Role of the Chief Research Informatics Officer in Academic Health Centers. *Appl Clin Inform* 2017; 8: 845–853.

[3] Nasjonal IKT. Clinical Knowledge Manager, http://arketyper.no/ckm/ (accessed 12 March 2018).

[4] Norwegian Centre for Research Data. Nesstar - publisher user guide version 4.0, http://www.nesstar.com/help/4.0/publisher/index.html (accessed 29 November 2018).

[5] DDI Alliance. Welcome to the Data Documentation Initiative | Data Documentation Initiative, https://www.ddialliance.org/ (accessed 30 November 2018).
[6] Direktoratet for e-helse. The Code of Conduct for information security in the healthcare and care services. *ehelse.no*, https://ehelse.no/personvern-og-informasjonssikkerhet/norm-for-informasjonssikkerhet/documents-in-english (accessed 30 November 2018).
[7] European Commission, Directorate-General for Research and Innovation. *European charter of access for research infrastructures: principles and guidelines for access and related services*. Luxembourg: Publications Office, http://bookshop.europa.eu/uri?target=EUB:NOTICE:KI0415085:EN:HTML (2016, accessed 30 November 2018).
[8] Bellika JG, Henriksen TD, Hurley JS, et al. Electronic health record data reuse infrastructure requirements - Ehealthresearch.no, https://ehealthresearch.no/prosjektrapporter/electronic-health-record-data-reuse-infrastructure-requirements (accessed 5 October 2017).