CrossMark
click for updates

SOFTWARE TOOL ARTICLE

Kvik: three-tier data exploration tools for flexible analysis of genomic data in epidemiological studies [version 1; referees: 2 approved with reservations]

Bjørn Fjukstad[1], Karina Standahl Olsen[2], Mie Jareid[2], Eiliv Lund[2], Lars Ailo Bongo[1]

[1]Department of Computer Science, UiT - The Arctic University of Norway, Tromsø, 9037, Norway
[2]Department of Community Medicine, UiT - The Arctic University of Norway, Tromsø, 9037, Norway

Abstract

Kvik is an open-source system that we developed for explorative analysis of functional genomics data from large epidemiological studies. Creating such studies requires a significant amount of time and resources. It is therefore usual to reuse the data from one study for several research projects. Often each project requires implementing new analysis code, integration with specific knowledge bases, and specific visualizations. Existing data exploration tools do not provide all the required functionality for such multi-study data exploration. We have therefore developed the Kvik framework which makes it easy to implement specialized data exploration tools for specific projects. Applications in Kvik follow the three-tier architecture commonly used in web applications, with REST interfaces between the tiers. This makes it easy to adapt the applications to new statistical analyses, metadata, and visualizations. Kvik uses R to perform on-demand data analyses when researchers explore the data. In this note, we describe how we used Kvik to develop the Kvik Pathways application to explore gene expression data from healthy women with high and low plasma ratios of essential fatty acids using biological pathway visualizations. Researchers interact with Kvik Pathways through a web application that uses the JavaScript libraries Cytoscape.js and D3. We use Docker containers to make deployment of Kvik Pathways simple.

This article is included in the Container Virtualization in Bioinformatics channel.

Open Peer Review

Referee Status: ✓ ? ?

	Invited Referees		
	1	**2**	**3**
REVISED version 2 published 16 Jun 2015	✓ report		? report
version 1 published 30 Mar 2015	? report	? report	

1 **Paul Klemm**, Otto-von-Guericke University Magdeburg Germany

2 **Zhenjun Hu**, Boston University USA

3 **Lilit Nersisyan**, National Academy of Sciences of Armenia Armenia

Discuss this article

Comments (0)

Corresponding author: Lars Ailo Bongo (larsab@cs.uit.no)

Competing interests: No competing interests were disclosed.

Introduction

Visual explorative analysis is essential to an understanding of biological functions in large-scale 'omics' datasets. However, enabling the inclusion of 'omics' data in large epidemiological studies requires collecting samples from thousands of people at different biological levels over a long period of time. It is therefore usual to reuse the data for different research questions and projects. Although an existing tool may be useful for one project, no tool provides the required functionality for several different projects.

We have therefore implemented Kvik, a framework that makes it easy to develop new applications to explore different research questions and data. We have identified five requirements for such applications:

Interactive The applications should provide interactive exploration of datasets through visualizations and integration with relevant information.

Familiar They should use familiar visual representations to present information to researchers.

Simple to use Researchers should not need to install software to explore their data through the applications.

Flexible Applications should provide support for data from several study designs. This requires the framework to adapt to the statistical analyses used by the applications.

Lightweight Applications and statistical analyses should be separated to make it possible for researchers to explore data without having to have the computational power to run the analyses.

There are several tools for exploring biological data in the context of pathways, such as VisANT (available online at visant.bu.edu) by[1], VANTED (available online at vanted.ipk-gatersleben.de)[2], enRoute by[3] or Entourage by[4] (both available online at caleydo.org). However, these tools do not provide the adaptability needed for exploration of multi-exposure datasets. Many existing tools place the visualization, data analysis and storage on the user's computer, making it necessary to have a powerful computer. In addition, the tools are often stand-alone applications that require users to install them and keep both application and data up to date. In this article we describe how we used Kvik to implement Kvik Pathways, a tool for exploring gene expression in the context of biological pathways. It solves the above requirements as follows:

Interactive Kvik Pathways provides interactive pathway visualizations and information from the popular Kyoto encyclopedia of genes and genomes (KEGG)[5] database (available online at kegg.jp).

Simple to use Kvik Pathways uses HTML5 and modern JavaScript libraries to provide an interactive application that runs in any modern web browser.

Familiar Kvik Pathways uses the familiar pathway representations from KEGG and graphical user interfaces found in modern web applications.

Flexible It uses the R programming language for statistical analyses (r-project.org) so that researchers can tailor analyses to fit the specific research question in each project.

Lightweight Kvik Pathways uses a powerful backend provided by the Kvik framework to perform statistical analyses.

Both Kvik and Kvik Pathways are open-sourced at github.com/fjukstad/kvik. We provide an online version of Kvik Pathways at kvik.cs.uit.no and a Docker image at registry.hub.docker.com/u/fjukstad/kvik to run Kvik Pathways in a local Docker instance or on a cloud service such as Amazon Web Services (aws.amazon.com) or Google Compute Engine (cloud.google.com/compute).

Methods

Kvik Pathways allows users to interactively explore a molecular dataset, such as gene expression, through a web application. It provides pathway visualizations and detailed information about genes and pathways from the KEGG databases (Figure 1). The Kvik framework provides a flexible statistics back-end where researchers can specify the analyses they want to run to generate data to be used for later visualization. For example, in Kvik Pathways we retrieve fold change for single genes every time a pathway is viewed in the application. This function is run ad-hoc on the back-end servers and generates output that is displayed in the pathways in the client's web browser. All of these functions are implemented in a simple R script and can make use of all available libraries in R, such as Bioconductor (bioconductor.org).

Researchers modify this R script to, for example, select a normalization method, or to tune the false discovery rate (FDR) used to adjust the p-values that Kvik Pathways uses to highlight significantly differentially expressed genes. Since Kvik Pathways is implemented as a web application and the analyses are run ad-hoc, researchers get an updated application by simply refreshing the Kvik Pathways webpage.

Implementation

We implemented interactive visualizations using the Cytoscape.js (cytoscape.github.com/cytoscape.js) library to generate the interactive pathway visualizations, and D3 (d3js.org) for Document Object Model (DOM) manipulation such as generating bar charts with *svg* elements. We integrate these with the popular Bootstrap front-end framework (getbootstrap.com) to provide a familiar and aesthetically pleasing user interface.

Kvik Pathways has a three-tiered architecture of independent layers (Figure 2). The browser layer consists of the web application for exploring gene expression data and biological pathways. A front-end layer provides static content such as HTML pages and stylesheets, as well as an interface to the data sources with dynamic content such as gene expression data or pathway maps to the web application. The back-end layer contains information about pathways and genes, as well as computational and storage resources to process genomic data such as the NOWAC data repository. The Kvik framework provides the components in the back-end layer.

In our setup the Data Engine in the back-end layer provides an interface to the NOWAC data repository stored on a secure server on our local Stallo Supercomputer Table 1 provides the interfaces). In Kvik Pathways all gene expression data is stored on the computer
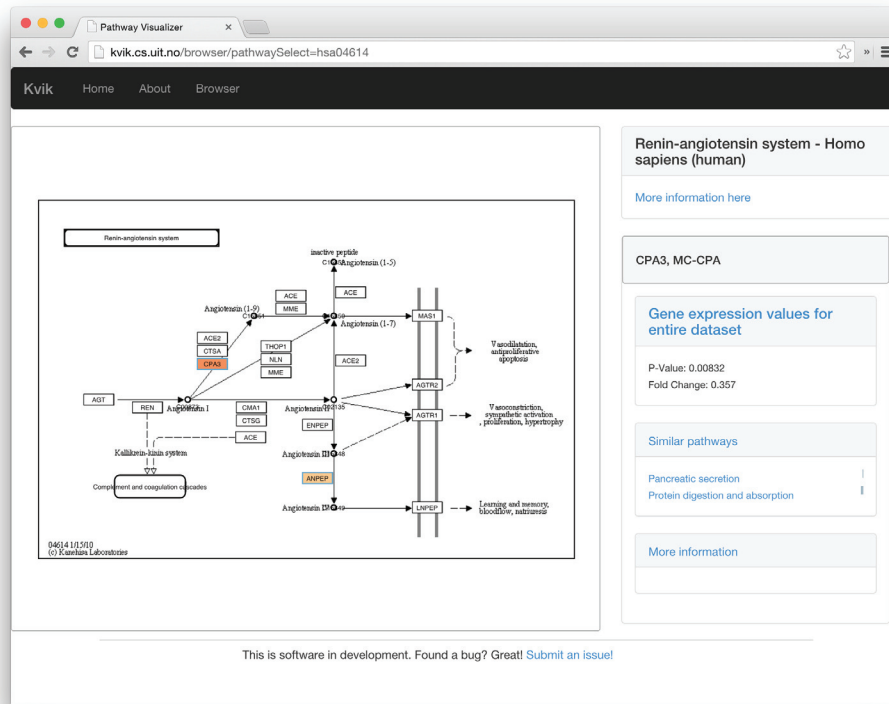
Figure 1. Screenshot of the renin-angiotensin pathway (KEGG pathway id hsa04614) in Kvik Pathways. The user has selected the gene CPA3 for further exploration.
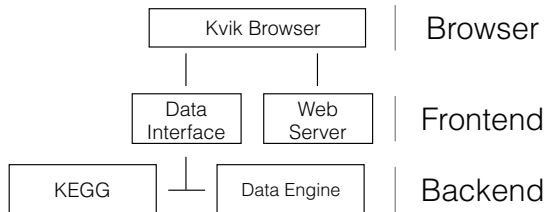

Figure 2. The three-tiered architecture of Kvik Pathways.

Table 1. The REST interface to the Data Engine. All URLs are relative to the hostname where the Data Engine server runs. On our public installation the Data Engine runs on kvik.cs.uit.no:8888. For example, use kvik.cs.uit.no:8888/genes/ to retrieve all available genes in our dataset.

URL	Description
/fc/[genes...]	Calculate and retrieve fold-change for the specified genes
/pvalues/[genes...]	Calculate and retrieve *p*-values for the specified genes
/exprs/[genes...]	Get the raw gene expression values from the dataset
/genes	Get a list of all genes in the dataset

that runs the Data Engine. The Data Engine runs an R session accessible over remote procedure calls (RPCs) from the browser layer using RPy2 (rpy.sourceforge.net) to interface with R.

To create pathway visualizations the Kvik backend retrieves and parses the KEGG Markup Language (KGML) representation and pathway image from KEGG databases through its REST API (rest. kegg.jp). This KGML representation of a pathway is an XML file that contains a list of nodes (genes, proteins or compounds) and edges (reactions or relations). Kvik parses this file and generates a JSON representation that Kvik Pathway uses to create pathway visualizations. Kvik Pathways uses the Javascript visualization library Cytoscape.js (js.cytoscape.org) to create a pathway visualization from the list of nodes and edges and overlay the nodes on the pathway image. To reduce latency when using the KEGG REST API, we cache every request on our servers locally. We use the average fold change between the groups in the sample set to color the genes within the pathway maps. To highlight *p*-values, the pathway visualization shows an additional colored frame around the node. We visualize fold change values for individual samples as a bar chart in a side panel. This bar chart gives researchers a global view of the fold change in the entire dataset.

Operation
Kvik Pathways runs in all modern web browsers and does not require any third-party software.

Use case

We used Kvik Pathways to repeat the analyses in a previous published project (6, doi: 10.1371/journal.pone.0067270) that compared gene expression in blood from healthy women with high and low plasma ratios of essential fatty acids. Gene expression differences between groups were assessed using *t*-tests (*p*-values adjusted with the Benjamini-Hochberg method). There were 184 differentially expressed genes significant on the 5% level. When exploring this gene list originally, functional information was retrieved from GeneCards and other repositories, and the list was analyzed for overlap with known pathways using MSigDB (available online at broadinstitute.org/gsea/msigdb). The researchers had to manually maintain overview of single genes, gene networks or pathways, and gather functional information gene by gene while assessing differences in gene expression levels. With this approach, researchers are limited by manual capacity, and the results may be prone to researcher bias.

Initially, Kvik Pathways was implemented to explore gene expression data from a not yet published dataset. To use Kvik Pathways to explore the data from the analyses in[6], we only needed to make small modifications to the R script used by the Data Engine. (The modified R script is found at github.com/fjukstad/kvik/blob/master/dataengine/data-engine.r). Instead of loading the unpublished dataset, we could load the dataset from[6] and reuse the functions that are accessible over RPC. Currently this script is less than 30 lines, consisting of four functions to retrieve data and a simple initialization step that reads the dataset. These functions are: *get*(*genes*), *genes*(), *f c*(*genes*) and *pvalues*(*genes*). *get* retrieves all information available for the given genes. *genes*() returns a list of all of the genes in the dataset. *f c*(*genes*) returns the fold change for the selected genes. *pvalues*(*genes*) returns the *p*-values for the given genes. After updating the R script in the Data Engine researchers using Kvik Pathways only had to reload a web page to get updated Kvik Pathways.

As an example of practical use of Kvik Pathways, we chose one of the significant pathways from the overlap analysis, the renin-angiotensin pathway (Supplementary table S5 in[6]). The pathway contains 17 genes, and in the pathway map we could instantly identify the two genes that drive this result. The color of the gene nodes in the pathway map indicates the fold change, and the statistical significance level is indicated by the color of the node's frame. We use this visual image of a biological process to see how these two genes (and their expression levels) are related to other genes in that pathway, giving a biologically more meaningful context as compared to merely seeing the two genes on a list.

Summary

Kvik Pathways is an open-source system for explorative analyses of functional genomics data from epidemiological studies. It uses R to perform on-demand data analyses providing a flexible back-end that can expand to new analyses and research projects. It uses modern visualization libraries and a powerful back-end for performing on-demand statistical analyses. Epidemiological researchers have used Kvik Pathways to analyze gene expression data. Kvik Pathways is open-sourced at github.com/fjukstad/kvik and is available as a Docker image at registry.hub.docker.com/u/fjukstad/kvik.

Data availability

Data used in the use case is available in the Gene Expression Omnibus (ncbi.nlm.nih.gov/geo), under accession number GSE15289.

Software availability

Latest source code
https://github.com/fjukstad/kvik

Source code as at the time of publication
https://github.com/F1000Research/kvik/releases/tag/1.0

Archived source code as at the time of publication
http://dx.doi.org/10.5281/zenodo.16375

Software license
The MIT license

Author contributions

LAB and BF designed the architecture of the system. BF implemented. All conducted the requirements analysis. EL, MJ, KSO contributed case study. BF drafted manuscript. All authors read, revised and approved the manuscript.

Competing interests

No competing interests were disclosed.

References

1. Hu Z, Chang YC, Wang Y, *et al.*: **VisANT 4.0: Integrative network platform to connect genes, drugs, diseases and therapies.** *Nucleic Acids Res.* 2013; **41**(Web Server issue): W225–W231.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**

2. Junker BH, Klukas C, Schreiber F: **VANTED: a system for advanced data analysis and visualization in the context of biological networks.** *BMC Bioinformatics.* 2006; **7**(1): 109.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**

3. Partl C, Lex A, Streit M, *et al.*: **enRoute: Dynamic path extraction from biological pathway maps for in-depth experimental data analysis.** In *Biological Data Visualization (BioVis), 2012 IEEE Symposium on,* pages 107–114.
   Publisher Full Text

4. Lex A, Partl C, Kalkofen D, *et al.*: **Entourage: visualizing relationships between biological pathways using contextual subsets.** *IEEE Trans Vis Comput Graph.* 2013; **19**(12): 2536–2545.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**

5. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Res.* 2000; **28**(1): 27–30.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**

6. Olsen KS, Fenton C, Frøyland L, *et al.*: **Plasma fatty acid ratios affect blood gene expression profiles--a cross-sectional study of the Norwegian Women and Cancer Post-Genome Cohort.** *PLoS One.* 2013; **8**(6): e67270.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Open Peer Review

Current Referee Status: ⁇ ⁇

Version 1

Referee Report 30 April 2015

Zhenjun Hu

Bioinformatics Graduate Program and Department of Biomedical Engineering, Boston University, Boston, MA, USA

The manuscript presents Kvik as an open-source system for explorative analysis of functional genomics data from large epidemiological studies. The authors seem have excellent ideas, but the implementation of the tool is far behind these ideas. I would like to approve the manuscript if the following points can be addressed:

1. The target of the tools. There are in general two type of tools: data provider and tool provider. The two of course can be combined. The prior in general provides knowledge, and the later provides functions to analyze users' own data. Kvik however seems lack the data to be a knowledge provider, and also does not provide enough functionality to be the later. To be the former, I will recommend authors to add more epidemiological data, to be the later, the author need to give clear instruction how user's own data can be analyzed using Kvik. For an example, the idea to connect to the cloud service is excellent, but how can Kvik to achieve this?

2. The implementation of Kvik seems to be improved, especially the performance. When I tried the Kvik, the browser tells me several time that the page is not responding. Yet I know the page is responding, but just take too much time. In addition, from data security point of view, it is not good to use RPCs from the browser layer to data engine directly, it shall be avoid in general in the three-tier architecture.

3. The manuscript need to focus more on the functionality of the tool. The current manuscript has too many technical details.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

Author Response 02 Jun 2015

Bjørn Fjukstad, University of Tromso, Norway

We would first like to thank the reviewer Zhenjun Hu for his thorough feedback and comments.

1. Since we have open-sourced the application we believe that Kvik Pathways can provide knowledge in itself, but also be used by other researchers to gain knowledge in their data. We have modified the use case to make it more clear how users can do so themselves. We also refer to the online repository where this information is also available. As we state in the Introduction we provide Docker images for running Kvik Pathways in a cloud service.

2. We agree, the current implementation can be improved. We are working on a second version where we have reduced the latency and that gives more feedback to the user if he has to wait. Regarding using RPCs from the browser layer we agree. We have updated the note with more details on how the system is implemented. The Data Engine provides an HTTP REST API that the browser layer queries. It does not send RPC requests from the browser layer, but from the frontend when it receives a request to the HTTP REST API. Also, the only RPCs that are allowed to run are defined in the R analysis script.

3. We agree. We have reduced the number of implementation details to make the functionality stand out more.

Competing Interests: No competing interests were disclosed.

Referee Report 23 April 2015

Paul Klemm

Department Simulation and Graphics, Faculty of Computer Science, Otto-von-Guericke University Magdeburg, Magdeburg, Germany

The work presented by Fjukstad *et al*. pursues in pushing the notion of open science in epidemiology. It describes Kvik, a web-based tool for analyzing genomic pathways. I really like the ideas behind it and value the detailed implementation section as well as the state-of-the-art techniques used. On the other hand I think that the paper should focus more on describing the epidemiological context, associated requirements and target groups to communicate the design choices for Kvik. Go into detail on the application *workflow*.

General Feedback

I like the approach the authors take with the paper. The tool they describe seems to be well suited for analyzing genomic pathways in the epidemiological context.

I miss a clear statement on the papers contribution. Maybe you can put a bullet list in the *Introduction* section to tell the reader what things can be done with your software, which could not be done before!

In my understanding, epidemiology is a very interdisciplinary in terms of associated experts. You have your clinicians deriving hypotheses from their day to day practice, statisticians deriving statistically sound conclusion as well as biologists and computer scientists associated with such projects. *Which of these are your target group?*

When you described your target users, describe what they are trying to find out. How does your tool help them doing that? Does it allow them do their work faster? Do they derive insights they could not get before? The latter would be a huge contribution! Please give more details on the *workflow* of you system!

I miss a clear distinction from the NIK-2014 paper "Kvik: Interactive exploration of genomic data from the NOWAC postgenome biobank". The paper was also not cited in this work. It seems to me that the majority of the content of the presented paper can already be found in the NIK-2014 paper. Please elaborate on the differences and cite the paper. *If you can not state clear differences, there is, in my opinion, no point in publishing this paper and I will rate it 'Not Approved'.*

Title
The title of the work is appropriate.

Abstract
The abstract motivates the need for new tools, which allow to assess the vast amount of epidemiological data well. In my opinion it can be improved by:
- reduce the amount of implementation detail. You tell the reader later on which frameworks and libraries you use
- explain who are your users.
- what can be done with your tool now, which could not be done before?

Minor comments on the abstract:
- "Existing data exploration tools do not provide all the required functionality for such multi-study data exploration." This is a dangerous statement, since you do not say anything about what the *required functionality* is! I think I know what you are trying to get at, in the introduction you describe it better with: "Although an existing tool may be useful for one project, no tool provides the required functionality for several different projects."

Introduction
The introduction can be improved by clearly stating the contributions (e.g., as bullet points).

I would like to see some reference or a method on how the five requirements were acquired. These are all things, which are important in almost all applications. Where is the difference of software in the epidemiological context towards other context and how does Kvik adapt to the arising requirements? You answer many of these questions, later on when you repeat all the requirements again, but to me it is not structured well.

Methods
The method section is written well. I would like to know how the users modify the R scripts (beginning second paragraph). Do they do this inside Kvik or do they have to switch into another software for it?

Figure 1 caption: What can the user do now after he or she selected the gene? The workflow is not clear to me.
Figure 2 was already presented very similarly in the NIK'14 paper.

Minor: Closing parenthesis in sentence "In our setup the Data Engine in the back-end layer provides an interface to the NOWAC data repository stored on a secure server on our local Stallo Supercomputer Table 1 provides the interfaces)."

Use Case
The use case section can be strengthened by reducing the amount of implementation details (in my

opinion mentioning the individual function names is not necessary to comprehend the functionality) and focusing more on the involved actors and tasks and contexts associated with the use case. What feedback was given by the user(s)?

Reusability
The effort of the authors to make the software publicly available is worth a special note. Modern state of the art techniques are combined with powerful back-end systems, which scale well on different application scenarios.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

Author Response 02 Jun 2015
Bjørn Fjukstad, University of Tromso, Norway

We would first like to thank the reviewer Paul Klemm for his thorough feedback and comments.

Difference between the NIK paper and the Application Note

When we wrote the NIK paper, Kvik was only a system for exploring genomic data in the context of pathways. Since then we have realized that exploring genomic data in pathways is not enough, that we need a framework that allows us to build different applications for exploring data from different studies with different designs. With this in mind we have refined the requirement analysis from our initial Kvik system and developed more general requirements that these applications should satisfy.

From our initial Kvik implementation we have now decoupled the application (Kvik Pathways) from the framework, allowing fast development of new applications. The Kvik framework provides interfaces to a Data Engine that provides statistical analyses, and interfaces to online databases such as KEGG. Kvik Pathways is the first application that we developed using the Kvik Framework. Using Kvik we have developed several other applications that will be published in the near future.

In the NIK paper we described from a computer science point of view the features of Kvik, both looking at the application itself, and the backend features that are now a apart of the Kvik Framework. The NIK paper was written to give a more in detail view of how the system works and performs, while in this application note we want to describe how our epidemiology researchers helped to develop the application and how they used it to reproduce results they found in an already published dataset.

Using an already published dataset was important to us since it allows us to provide a publicly available Kvik installation for others to use. We will revisit the second paragraph of the Use case section where we discussed how we used the initial Kvik system to explore different data from a different study design.
To sum up, our new contributions in the application note are as follows:
- Publicly available application

- Publicly available Docker containers that researchers can use to set up local installations of Kvik Pathways.

- Reproduced the results from an already published dataset to make the system publicly accessible at kvik.cs.uit.no

- A more refined requirement analysis that reflects our experiences after publishing the NIK paper. The important changes are:

   i) emphasis on integration of online knowledge bases (interactive requirement),

   ii) emphasis on the system being flexible to adapt to data and different statistical analyses,

   iii) we removed the security since we believe that data should be publicly available, and

   iv) put emphasis on separating computation and visualization (lightweight).

We have cited the NIK paper in the application note and improved the text to highlight the differences between the framework and the Kvik Pathways application.

We have included a list of contributions in the Introduction section.

We agree that epidemiology is a very interdisciplinary. Kvik Pathways has been developed in a team of epidemiologists, biologists, statisticians and computer scientists. The application note targets such groups of researchers working together to develop systems for gaining biological insights in genomic data. We have re-written parts of the note to clarify how researchers have used the application.

Abstract : We agree and have reduced the amount of implementation detail and made it more specific what our users have done with the application.
Introduction: We have revisited the requirements and specified how these are different from regular applications. We believe that it is best to separate the requirements and how we solved them in two different lists.

Methods: As of today users modify R scripts outside Kvik. We have made it clear in the methods section. We will expand Figure 1 caption to clearify the workflow. Regarding figure 2. We chose to include it since it highlights the important three-tiered architecture with applications that use the Kvik framework. We have modified the figure to highlight connection between the application and the framework.

Use case: As mentioned we will expand this section with a more detailed workflow. Since it is an app note targeted towards users we will reduce implementation details and refer to the source code.

Competing Interests: No competing interests were disclosed.