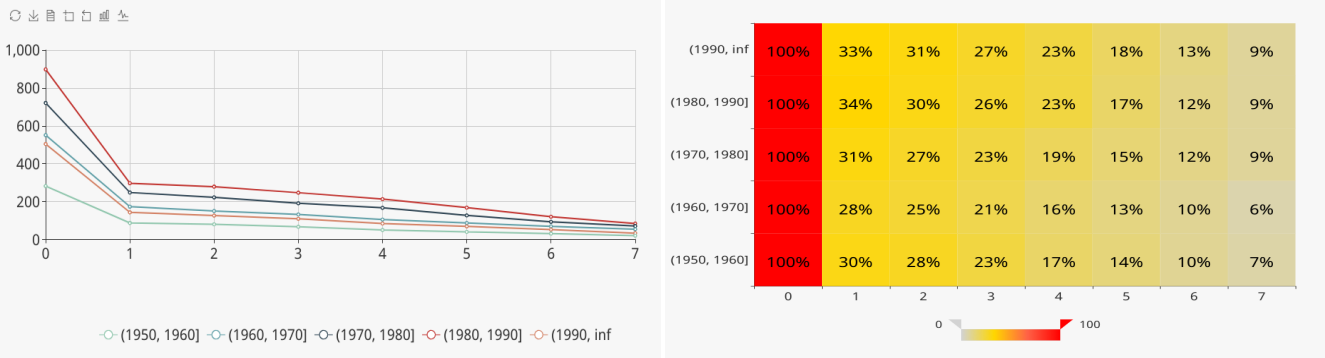

Figure 4: Result Visualization

small range can be chosen for short-term effects while a large one can be used for long-term effects.

The next selection is to decide the patients to be included in the analysis, which is specified in the "User Selection" panel. By default, all patients will be included. In Figure 3, those who experienced event "diagnose" with diagnosing result "Disease-B" within any week are specified as the target patients. More constraints on patient selection can be cast by ticking the + button on the "Event" line.

For the "Birth Criteria" panel, birth event "prescribe" and the corresponding medicine "Medicine-A" are specified. In addition, a qualified patient selected into a cohort should take the medicine at least twice since the frequency of this birth event is specified to be larger than or equal to 2. If multiple birth events are assigned by ticking the + button, a patient will be selected into cohorts only when all the birth events are met. In the last part of this panel, it is indicated that the qualified patients should be grouped by their "birthyear" in the range of *1950 to 2000* with a scale of 10 years.

3.3 Result Visualization

Within seconds of submitting the cohort analysis task, the analysis result is returned and visualized in various ways in frontend to help doctors better understand it. Here we choose a line chart and a heat map, as shown in Figure 4. In the line chart, the x-axis represents the *age* of the patients since their birth, i.e., the number of days after taking medicine A twice, and the y-axis represents the aggregation result, i.e., the number of patients with abnormal values detected for Labtest-C. The value of y-axis on age 0 is the total number of patients selected into the cohort. Each line stands for a cohort of patients that were physically born in a same decade. This line chart, answering the cohort query numerically, not only illustrates the trend of patient behavior along the time axis, but also offers a view of the difference in the behavior of different cohort patients.

The heat map is presented along age and cohort group dimensions. A cell (i, j) in the map represents the proportion of patients in cohort i that are detected with an abnormal Labtest-C value at age j. Having cells be shaded with different color depths in accordance to its metric value gives spontaneous expression on the evolvement of patient behavior in terms of Labtest-C result, and indicates deep insight into patient behavior between different cohorts.

The two charts explain the result of the submitted cohort query in an absolute and relative manner respectively. We can observe that younger patients are actively exhibiting side effects, as suggested by the high values in group (1980, 1990] in the line chart, while elder patients take longer to get accustomed to the medicine.

Additionally, with the help of ECharts, we provide a wide variety of operations to manipulate the chart, such as zooming in and changing the chart type, as shown in the toolbar above the line chart. These functions enable immediate responses to visualization exploration, which is helpful for further data excavation.

4 RELATED WORK AND CONCLUSION

Recently, due to the increasing volume of web user behavior data, cohort analysis has been introduced to mine unusual user behavior and improve user retention as proposed in [1, 5, 6]. However, these products directly follow the single birth event specification, restricting the diversity and representativeness of the cohorts generated. Moreover, they all employ simple SQL-based approaches, which are much slower than COHANA.

This demonstration shows a powerful and comprehensive tool on top of COHANA for cohort analysis while keeping operations as simple and intuitive as possible. The extended cohort concepts cover a broad range of practical data analysis needed in many domains concisely. Benefiting from the high efficiency of the COHANA engine, analysts without any backgrounds can conduct reporting or verify their ideas in minimal time with a user-friendly interface. Finally, the visualization of query results provides deep understanding and insights into user behavior.

ACKNOWLEDGMENT

This research was in part supported by the National Research Foundation, Prime Ministers Office, Singapore, under its Competitive Research Programme (CRP Award No. NRFCRP8-2011-08). Zhongle Xie's work was partially supported by the National Research Foundation Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme (E2S2-SP2 project).

REFERENCES

[1] Amplitude. 2017. Amplitude Official Site. https://amplitude.com.
[2] Qingchao Cai, Zhongle Xie, Meihui Zhang, Gang Chen, H.V. Jagadish, and Beng Chin Ooi. 2018. Efficient Temporal Dependence Discovery in Time Series Data. *PVLDB*.
[3] Norval D Glenn. 2005. *Cohort analysis*. Vol. 5. Sage.
[4] Dawei Jiang, Qingchao Cai, Gang Chen, H.V. Jagadish, Beng Chin Ooi, Kian-Lee Tan, and Anthony K.H. Tung. 2016. Cohort query processing. *PVLDB* 10, 1–12.
[5] Mixpanel. 2017. The definition of retention. https://mixpanel.com/retention/.
[6] RJMetrics. 2017. RJMetrics Official Site. https://rjmetrics.com/.